ICAIBDA

October 27-29, 2021

1st INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE & BIG DATA ANALYTICS (ICAIBDA)

PROCEEDING BOOK

"Breakthrough Research of **Artificial Intelligence** and **Big Data** for Post-Pandemic Recovery"

ISBN 978-1-6654-0890-5

Sponsored By : IEEE ComSoc IEEE Communications Society Indonesia Chapter; UNIVERSITAS PADJADJARAN

Organized By : AIDA RESEARCH CENTER FOR ARTIFICIAL INTELLIGENCE AND BIG DATA UNPAD

Co-organized By : Fakultas MIPA Universitas Padjadjaran

2021 International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA)

October 27th-29th 2021

Bandung, Indonesia

ISBN 978-1-6654-0890-5

Publisher IEEE

Secretariat Research Center for Artificial Intelligence and Big Data

Universitas Padjadjaran

Jalan Dipati Ukur No. 35 Bandung 40132

E-mail: pusris.aida@unpad.ac.id

Author Index

TABLE OF CONTENTS

2021 International Conference on Artificial Intelligence and Big Data Analytics Committee

1. Honorary Chairs
   - Prof. Dr. Rina Indiastuti, M.SIE., Rector of Universitas Padjadjaran, Indonesia
   - Prof. Rizky Abdullah, Ph.D., Apt, Universitas Padjadjaran, Indonesia
2. General Chair
   - Dr. Intan Nurma Yulita, M.T, Universitas Padjadjaran, Indonesia
3. Secretariat Chair
   - Herlina Napitupulu, Ph.D, Universitas Padjadjaran, Indonesia
4. Finance Chair
   - Dr. Afrida Helen, M.T, Universitas Padjadjaran, Indonesia
5. Technical Program Chair
   - Dr. Anindya Apriliyanti Pravitasari, Universitas Padjadjaran, Indonesia
   - Gandeva Bayu Satrya, Ph.D, Telkom University, Indonesia
6. Publication Chair
   - Fajar Indrayatna, Universitas Padjadjaran, Indonesia
   - Linda Kurniawati, Universitas Padjadjaran, Indonesia
7. Technical Committee
   - Dr. Khoirul Anwar, Telkom University, Indonesia
   - Dr. Radial Anwar, Telkom University, Indonesia
   - Ravimal Bandara, University of Sri Jayewardenepura, Sri Lanka
   - Dr. Joan Bas, Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Spain
   - Dr. Achmad Syahrul Choir, Politeknik Statistika STIS, Indonesia
   - Dr. Noriko Etani, All Nippon Airways Co., Ltd., Japan
   - Alireza Ghasempour, ICT Faculty, USA
   - Dr. Irlandia Ginanjar, Universitas Padjadjaran, Indonesia
   - Prof. Pitoyo Hartono, Chukyo University, Japan
   - Dr. Diptarama Hendrian, Tohoku University, Japan
   - Hendry Hendry, Chaoyang University of Technology, Taiwan
   - I Dewa Putu Hermida, Indonesian Institute of Sciences (LIPI), Indonesia
   - Dr. Nyoman Karna, Telkom University, Indonesia
   - Dr. Ahmed Kawther, Mustansiriyah Universtiy, Iraq
   - Dr. Adrian Kliks, Poznan University of Technology, Poland
   - Dayat Kurniawan, Indonesian Institute of Sciences, Indonesia
   - Dr. Sidath R Liyanage, University of Kelaniya, Sri Lanka
   - Dr. Wayan Firdaus Mahmudy, Universitas Brawijaya, Indonesia
   - Dr. Ratheesh Kumar Meleppat, University of California Davis, USA
   - Dr. Devarani Devi Ningombam, Gandhi Institute of Technology and Management (GITAM) University, India
   - Prof. Indumathi Raghavan, Mepco Schlenk engineering college, India
   - Dr. Juli Rejito, Universitas Padjadjaran, Indonesia
   - Indra Riyanto, Universitas Indonesia, Indonesia
   - Dr. Sarah Flora Samson Juan, UNIMAS, Malaysia
   - Dr. Arief Suryadi Satyawan, Waseda University, Japan
   - Prof. Chin-Shiuh Shieh, National Kaohsiung University of Science and Technology, Taiwan
   - Dr. Suherman Suherman, Universitas Sumatera Utara, Indonesia
   - Dr. Arwin Datumaya Wahyudi Sumari, State Polytechnic of Malang, Indonesia
   - Andang Sunarto, Tadris Matematika, Institut Agama Islam Negeri (IAIN) Bengkulu Indonesia, Indonesia
   - Dr. Sugiyarto Surono, University of Ahmad Dahlan, Indonesia
   - Dr. Hadi Sutopo, Kalbis Institute, Indonesia
   - Suyoto Suyoto, Lembaga Ilmu Pengetahuan Indonesia, Indonesia

- Prof. Eirini Eleni Tsiropoulou, University of New Mexico, USA
- Dr. Adriaan J. van Wijngaarden, Bell Laboratories, Nokia, USA
- Dr. Vinod Kumar Verma, SLIET, LONGOWAL, India
- Dushyanthi Vidanagama, General Sir John Kotelawala Defence University, Sri Lanka
- Dr. Nimalka Wagarachchi, Faculty of Engineering, University of Ruhuna, Sri Lanka
- Dr. Ameen Aledani, University of Basrah, Iraq
- Prof. Mikulas Alexik, University of Zilina, Slovakia
- Dr. Norma Alias, Universiti Teknologi Malaysia, Malaysia
- Prof. Hasan Aydogan, Selcuk University Kampus Selcuklu, Turkey
- Dr. Roman Dunaytsev, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, Russia
- Heba Fadhil, University of Baghdad, Iraq
- Handoko Handoko, Universitas Kristen Satya Wacana, Indonesia
- Prof. Shi-Jinn Horng, National Taiwan University of Science and Technology, Taiwan
- Dr. Yong-Jin Lee, Korea National University of Education, Korea (South)
- Murinto Murinto, Universitas Ahmad Dahlan, Indonesia
- Dr. Dmitry Namiot, Moscow State University, Russia
- Syafrial Fachri Pane, Applied Bachelor Program of Informatics Engineering, Politeknik Pos Indonesia, Indonesia
- L. M. Rasdi Rere, University of Indonesia, Indonesia
- Taufik Ridwan, Universitas Pendidikan Indonesia, Indonesia
- Mujiono Sadikin, University of Mercu Buana, Indonesia
- Dr. Selvathi Selvathi, Mepco Schlenk engineering college, India
- Dr. Joni W. Simatupang, President University, Indonesia
- Darma Tasia, Universitas Islam Negeri Alauddin Makassar, Indonesia
- Prof. Kasturi Vasudevan, Indian Institute of Technology Kanpur, India
- Ionia Veritawati, Universitas Pancasila, Indonesia
- Theophilus Wellem, Satya Wacana Christian University, Indonesia
- Prof. Eduard Babulak, Liberty University, USA
- Prof. Dragana D. Bajić, University of Novi Sad, Serbia
- Dr. Setiawan Hadi, Universitas Padjadjaran, Indonesia
- Dr. Dimitrios Koukopoulos, University of Patras, Greece
- Dr. Rathnayaka Mudiyanselage Kapila Tharanga Rathnayaka, Sabaragamuwa University of Sri lanka, Sri Lanka
- Dr. Astari Retnowardhani, Bina Nusantara University, Indonesia
- Dr. Ali Othman Al Janaby, Ninevah University, Iraq
- Rani Megasari, Indonesia University of Education, Indonesia
- Fitra A. Bachtiar, Brawijaya University, Indonesia
- Didit Didit Widiyanto, UPN Veteran Jakarta, Indonesia
- Toni Toharudin, Universitas Padjadjaran, Indonesia
- Nisa Hanum Harani Harani, Politeknik Pos Indonesia, Indonesia
- Dr. Muhammad Aufaristama, University of Twente, The Netherlands
- Rolly Maulana Awangga, Institut Teknologi Bandung, Indonesia
- Dr. Dwina Kuswardani, Institut Teknologi PLN, Indonesia
- Dr. Rindang Bangun Prasetyo, Politeknik Statistika STIS, Indonesia
- I Gede Eka Wiantara Putra, Politeknik Nasional Denpasar, Indonesia
- Dr. Shahzad A, NFC Institute of Engineering & Technology, Pakistan
- Dr. Devi Fitrianah, Universitas Mercu Buana, Indonesia
- Prof. Atje Setiawan Abdullah, Universitas Padjadjaran, Indonesia

Welcome Message from General Chair of ICAIBDA 2021

On behalf of organizing committee, I am honored and delighted to welcome all distinguished guests, keynotes and invited speakers, and participant to the 2021 IEEE International Conference on Artificial Intelligence and Big Data Analytics. This event is the first international conference which organized by Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran, sponsored by IEEE ComSoc. In particular, the theme of ICAIBDA 2021 is "Breakthrough Research of Artificial Intelligence and Big Data for Post-Pandemic Recovery". We can share and learn many current topics in various tracks including Artificial Intelligence, Big Data, Communication, and others research fields. We do hope this event could facilitate all participants to interact each other intensively for extending scientific network in the future.

We are pleased to inform that all participants attending this conference are from several countries, i.e. Indonesia, China, Japan, and Turkey. In closing, we wish to express our gratitude to all keynotes and invited speakers who have come to share their knowledges in this event. Also, we gratefully thank to all participant for your contribution to ICAIBDA 2021. In advance, we would like to extend our appreciation and sincerely thanks to Universitas Padjadjaran for supporting this conference. Finally, I would like to thank the advisory board, steering and organizing committee members for making this event occurred. We wish all participants a very fruitful and pleasant scientific programs of this conference

Thank you,

Dr. Intan Nurma Yulita, M.T

(ICAIBDA 2021, Chair)

Technical Program

WEDNESDAY, 27 OCTOBER 2021 (DAY 1)				
TIME (UTC +7)	**SESSION**	**SPEAKER**		
12:40-12:55	Registration	**Committee**		
12:55-13:00	Opening Session	**MC: Marsha Stephanie Manurung**		
13.00-13.05	Indonesia Raya			
13:05-13:10	Chair of the IEEE Communication Society, Indonesia Chapter	**Dr. Wiseto Agung**		
13:10-13:15	Vice Rector for Research and Inovation	**Prof. Dr. Ir. Hendarmawan, M.Sc.**		
13:15-13:25	ICAIBDA General Chair	**Dr. Intan Nurma Yulita, M.T.**		
13.25-13.30	MC introduces moderator	**MC: Marsha Stephanie Manurung**		
TIME (UTC +7)	**SESSION**	**SPEAKER**	**TITLE**	**MODERATOR**
13:30-14:30	Keynote Speaker 1	**Prof. Boris Mirkin** National Research University Higher School of Economics, Moscow RF & Birkbeck University of London UK	"Method for Taxonomic Content-Analysis of Text Collections"	Dr. Setiawan Hadi, M.Sc.CS.
14:30-15:30	Keynote Speaker 2	**Dr. Mahardika Pratama** Nanyang Technological University, Singapore	“Continual and Autonomous Machine Learning”	
15:30	Announcement	**MC: Marsha Stephanie Manurung**		
PARALEL SESSION 1 (BREAKOUT ROOM)				
ROOM 1 (5 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
	1570758102			

15:45-16:00		Md Sabbir Hosen (China)	Dynamic Channel Allocation Technique for Cognitive Radio Based UAV Networks	Mira Suryani, S.Pd., M.Kom.
16:30-16:45	1570764569	Majan Abdullah Al Jahdhami (Turkey)	Performance Analysis of Mobile Broadband Networks in Ibra City, Oman	
16:45-17:00	1570768872	Riki Murakami (Japan)	Dynamic Topic Tracking and Visualization Using COVID-19 Related Tweets in Multiple Languages	
ROOM 2 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
16:15-16:30	1570760920	Faizal Sudrajat	Simulation of A Decision Support System Using Data Mining Method with C4.5 Algorithm: A Case Study	Dr. Intan Nurma Yulita, M.T
16:30-16:45	1570762124	Eka Sari Oktarina	A New Tele-Healthcare System of Heart Sound Signal Classification Based on Gated Recurrent Unit	
16:45-17:00	1570763867	Endang Wahyuni	Outlier Detection Using K-Means Clustering at the Inquiry-Based Learning Results Dataset in Students	
17:00-17:15	1570763934	Dita Fadma Ristianti	Optimization of Fuzzy Support Vector Machine (FSVM) Model in Multiple Metric Spaces	
ROOM 3 (5 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
15:45-16:00	1570756842	M. Aldiki Febriantono	Transfer Learning-Based Mobile-Focused Automated COVID-19 Detection from Chest X-Ray	Dessy Novita ST, MT, Ph.D
16:00-16:15	1570766172	Muhamad Fahriza Novriansyah	The Impact of Large-Scale Social Restriction and Odd-Even Policies During Covid-19 Pandemic to Traffic Congestion and Air Pollution in Jakarta	

16:15-16:30	1570766803	Zein Himami	Pathologic Myopia Detection from Retinal Fundus Images Using Deep Learning Algorithms	
16:30-16:45	1570766855	Dennis Dennis	Sentiment Classification Against the Public Activity Restrictions Policy in Jakarta Using Machine Learning Models	
16:45-17:00	1570761061	Heliza Rahmania Hatta	Diagnosis of Aglaonema Plant Disease Using Forward Chaining and Naive Bayes Methods	

THURSDAY, 28 OCTOBER 2021 (DAY 2)				
TIME (UTC +7)	**SESSION**	**SPEAKER**	**TITLE**	**MODERATOR**
08.55-09.00	Opening			
09:00-09:05	MC introduces moderator	**MC: Brilliant Dimas Akbarrezza Purwadi**		
09:05-10:05	Keynote Speaker 3	**Assc. Prof. Norma Alias** Universiti Teknologi Malaysia, Malaysia	"Predictive Method of Artificial Intelligence for Big Data Enhance Decision Making"	Dessy Novita ST, MT, Ph.D
10:05-10:35	Invited Speaker 1	**Prof. Adiwijaya** Rector of Telkom University, Indonesia	“A Survey : Cancer Classification Based on Microarray Data”	
10:35-11:00	Break			
TIME (UTC +7)	**SESSION**	**SPEAKER**	**TITLE**	**MODERATOR**
11:00-11:30	Invited Speaker 2	**Yudhie Andriyana, Ph.D.** Vice Dean of FMIPA Universitas Padjadjaran	"A robust flexible modeling technique to overcome a curse of dimensionality"	Dr. Anindya Apriliyanti Pravitasari, M.Si
11:30-12:00	Invited Speaker 3	**Prof. Nur Iriawan, M.Ikom, Ph.D** Institut Teknologi Sepuluh November, Indonesia	"AI in VSD of Ultrasound Video: Detection and Segmentation"	
12:00-12:10	Setelah pemberian sertif: 1. Adlibs - Sponsored by and Supported by 2. TVC Dewaweb 3. Adalibs Dewaweb			

12.10.13.00	Break			
13:00-13:30	Invited Speaker 4	**Assc. Prof. Ts. Dr. Wan Muhamad Amir** Universiti Sains Malaysia, Malaysia	"The Craniofacial Research: A Methodological Development in Biostatistics"	Mira Suryani, S.Pd., M.Kom.
13:30-14:00	Invited Speaker 5	**Assc. Prof. Ghosh Ankush** The Neotia University, India	"Machine Learning for Autonomous Driving"	
14:00-14:10	Announcement	**MC: Brilliant Dimas Akbarrezza Purwadi**		
PARALEL SESSION 2 (BREAKOUT ROOM)				
ROOM 1 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
15:15-15:30	1570768750	Dicky Rahma Hermawan	Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle Coupon Recommendation Data	Erick Paulus, M.Kom
15:30-15:45	1570768779	Yuela Thahira	Comparison of Adolescent Vaccination Data Accuracy by Urban Village in DKI Jakarta Province in July 2021 Using Several Data Mining Methods	
15:45-16:00	1570769382	Difa Bagasputra Maulana	Data Prediction Analysis of the Third Wave Covid-19 in Bogor Regency	
16:00-16:15	1570769074	Kurniadi	Indonesian Digital Wallet Sentiment Analysis Using CNN and LSTM Method	
16:15-16:30	1570767989	Ayu Masyita Putri	Sentiment Analysis of YouTube Video Comments with the Topic of Starlink Mission Using LSTM	
16:30-16:45	1570767843	Luki Setiawan	Prediction of Stock Price Data of PT. Ramayana Lestari Sentosa Tbk. Using Long Short Term Memory Model	
ROOM 2 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**

15:15-15:30	1570767273	Syahrul Qolbi Nur Septi	Face Recognition Using Fisherface and Support Vector Machine Method	Dr. Afrida Helen, M.Kom
15:30-15:45	1570768503	Daffa Rahmansyah Danistya	Prediction New Cases of Covid-19 in Indonesia Using Vector Autoregression (VAR) and Long-Short Term Memory (LSTM) Methods	
15:45-16:00	1570768735	Fahri Zulkarnaen	Indonesian Abusive Tweet Classification Based on CNN and LSTM Method	
16:00-16:15	1570769387	Nugroho Sinung Adi	Stroke Risk Prediction Model Using Machine Learning	
16:15-16:30	1570767283	Dio Satyaloka	SK-MOEFS Multi-Objective Evolutionary Fuzzy System Library Effectiveness as User-Friendly Cryptocurrency Prediction Tool	
16:30-16:45	1570767294	Muhamad Farid Ridho Rambe	Wavelet Neuro Fuzzy System (WNFS) in Predicting the Exchange Rate of the Rupiah Against the US Dollar	
ROOM 3 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
15:15-15:30	1570767295	Sheila Azhar	Indonesian Food Price Prediction Method with Adaptive Neuro-Fuzzy Inference System (ANFIS)	Herlina Napitupulu, Ph.D.
15:30-15:45	1570767608	Anne Fernanda	Stock Price Prediction of ITMG Issuers Using the Adaptive Neuro-Fuzzy System (ANFIS) Method	
15:45-16:00	1570746601	Fauzi Faruq Nabbani	Twitter Social Media Sentiment Analysis of Telecominications Service Provider Using Long Short-Term Memory Method	
16:00-16:15	1570764220	Raihan Badrahadipura	COVID-19 Detection in Chest X-Rays Using Inception Resnet-v2	
16:15-16:30	1570766586	Arif Rhizky Gilang Purnama	Search System for Translation of Al-Qur'an Verses in Indonesian	

			Using Bm25 and Semantic Query Expansion	
16:30-16:45	1570767275	Farah Savitri	Forecast Inflation in Indonesia Using Long Short Term Memory	
ROOM 4 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
15:15-15:30	1570768716	Khansa Afifah	Sentiment Analysis on Telemedicine App Reviews Using XGBoost Classifier	Dr. Anindya Apriliyanti Pravitasari, M.Si
15:30-15:45	1570767644	Rividya Aluna	Electronic News Sentiment Analysis Application to New Normal Policy During the Covid-19 Pandemic Using Fasttext and Machine Learning	
15:45-16:00	1570767910	Cryssa A. E. Piter	Multi-Label Classification for Scientific Conference Activities Information Text Using Extreme Gradient Boost (XGBoost) Method	
16:00-16:15	1570767987	Mochamad Lutfi	LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, and Random Forest	
16:15-16:30	1570768032	Sarah Hasna Azzahra	Text Categorization of Job Vacancy Using Recurrent Neural Network Method	
16:30-16:45	1570768715	Sachi Hongo	Design of Museum Historical Heritage Management System Using Blockchain Digital Certificate and Hyperledger Composer	
FRIDAY, 29 OCTOBER 2021 (DAY 3)				
PARALEL SESSION 3 (BREAKOUT ROOM)				
ROOM 1 (5 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
09:00-09:05	Opening			Dr. Asep Sholahuddin, MT

09:05-09:20	1570769379	Sulthan Nior	Star Type Classification Using A 6 Class Star Dataset with J48 Decision Tree, Naïve Bayes & Random Forest	
09:35-09:50	1570769385	Farhan Gunadi	Preprocessing Application for Car Insurance Claim Classification Model	
10:05-10:20	1570770971	Dimas Ananda	Application of the Bidirectional Gated Recurrent Unit (BiGRU) in the Sentiment Analysis of Tokopedia Application Users	
10:20-10:35	1570769352	Syifa Auliyah Hasanah	A Comparison of SVM and Naïve Bayes Classifier in Binary Sentiment Reviews for PeduliLindungi Application	
10:35-10:50	1570769048	Muhammad I'tikafi Khoirul Haq	Classification of Water Potability Using Machine Learning Algorithm	
ROOM 2 (6 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
09:00-09:05	Opening			Dr. Intan Nurma Yulita, M.T
09:05-09:20	1570737018	Intan Nurma Yulita	AdaBoost Support Vector Machine Method for Human Activity Recognition	
09:35-09:50	1570769456	Sari Widya Sihwi	Automatic Text Summarization with Categorization on Online News About Indonesia's Public Figure Using Fuzzy Logic Method	
10:05-10:20	1570769698	Intan Nurma Yulita	Utilization of Data Warehouse in Business Intelligence with Kimball Method	
10:20-10:35	1570770556	Refa Annisatul Ilma	Twitter's Hate Speech Multi-Label Classification Using Bidirectional Long Short-Term Memory (BiLSTM) Method	
10:35-10:50	1570769396	Aditya Rizky Fadillah	Data Mining Implementation Using Frequent Pattern Growth on Transaction Data for	

			Determining Cross-Selling and Up-Selling (Case Study: Cascara Coffee)	
10:50-11:05	1570766599	Mulya Nurmansyah Ardisasmita	COVID-19 Social Safety Nets Sentiment Analysis on Twitter Using Gated Recurrent Unit (GRU) Method	
11:05-11:20	1570769700	Intan Nurma Yulita	Market Basket Analysis on Sales Transactions for Micro, Small and Medium Enterprises Using Apriori Algorithm to Support Business Promotion Strategy in RDA Hijab	
ROOM 3 (5 presenter)				
TIME (UTC +7)	**PAPER ID**	**PRESENTER**	**TITLE**	**MODERATOR**
09:00-09:05	Opening			Linda Kurniawati, B.Eng., MM
09:05-09:20	1570768720	Natalia Kustanto	Sentiment Analysis of Indonesia's National Health Insurance Mobile Application Using Naïve Bayes Algorithm	
09:35-09:50	1570769392	Dhiya Uzh Zharief	The Effect of Facial Attributes in Identifying Gender Using Facial Recognition	
10:05-10:20	1570769390	Shafira Khoirunnisa	The Effect of Educational Background on High Jobs and Income	
10:20-10:35	1570770995	Paquita Putri Ramadhani	Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method	

Keynote Speaker 1

Method for Taxonomic Content-Analysis of Text Collections

Boris Mirkin

NRU HSE Moscow RF and Birkbeck University of London UK

The core of the method is a formalization of the notion of generalization, a property of higher cognitive systems. To achieve that, we use a pre-defined domain taxonomy -- a rooted tree of the domain whose nodes are labeled with domain concepts so that more general concepts correspond to nodes of higher ranks.

The method includes the following steps: (1) computing matrix of relevance between texts and taxonomy leaf topics using a purely structural string-to-text relevance measure based on suffix trees representing the texts and annotated by substring frequencies, (2) obtaining fuzzy clusters of taxonomy leaf topics using a method involving both additive and spectral properties, and (3) finding most specific generalizations of the fuzzy clusters in a rooted tree of the taxonomy. Such a generalization parsimoniously lifts a cluster to its ``head subject'' in the higher ranks of the taxonomy, to tightly cover the cluster by minimizing the number of errors, ``gaps'' and ``offshoots''. The method applies to two collections of research papers in the data science domain: (a) ~18000 research papers published in 17 Springer journals for 20 years, and (b) ~27000 research papers retrieved from Springer and Elsevier journals in response to data science related queries. Our taxonomy of Data Science is derived from the Association for Computing Machinery Classification of Computing System (ACM-CCS). We observe some tendencies of research that cannot be seen by using more conventional techniques [1].

Currently we explore a version of the lifting method based on the maximum likelihood criterion; the event probabilities are computed using multiple runs of the parsimonious lifting method. If time permits, I am going to tell of a different application of the lifting method: audience extension in targeted advertising over internet [2]. Joint work with T. Fenner (University of London, UK), S. Nascimento (New University of Lisbon), D. Frolov (NRU HSE Moscow), Z. Airapetyan (NRU HSE Moscow), Z. Taran (Delta University, MS USA).

References

Frolov, D., Nascimento, S., Fenner, T., & Mirkin, B. (2020). Parsimonious generalization of fuzzy thematic sets in taxonomies applied to the analysis of tendencies of research in data science. Information Sciences, 512, 595-615.

Frolov, D., Taran, Z., & Mirkin, B. (2021, August). Using the IAB Contents Taxonomy and Optimal Lifting for Efficient Audience Extension. In International Conference on Intelligent and Fuzzy Systems (pp. 596-603). Springer, Cham.

Keynote Speaker 2

A Big Data Implementation on Industrial Computer Platform and its Performance

Norma Alias

Universiti Teknologi Malaysia

normaalias@utm.my

The new challenges and future directions facing the use of big data and artificial intelligence (AI) in new norm research and innovation. The motivation of transition from small data to big data is to improve the prediction accuracy, enhance the capability of data-driven models, speed the training and learning performance measurement. With massive industrial data, the implementation of big data is addressed on the 7Vs features such as variety, volumes, velocity, visibility, variability, veracity and value. Education research dealing fully data-driven decision-making organizations in industries such as banking and security, media and entertainment, healthcare analysis, manufacturing industry 4.0. The investigation of the big data implementation involving the capability of mathematical model, discretization process, machine learning algorithm, numerical schemes, statistical tests and big data analytics performance. This requires a parallel computing architecture optimized for industrial big data. The impact of high performance computing (HPC) with high inter-processor communication protocol in big data simulation will be achieved beyond expectations. The HPC supported by the industrial computer platform integrated with parallel computing architecture, industry-leading performance, huge storage requirements, high speed processors, compact embedded PC with multi-core and CPU-GPU based. The performance indicators of industrial computer can to be measured in terms of run time, speedup, efficiency, effectiveness and some numerical results.

Keynote Speaker 3

Continual & Autonomous Machine Learning

Mahardika Pratama

Nanyang Technological University

Continual learning is a rapidly growing research topic where the underlying goal is to build a resource-efficient learning algorithm learning forever from data streams with the absence of old data samples. That is, it is capable of accepting the presence of new data samples, new patterns, new classes or even new tasks. The underlying challenge is not only to adapt efficiently to new environments but also to combat the catastrophic forgetting problem leading to performance loss of previously seen concepts because network parameters are over-written when learning new concepts. This talk will start from introduction of continual learning algorithms to handle various complex problems of data streams under a single task. It will continue with the case of streaming tasks (multi-task) afterward where solution of catastrophic forgetting, namely Inter-Task Synaptic Mapping (ISYANA) will be discussed followed by introduction of Knowledge Retention in Self-Adaptive Deep Continual Learner (KIERA) to cope with a sequence of different unsupervised learning tasks.

Artificial Intelligence for Future Generation Robotics

Ankush Ghosh

School of Engineering and Applied Sciences, The Neotia University, India

Artificial Intelligence has become one of the most prevalent topics in recent years. The applications of AI we see today is a tip of the iceberg. The AI revolution has just begun. It is becoming an integral part of all modern electronic devices. Application in automation areas like automotive, security and surveillance, augmented reality, smart home, retail automation and healthcare are few of them. But, Artificial Intelligence wouldn't be that popular if it hadn't found so many applications across so many sectors, mostly in the software sector. However, we are betting on hardware, since AI is becoming an integral part of almost all modern automated devices. And it is impossible to discuss the future of automation without talking about robots. New generation Robotics will play a major role in this automated world. The future generation robotics would get a relutionary change due to advancement of Artificial Intelligence and meachine learning in recent times.

AdaBoost Support Vector Machine Method for Human Activity Recognition

Intan Nurma Yulita
Research Center for Artificial Intelligence and Big Data
Universitas Padjadjaran
Bandung, Indonesia
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
intan.nurma@unpad.ac.id

Erick Paulus
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
Leiden Institute Advanced Computer Sciences
Leiden University
Leiden, Netherland
e.paulus@liacs.leidenuniv.nl

Asep Sholahuddin
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
asep.sholahuddin@unpad.ac.id

Dessy Novita
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
dessy.novita@unpad.ac.id

***Abstract*— Human activity recognition research is being implemented more and more as technology advances in computer vision. Many fields require activity recognition technology, such as theft detection or online exam cheating detection. One method that is widely used is AdaBoost. This study proposes the AdaBoost Support Vector Machine Method, a combination of the AdaBoost Method and the Support Vector Machine. The evaluation uses datasets for human activity recognition and compares them with other machine learning algorithms. The results obtained indicate that the proposed method has the highest performance compared to the tested algorithms. The highest accuracy in this study was 96.06%. It shows that SVM as an AdaBoost component is proven to be able to improve the performance of AdaBoost.**

Keywords*— human activity recognition, AdaBoost, Support Vector Machine, computer vision

I. INTRODUCTION

With the development of technology, both hardware and software, computer vision technology is also growing [1]. One of the applications of computer vision is activity recognition [2]. It has been widely used in various fields such as robotics, security, and education. In robotics, human activity recognition can produce robots that have interaction capabilities. Robots can capture the movements of humans in front of them [3]. The robot system processes the movement data so that later the robot can respond to the action, for example, replying to the greeting if a human greets the robot. Human activity recognition can be used to detect theft [4]. Closed Circuit Television (CCTV) records all human activities in the room. Of course, it takes time, and the effort to monitor human activity recognition is made manually. In this case, computer vision can automatically detect suspicious activity such as theft. In the case of education, computer vision helps see suspicious activities during exams [5]. The system records the activities of the examinees through the camera. Through this recording, the system processes and identifies participants and evidence of fraud. Implementation of this system streamlines the number of supervisors. Especially with the COVID-19 pandemic, exams are conducted online, making it even more difficult if traditionally supervised. The involvement of computer vision allows for individual supervision even though the exam involves many examinees [6].

Computer vision can identify human activity recognition using Artificial Intelligence (AI). The essence of AI research in machine learning. It is an AI approach that focuses on creating machines that can learn without being explicitly programmed. Learning is an essential part of what makes us human. If we're going to build AI that can perform tasks with human-like intelligence, then we need to build machines that can learn on their own, based on their past experiences.

Machine learning has dramatically affected the industrial world; most of the industrial world who work with large amounts of data have felt the influence of machine learning in their industry. The goal is to gain insight from the data they have, and this technology can make work more efficient or faster with new data that tends to be the same.

Several machine learning algorithms have been implemented in computer vision, such as Adaboost [7], Naïve Bayes [8], Random Forest [9], Bayesian Networks [10]. This study aims to compare the above algorithms in recognizing human movement. This study also proposes a combination of AdaBoost by using a base learner in a Support Vector Machine (SVM) to produce better modeling. SVM is used for base learners because several studies have shown that SVM has high accuracy for motion recognition [11-13]. By making SVM a base learner, it is hoped that the combination can improve the work of AdaBoost.

II. LITERATURE STUDY

A. Support Vector Machine

Vladimir Vapnik, Bernhard Boser, and Isabelle Guyon are the researchers who developed SVM. These researchers have a background in structural risk minimization (SRM) [14]. Thus SVM also expands based on SRM. At the beginning of its development, SVM was used to solve linear problems. But now, SVM has also solved non-linear issues [15]. SVM has a kernel trick that transforms data into a higher-dimensional space. A hyperplane then separates the data on this new dimension. This separation identifies data with different classes. Of course, if the hyperplane is terrible, other classes are put into the same group. This hyperplane is found using support vectors and margins. To avoid

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

misclassification of each training example, SVM uses parameter C to control it. The optimization will choose a smaller margin if the hyperplane correctly classifies all training data for large C values [16]. However, if the value of C is minimal, the optimization will choose a more significant margin even though it has managed to classify all training data correctly. Therefore, the value of C dramatically affects the performance of SVM in recognition.

B. AdaBoost

The ensemble method is a method that combines several weak learners to produce strong learners [17]. Of course, merging several models intuitively will help if the models are different. Boosting is one method of the ensemble method. Another technique of the ensemble method is Bagging. Boosting uses the same algorithm to become a weak learner. The training focused on the weight of the data that failed to be appropriately classified. Adaptive boosting (AdaBoost) is one of the most widely used boosting algorithms. The implementation of AdaBoost and its variants has solved classification problems or pattern recognition in everyday life [18]. This is because although this method is simple, it produces high accuracy. Its success is supported by the data mechanism that failed to be appropriately classified in the previous stage to be given more attention at the next step.

The main idea of AdaBoost is:

- AdaBoost has several training iterations. Weak learners will be trained repeatedly according to this number of iterations.
- Training data on weak learners based on the success in the previous stage.
- If the data failed to be classified in the previous stage, this data would be given more attention which is indicated by an increase in the weight of the data.
- Each weak learner will be given a weight according to their success in classifying. The better the classification results, the higher the importance of the weak learner.
- The AdaBoost training process ends when all data is classified correctly.
- In the final stage, determining the class of data-based voting from all weak learners. However, weak learners with higher weights will have more influence on this voting.

III. METHODOLOGY

The research methodology is shown in Fig. 1. The dataset came from 30 participants involved in the research conducted by Anguita et al. [13]. Age ranged from 19-48 years. Experiments were carried out by recording the activities of each participant in the following actions: sitting, laying, walking-down-stairs, walking-up-stairs, standing, walking. The recording was done using a gyroscope, accelerometer, and video recorder. Signals from the gyroscope and accelerometer were preprocessed with a noise filter. This filtering is to eliminate noise that appears during recording or data storage. If not stopped, the model has poor performance. Next, the fixed-sampling was performed every 0.256 seconds using an overlapping window with a size of 0.5 of the sampling size. Butterworth low-pass filters were also used in this study to separate the components of gravity and body motion. Butterworth Filter is a type of signal processing designed to produce a frequency response that is as flat as possible in the passband. At the same time, the Butterworth low pass filter is a filter that has a flat maximum frequency response (no ripple) from 0 Hz frequency to the end of the passband area, namely the cut frequency. Off which suffers from - 3dB attenuation. This preprocessing produced 563 attributes for each sample. The dataset consists of 10299. This dataset was divided into 7352 samples as training data and 2947 samples as test data.

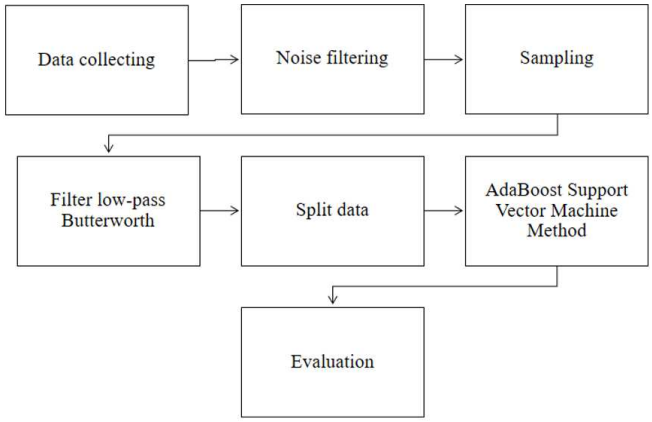


Fig. 1. Research methodology

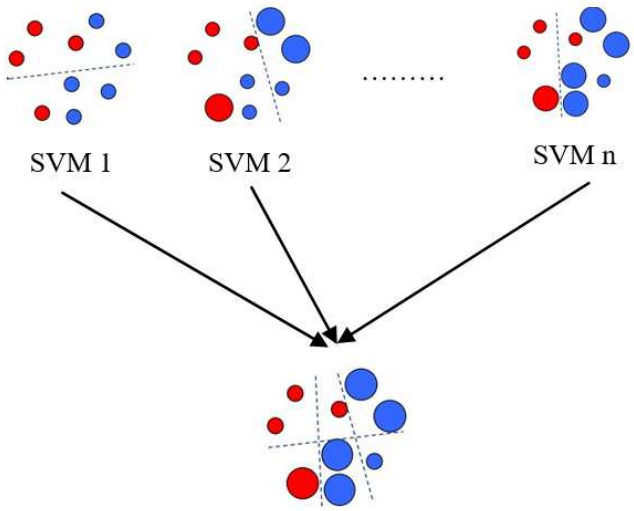


Fig. 2. AdaBoost Support Vector Machine algorithm

This study focuses on training SVM as a base learner on AdaBoost. The illustration of the combination is shown in Figure 2. There are n iterations to form the final classifier. At each iteration, SVM performs a classification. If an iteration. If data fails to be appropriately classified, then the data in the next iteration will be given a higher weight. The higher the weight of the data is intended so that the SVM in the next stage can focus on classifying this incorrect data with better modeling. This study also compared other algorithms such as Naïve Bayes, Random Forest, Bayesian Networks, Bagging, Extra Tree, K-Nearest Neighbor (KNN), and Random Tree. The evaluation was done by calculating accuracy. Accuracy is based on the amount of properly classified data divided by the whole data.

IV. RESULTS

A. Analysis of C in SVM

The implementation was done using 200 iterations on the SVM kernel in Poly Kernel. Based on Table 1, it is found that the size of C is the most optimal when the value is 0.70. If C=0.60 and 0.65, the accuracy is lower than if C=0.70. However, the value of C above 0.70 does not indicate a higher accuracy. The success of AdaBoost SVM can be assessed through the confusion matrix in Table 2. This method often fails to classify data with the label "standing" and also "Walking-Upstairs." But the primary failure when classifying data with the class "sitting." There were 56 data from this label that failed to be classified. Fifty-five of this data was classified as data with the type "standing." A, B, C, D, E, and F in Table 2 are symbols of the class of standing, sitting, laying, walking, walking downstairs, walking upstairs.

TABLE I. ANALYSIS OF C IN SVM

No.	Method	Kernel	C	Accuracy
1	AdaBoost SVM	Poly Kernel	0.60	95.86%
2	AdaBoost SVM	Poly Kernel	0.65	95.79%
3	AdaBoost SVM	Poly Kernel	0.70	96.06%
4	AdaBoost SVM	Poly Kernel	0.75	95.59%
5	AdaBoost SVM	Poly Kernel	0.80	95.96%

TABLE II. CONFUSION MATRIX

	Predicted Class					
Actual Class*	*A*	*B*	*C*	*D*	*E*	*F*
A	513	19	0	0	0	0
B	55	435	0	0	0	1
C	0	0	537	0	0	0
D	0	0	0	491	4	1
E	0	0	0	5	405	10
F	0	0	0	20	1	450

B. Analysis of The Base Learner

Based on the data in Table 1, the most optimal AdaBoost SVM was obtained when C = 0.70. This model was then compared using other algorithms as a base learner, such as Decision Tree, Decision Stump, Random Forest, Hoeffding Tree. Based on Table 3, Decision Stump had the lowest performance despite being another familiar base learner used for AdaBoost. Other tree-based algorithms in this study produced a better performance with a minimum achievement of 87.92% on the Hoeffding Tree. However, these algorithms had lower accuracy than AdaBoost SVM.

C. Comparison to Other Methods

This study also compared other machine learning methods, namely Naïve Bayes, Random Forest, Bayesian Networks, Bagging, Random Tree, Extra Tree, and KNN. Table 4 shows that the method proposed in this study had the best performance. Another best algorithm was Random Forest, but its accuracy had a difference of 3% compared to the method proposed in this study.

TABLE III. ANALYSIS OF BASE LEARNER

No.	Method	Base Learner	Accuracy
1	AdaBoost	SVM	96.06%
2	AdaBoost	Decision Tree	93.65%
3	AdaBoost	Decision Stump	35.05%
4	AdaBoost	Random Forest	93.82%
5	AdaBoost	Hoeffding Tree	87.92%

TABLE IV. ANALYSIS OF BASE LEARNER

No.	Method	Accuracy
1	AdaBoost SVM	96.06%
2	Naïve Bayes	78.96%
3	Random Forest	93.62%
4	Bayesian Networks	82.93%
5	Bagging	89.24%
6	Random Tree	78.59%
7	Extra Tree	79.91%
8	KNN (K=1)	88.40%
9	KNN (K=2)	85.65%

V. CONCLUSION

Based on the research conducted by this study, it can be seen that:

- AdaBoost SVM had the highest performance compared to Naïve Bayes, Random Forest, Bayesian Networks, Bagging, Random Tree, Extra Tree, and KNN in this study.
- SVM was the right choice as a base learner of AdaBoost. The performance of SVM as a base learner was higher than other base learners in this study.
- AdaBoost SVM produced the highest performance if C=0.70, and it made this method able to achieve an accuracy of 96.06%. It shows the proposed method can be promoted for other human activity recognition

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran, that supported the study. This work was also supported by the Center for Excellence Research Grant for Higher Education, funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, contract number: 1207/UN6.3.1/PT.00/2021.

REFERENCES

[1] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," Foundations and Trends® in Computer Graphics and Vision, 12(1–3), 1-308, 2020.

[2] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," Multimedia Tools and Applications, 79(41), 30509-30555, 2020.

[3] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita, "Perceiving the person and their interactions with the others for social robotics–a review," Pattern Recognition Letters, 118, 3-13, 2019.

[4] V. Singh, S. Singh, and P. Gupta, P, "Real-time anomaly recognition through CCTV using neural networks," Procedia Computer Science, 173, 254-263, 2020.

[5] S. Maniar, K. Sukhani, K. Shah, K., and S. Dhage, "Automated Proctoring System using Computer Vision Techniques," 2021 International Conference on System, Computation, Automation, and Networking (ICSCAN) (pp. 1-6), 2021.

[6] H. Li, M. Xu, Y. Wang, H. Wei, and H. Qu, "A Visual analytics approach to facilitate the proctoring of online exams," 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-17), 2021

[7] A. Subasi, D.H. Dammas, R. D. Alghamdi, R. A. Makawi, E. A. Albiety, T. Brahimi, T., and A. Sarirete, "Sensor-based human activity recognition using AdaBoost ensemble classifier," Procedia computer science, 140, 104-111, 2018

[8] P. Mohan, and S. Chinchu, "A Comparison of human activity Recognition [HAR] based on machine learning classifiers," International Journal of Research in Engineering, Science and Management, 4(6), 245-248, 2021

[9] S. Balli, E. A. Sağbaş, and M. Peker, M. "Human activity recognition from smartwatch sensor data using a hybrid of principal component analysis and random forest algorithm," Measurement and Control, 52(1-2), 37-45, 2019.

[10] L. Liu, S. Wang, G. Su, Z. G. Huang, and M. Liu, M, "Towards complex activity recognition using a Bayesian network-based probabilistic generative framework," Pattern Recognition, 68, 295-309, 2017.

[11] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," IEEE Transactions on Industrial Informatics, 13(6), 3070-3080, 2017.

[12] S. Abudalfa, and K. Bouchard, K, "Hybrid deep-readout echo state network and support vector machine with feature selection for human activity recognition," Big Data Technologies and Applications, pp. 150-167, 2020.

[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," International workshop on ambient assisted living, pp. 216-223, 2012.

[14] M. M. Hassan, M. Z. Uddin, A. Mohamed, A., and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," Future Generation Computer Systems, 81, 307-313, 2018.

[15] K. S. Kumar, and R. Bhavani, "Human activity recognition in egocentric video using PNN, SVM, kNN and SVM+ kNN classifiers," Cluster Computing, 22(5), 10577-10586, 2019.

[16] C. M. Patil, B. Jagadeesh, and M. N Meghana, "An approach of understanding human activity recognition and detection for video surveillance using HOG descriptor and SVM classifier," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 481-485, 2017.

[17] J. W. Kasubi, and M. D. Huchaiah, "Human activity recognition for multi-label classification in smart homes using ensemble methods," International Conference on Artificial Intelligence and Sustainable Computing, pp. 282-294, 2021.

[18] I. Akhter, A. Jalal, and K. Kim, "Pose estimation and detection for event recognition using sense-aware features and Adaboost classifier," 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), pp. 500-505, 2021

Sentiment Classification Against the Public Activity Restrictions Policy in Jakarta Using Machine Learning Models

Dennis Dennis[a1], Yudhistira Nugraha[ab2], Andi Sulasikin[a3]
Farhan Adham Irawan[a4], Clarissa Febria Finola[a5], Fajar Kurniawan[a6], Juan Intan Kanggrawan[a7], Alex L. Suherman[c8]
[a]*Jakarta Smart City, Department of Communicatiton, Informatics, and Statistics,* Jakarta, Indonesia
[b]*School of Computing, Telkom University*, Bandung, Indonesia
[c]*Directorate of Research and Community Services, Telkom University*, Bandung, Indonesia
e-mail: [1]michaeldennisw@gmail.com, [2]yudhistiranugraha@telkomuniversity.ac.id , [3]andisulasikin.jsc@gmail.com,
[4]farhanadham04@gmail.com, [5]clarissa.jsc@gmail.com, [6]fajar.kurniawan.jsc@gmail.com,
[7]juan.tan.kang@gmail.com, [8]alexsuherman@telkomuniversity.ac.id

***Abstract*—To stop the spread of the COVID-19, the Indonesian government implemented community activities restrictions enforcement (in Indonesian language: *Pemberlakuan Pembatasan Kegiatan Masyarakat* or PPKM) starting from January 2021. The term *PPKM* applied are PPKM *Mikro* (in Indonesian language) or Micro PPKM, *PPKM Darurat* (in Indonesian language) or Emergency PPKM, and *PPKM Level 1-4* or Level 1-4 PPKM. On the other hand, the existing research mostly used Twitter as the data source to do sentiment classification. Therefore, we aimed to classify social media comments on Facebook and YouTube on Level 1-4 PPKM policy in Jakarta. We used "*PPKM Jakarta"* as the keyword topic in August - September 2021 when Level 1-4 PPKM was ongoing. In addition, we compared datasets composition, machine learning models, and features extraction. Random Forest, Naive Bayes, and Logistic Regression were performed as the machine learning models due to they were the top three models on the previous research. We extracted word unigram, word bigram, character trigram, and character quadrigram as the feature extraction. The highest average F-measure was obtained with a 79.6% score of the Logistic Regression model using character quadrigram extraction. We found that comments from Facebook and YouTube were dominated by neutral sentiment (49.8%) with this setup. It means the people of Jakarta started to trust the government in handling the COVID-19 pandemic. Through word cloud analysis, it is recommended that social assistance be reviewed for those directly affected.**

Keywords—sentiment, PPKM, Facebook, YouTube, wordcloud

I. INTRODUCTION

COVID-19 cases in Indonesia were first reported in March 2020 [1]. Over time, active positive cases in Jakarta increased to 675 on April 1, 2020 [2]. To stop the spread from worsening, Jakarta became the first province in Indonesia to implement the Large-Scale Social Restrictions (in Indonesian language: *Pembatasan Sosial Berskala Besar* or PSBB) policy starting on April 10, 2020 [3]. The key to this restriction was studying, working, and praying from home [4].

Jakarta's active positive cases trend increased to 109.963 cases on December 31, 2020 [2]. In January 2021, the Indonesian government adopted the policy of *Enforcement of Community Activity Restrictions* (in Indonesia language: *Pemberlakuan Pembatasan Kegiatan Masyarakat* or PPKM). The critical difference lies in government initiation. In PSBB, the initial initiative is in the regional government (in this case, the Provincial Government of Jakarta), while in PPKM, the initiative is in the central government [5]. Hence, PPKM policy could be applied uniformly [6].

Jakarta became one of the provinces to implement the PPKM [7]. Throughout 2021, PPKM terms that have been used are Micro PPKM (February 9 to July 25, 2021), Emergency PPKM (July 3-20, 2021), and Level 1-4 PPKM (July 21, 2021 to September 2021) [7][8][9][10]. On July 16, 2021, Jakarta reached the highest active case of 113.138 cases. From the existing PPKM terms, Emergency PPKM and Level 1-4 PPKM succeeded in suppressing active cases of COVID-19 in Jakarta to 2.558 cases on September 20, 2021 [2].

Indonesia is the fourth country with the most significant social media users in the world [11]. The most used social media and video streaming applications are Facebook and YouTube, with an average time of 17 hours per month and 25.9 hours per month, respectively [12]. 140 million Indonesians used social media Facebook [13], and 132 million Indonesians used YouTube [13]. Content on social media provides not only entertainment and product information but also the latest news. One of the news that is being discussed frequently is about handling the COVID-19 pandemic.

Sentiment analysis is a research area to extract information from texts [14]. It is used to understand public responses related to companies or governments' products, services, or policies [15]. By utilizing social media activities, public points of view towards an approach can be known.

Many sentiments analysis studies have been conducted using Twitter as the data source. In [16], the sentiment analysis about COVID-19 in Indonesia was studied. The dataset was taken in April 2020 with the keyword "Indonesian Corona Virus," classified into negative, neutral, and positive sentiment. They found that most of the tweets had opposing opinions. The main concern was doubting Indonesia's ability to handle the pandemic. These finding was different from [17] where tweets were dominated by positive sentiment.

The most recent study was conducted by Pradhana [18]. He investigated sentiment analysis related to Micro PPKM. He also used Twitter as the data source, Support Vector Machine as the model, and term frequency-inverse document frequency (TF-IDF) as the feature and classified tweets into two sentiments, i.e., negative and positive sentiment. He

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

discovered that 12.8% of tweets were negative sentiment and 87.2% of tweets were positive sentiment.

Twitter tends to be easier to retrieve data because of the Application Programming Interface (API) existence. However, with the abundance of Facebook and YouTube users in Indonesia, it would be interesting to analyze these two social media sentiments. As far as we know, there is no research study about Level 1-4 PPKM policy implementation in Jakarta. Therefore, we aimed to classify the sentiment of Level 1-4 PPKM policy in Jakarta using Facebook and YouTube comments. Also, we compared several datasets composition, machine learning models, and features extraction to obtain what combination produces the best performance. We adopted the machine learning models from [19] because those were the top three best performing models in the Indonesian language tweets dataset. We also implemented features extraction as [19] did to be compared using Facebook and YouTube datasets. Moreover, we provided word cloud analysis for each sentiment to observe the most frequent word. These words could give recommendations to the government for future policy.

This paper is organized as follows. Section 2 contains the methodology explanation. Section 3 shows experiment results. Section 4 gives research discussion. Lastly, section 5 shows conclusions, recommendations, and future work.

II. METHODOLOGY

This study aimed to classify the sentiment about Level 1-4 PPKM in Jakarta using Facebook and YouTube comments as the data source. Three datasets were experimented with, i.e., the imbalanced, balanced, and balanced dataset with low document frequency (DF) removal. We compared the top three best performance machine learning models from [19], Random Forest, Naïve Bayes, and Logistic Regression. The features used in this study are word unigram, word bigram, word n-gram (the union of word unigram and bigram), character trigram, character quadrigram, and character n-gram (the union of character trigram and quadrigram). Lastly, we showed the word cloud analysis to give an evaluation to the government. Fig. 1 shows the methodology approach.

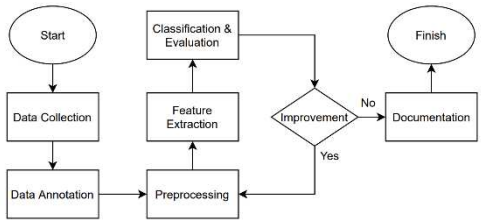


Fig. 1. Methodology Approach

A. Creating the Dataset

1) Data Collection

3.583 comments were crawled as the dataset source from August 14 to September 14, 2021 with 1.125 comments from Facebook and 2.458 comments from YouTube. "*PPKM Jakarta*" was chosen as the keyword topic. Comments from Facebook came from two posts shared by the official Jakarta government account (*Pemprov DKI Jakarta*). YouTube's comments were crawled from the top ten videos from *Kompas TV*, *CNN Indonesia*, *tvOneNews*, *Berita Satu*, and the government's channel (*Sekretariat Presiden*). The NaN and duplicate removal were applied. Therefore, from 3.583 comments, we got 3.255 comments to be trained and tested. From 3.255 comments, 2.288 comments were used for training, and 967 comments were used for testing.

R programming language was used to scrape the data. Rselenium package and Rvest package were utilized to support the data collection. Rselenium was used to collect data from social media. Rvest was used to accelerate the data collection process. Using Rselenium only will slow down the process because the text is taken one by one.

TABLE I. COMMENTS EXAMPLE FROM FACEBOOK (FB) AND YOUTUBE (YT)

Text	Source
Agustusan kita bikin lomba rame-rame lomba siapa yg bisa bertahan hidup<U+0001F923><U+0001F602>	YT
Insyaa Allah saya yang menang karena saya akan mengikuti peraturan yang ada kalo saya berarti anda kalah atau menang yaa?	YT
Lomba cara pake masker dan lomba swab paling banyak	YT
Hanya membuat masalah baru, sebaiknya tutup pintu bagi pengungsi afganistan untuk indonesia,negeri ini sedang dlm bencana covid,yg berdampak memporak porandakan seluruh perekonomian,apa untungnya menerima teroris pembawa masalah yg hanya akn membuat ki… Lihat Selengkapnya	FB
Sakit hati berjaMAAh	FB
betul kenapa nkri harus patuh dan tunduk pengungsi hajar saja mereka	FB

2) Data Annotation

2.288 comments from Facebook and YouTube were labeled manually by two people into negative (0), neutral (1), and positive sentiment (2). The optimistic view contained supportive, hopeful, pray, and constructive criticism statements. The negative message had blasphemy, unpleasant, and satire statements. The neutral sentiment included questions and ideas that cannot be polarized into positive and negative views. Hence, 2.288 comments were classified into three classes: 1.239 comments labeled as negative sentiment, 678 comments labeled as neutral sentiment, and 371 comments tagged as positive sentiment.

B. Sentiment Classification

Python was used to apply the sentiment classification. For feature extraction, CountVectorizer from Sklearn was applied.

1) Preprocessing

There were seven steps in the preprocessing section, i.e., convert to lowercase, remove links, convert (at)username to AT_USER, additional white spaces removal, replace hashtag with word (for example, *#ppkm* to *ppkm*), punctuation and stopwords removal, and change slang words. The slang words dictionary was taken from [20]. The stopwords dictionary and the code were adapted from [21] with some modifications.

NaN values and duplicate comments were removed from the dataset. Hence after the preprocessing step, we got 1.214 comments as negative sentiment, 659 comments as neutral sentiment, and 343 comments as positive sentiment. This first dataset was used as the imbalanced dataset.

The balanced dataset was obtained by implementing an under-sampling technique. A total of 340 comments from each sentiment class were chosen randomly. Therefore, we got 341 comments as negative sentiment, 340 comments as neutral sentiment, and 339 comments as positive sentiment.

The balanced dataset with low DF removal was achieved by collecting words with one occurrence and removing them. NaN values removal and re-balanced were implemented. In the end, we got 191 comments as negative sentiment, 190 comments as neutral sentiment, and 189 comments as positive sentiment.

2) Feature Extraction

Six features were used: word unigram, word bigram, word n-gram (the union of word unigram and bigram), character trigram, character quadrigram, and character n-gram (the union of character trigram and quadrigram). Word bigram was chosen to support the word cloud analysis that should be considered the policy's evaluation.

3) Classification and Evaluation

Random Forest, Naive Bayes, and Logistic Regression were used to classify the sentiment because in [19], they were the top three models with the best performance. The maximum iteration (*max_iter*) parameter in Logistic Regression was adjusted into 500 counts to converge. The model was evaluated using the 10-fold cross-validation method due to the limited amount of data. The weighted average F-measure was used to be the indicator of model performance because we were interested in checking how far the model could classify each text on every sentiment.

III. Results

We organized three experiments. First, we compared machine learning models in three different dataset compositions (i.e., imbalanced dataset, balanced dataset, and balanced dataset with low DF removal) to determine the best performance of models and improvement strategy. Second, we compared the performance of feature selection using the best algorithm performance from the previous result. Third, we created word clouds based on text with the corresponding sentiment.

At first, we compared the performance of the model in three datasets. The weighted F-measure score was achieved by averaging all the features used in this research. Table 2 shows the experiment result. The weighted average F-measure increased significantly by 17.6%-point and 4.4%-point based on the average comparison for each dataset. Looking into the third dataset, the Logistic Regression performed better than the two others, with a 76.3% score. It suggests that we may use the Logistic Regression model in balanced dataset with low DF removal to train the model.

TABLE II. The Weighted Average F-measure for Each Model

Model	Imbalanced Dataset (%)	Balanced Dataset (%)	Balanced Dataset with Low DF Removal (%)
Random Forest	51.7	70.8	75.7
Naïve Bayes	53.7	68.8	73.0
Logistic Regression	53.7	72.3	76.3
Average	53.0	70.6	75.0

Second, the comparison between feature extraction was conducted. The weighted average F-measure score was achieved using the Logistic Regression model in the balanced dataset with low DF removal. Table 3 compares all feature extraction. The highest score was obtained by character quadrigram with a 79.6% score, while the word bigram performed poorly. The character n-gram outperformed the word n-gram with a 0.3%-point difference. It suggests that we may apply character quadrigram for sentiment classification in the test set.

TABLE III. The Weighted Average F-measure for Each Feature Selection with Logistic Regression Model in Balanced Dataset with Low DF Removal

Feature Extraction	Weighted Average F-measure (%)
Word Unigram	77.1
Word Bigram	64.8
Word n-gram	79.0
Character Trigram	78.2
Character Quadrigram	79.6
Character n-gram	79.3

With the Logistic Regression model and character quadrigram feature, each comment in the test set was classified into three classes. From 967 comments in the test set, we found that 307 comments tagged as a negative sentiment (31.7%), 482 comments tagged as neutral sentiment (49.8%), and 178 comments tagged as positive sentiment (18.4%). It shows that comments were dominated by neutral sentiment.

Fig. 2 Word Cloud of Negative Sentiment

Third, word cloud analysis was conducted to gain text information for each sentiment. Each comment for each opinion was collected. Then, the word cloud was generated using the comments. Also, we sorted word bigram from the most occurrences to support the word cloud analysis. Fig. 2 shows the word cloud of negative sentiment. The most famous words highlighted are 'rakyat' (people) and 'ppkm' (the social restriction policy). This finding confirmed the critical topic of the research area. We spotted two exciting words like 'makan' (eat) and 'mati' (death). From bigram, we found the word 'mati kelaparan' (starve to death) with 14 appearances. It interprets that the people of Jakarta need to meet basic needs like food to deal with the COVID-19 pandemic.

Fig. 3 Word Cloud of Neutral Sentiment

Fig. 3 shows the word cloud of neutral sentiment. Beyond the word from the previous word cloud, two highlighted words are 'vaksin' (vaccine) and 'perpanjang' (extend). Interestingly, we found the word 'respon cepat' (fast respond) with seven appearances from the frequent bigrams. It interprets that the people of Jakarta are looking for vaccine information and news update related to COVID-19.

Fig. 4. Word Cloud of Positive Sentiment

Fig. 4 shows the word cloud of positive sentiment. The highlighted words are 'semoga' (hopefully), 'aamiin' (amin), and 'alhamdulillah' (Arabic phrase meaning praise to God). Overall, the intention of these words contains positivity, i.e., pray, hope, and support. From the frequent bigrams, we found the word 'terima kasih' (thank you) with 11 appearances and 'kerja nyata' (real work) with four appearances. It interprets that the people of Jakarta keep supporting the government by praying and hoping for the best in handling the pandemic.

IV. Discussion

This study aimed to compare datasets composition, machine learning models and features extraction to get the best combination. From the combination, we implemented it to classify the Facebook and YouTube comments. In addition, we analyzed the word cloud for each sentiment to evaluate the policy.

First, we will discuss the comparison between each dataset. Based on Table II, there has been a 17.6%-point increase in average F-measure from the imbalanced to balanced dataset. It shows that an unbalanced dataset leads to poor classification performance. The significant difference between the majority and minority class will make the model performs better in the majority than the minority class [22]. Therefore, balancing the dataset by using an under-sampling method is recommended to improve the model performance. We suggest experimenting with an over-sampling approach to balance the dataset.

From balanced dataset to balanced dataset with low DF removal, there has been a 4.4%-point increasing average F-measure. From 3.134 words in total, 2.122 words (67.71%) only appear once. These words are non-informative; hence, removing these words will reduce the dimensionality of the feature space [23]. The preliminary researchers found that this method is the most straightforward technique for vocabulary reduction with the lowest cost computation. Therefore, we suggest removing the inferior DF technique to improve model performance.

Second, we would analyze the comparison of machine learning models and features extraction. Table II shows that the Logistic Regression performs better than Random Forest and Naive Bayes. It happens because our dataset still has lots of noise and informal words even after the preprocessing step. For example, *iyain deh* can be changed to *ya, trmksh* can be changed to *terima kasih*, and *smoga* can be changed to *semoga*. When there are noise variables, Logistic Regression performs better than Random Forest [24]. On the other hand, Naive Bayes assumes that each word in a sentence is independent between different words and between multiple occurrences of the same word [25]. This assumption does not apply to real text data problems because each sentence is always related. Therefore, we suggest using Logistic Regression if there are lots of noisy words.

Table III shows that the character quadrigram is the best feature for the model. The advantages of these features are language independence and capturing sub-words, which is convenient in a noisy text dataset [26]. This result is different from a hate speech detection study in which the word n-gram outperformed the character n-gram [19]. Therefore, further research is needed to study the implementation of the character n-gram in the Indonesian language.

Third, we will show the classification result and word cloud analysis. The classification result points out that neutral sentiment dominates the comments in social media Facebook and YouTube. This finding is different from the previous research [16], which was dominated by the opposing view. It means there is a tendency to trust and support the government in dealing with the COVID-19 pandemic by applying Level 1-4 PPKM.

From the negative sentiment word cloud, one of the highlighted topics is the fulfillment of basic needs. The social restriction implementation impacts the difficulty of earning income so that basic food needs are challenging to meet. Therefore, we recommend the government gives attention to fundamental social assistance 'bantuan sosial' for those directly affected.

In the neutral sentiment word cloud, we highlighted the word 'vaksin' (vaccine). It points out that lots of people discuss vaccinations. This finding needs to be responded to by spreading the information about the COVID-19 vaccine and expanding vaccination reach that leads to herd immunity.

The positive sentiment contains lots of supporting and hopeful words. The positive comments achieve the least one (18%) in classification results. This finding shows that the people of Jakarta have started to support the government policy. One of the exciting bigrams is 'kerja nyata' (real work). It shows appreciation from the public that the government has made a great effort to control the COVID-19 pandemic. This effort needs to be maintained and improved in the future.

V. Conclusions

In summary, the sentiment toward the Level 1-4 PPKM policy was dominated by the neutral view, which means that the people of Jakarta started to trust the government in dealing with the COVID-19 pandemic. According to word cloud analysis, the government needs to execute and re-examine the social assistance program for those directly impacted. We have shown that implementing a balanced dataset with low DF removal gives the best performance. Logistic regression is a valuable model to deal with noisy text classification. The character n-gram is suitable to be used as feature selection in the Indonesian language. However, further studies are needed to add more insight to dealing with Bahasa Indonesia's noisy text from social media comments.

REFERENCES

[1] R. N. Velarosdela, "Kilas Balik Kronologi Munculnya Kasus Pertama Covid-19 di Indonesia," 2021. https://megapolitan.kompas.com/read/2021/03/02/05300081/kilas-balik-kronologi-munculnya-kasus-pertama-covid-19-di-indonesia?page=all (accessed Sep. 26, 2021).

[2] Jakarta Smart City, "Jakarta Tanggap Covid-19," 2020. https://corona.jakarta.go.id/id (accessed Sep. 26, 2021).

[3] R. N. Hakim, "Disetujui Menkes, PSBB DKI Jakarta Mulai Berlaku Selasa, 7 April 2020," 2020. https://nasional.kompas.com/read/2020/04/07/11582841/disetujui-menkes-psbb-dki-jakarta-mulai-berlaku-selasa-7-april-2020?page=all (accessed Sep. 26, 2021).

[4] Pemerintahan Pusat, "PP REPUBLIK INDONESIA NOMOR 21 TAHUN 2O2O," 2020. [Online]. Available: https://jdih.setkab.go.id/PUUdoc/176085/PP_Nomor_21_Tahun_2020.pdf.

[5] I. Safutra, "PSBB Diganti PPKM, Prioritas di Ibu Kota Tujuh Provinsi," 2021. https://www.jawapos.com/nasional/08/01/2021/psbb-diganti-ppkm-prioritas-di-ibu-kota-tujuh-provinsi/ (accessed Sep. 26, 2021).

[6] T. F. Arbar, "Sudah Ada PPKM, Luhut: Kita tidak akan mau PSBB lagi!," 2021. https://www.cnbcindonesia.com/news/20210204200314-4-221180/sudah-ada-ppkm-luhut-kita-tidak-akan-mau-psbb-lagi (accessed Sep. 26, 2021).

[7] CNN Indonesia, "Daftar 7 Provinsi dan 73 Kabupaten/Kota yang Berlakukan PPKM," 2021. https://www.cnnindonesia.com/nasional/20210111150933-20-592217/daftar-7-provinsi-dan-73-kabupaten-kota-yang-berlakukan-ppkm (accessed Sep. 26, 2021).

[8] Sekretariat Kabinet Republik Indonesia, "Mendagri Keluarkan Instruksi Mengenai PPKM Mikro dan Pembentukan Posko COVID-19 Tingkat Desa & Kelurahan," 2021. https://setkab.go.id/mendagri-keluarkan-instruksi-mengenai-ppkm-mikro-dan-pembentukan-posko-covid-19-tingkat-desa-kelurahan/ (accessed Sep. 26, 2021).

[9] Sekretariat Kabinet Republik Indonesia, "Mendagri Terbitkan Instruksi tentang PPKM Darurat Jawa-Bali," 2021. https://setkab.go.id/mendagri-terbitkan-instruksi-tentang-ppkm-darurat-jawa-bali/ (accessed Sep. 26, 2021).

[10] M. C. Rosa, "PPKM Level 4 Resmi Diperpanjang Hingga 2 Agustus 2021, Ini Aturan dan Wilayahnya," 2021. https://www.kompas.com/tren/read/2021/07/25/215500665/ppkm-level-4-resmi-diperpanjang-hingga-2-agustus-2021-ini-aturan-dan (accessed Sep. 26, 2021).

[11] Statista Research Department, "Leading countries based on Facebook audience size as of July 2021," 2021. [Online]. Available: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

[12] S. Kemp, "Digital 2021: Indonesia," 2021.

[13] Greenhouse Team, "Indonesia's Social Media Landscape: An Overview," 2019. https://greenhouse.co/blog/indonesias-social-media-landscape-an-overview/ (accessed Sep. 26, 2021).

[14] F. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment Analysis in Social Networks*. Elsevier Science, 20016.

[15] A. R. Naradhipa and A. Purwarianti, "Sentiment classification for Indonesian message in social media," *Proc. - Int. Conf. Cloud Comput. Soc. Netw. 2012 Cloud Comput. Soc. Netw. Smart Product. Soc. ICCCSN 2012*, no. July, pp. 2–5, 2012, doi: 10.1109/ICCCSN.2012.6215730.

[16] T. N. Simanjuntak and S. Pramana, "Sentiment Analysis on Overseas Tweets on the Impact of COVID-19 in Indonesia," *Indones. J. Stat. Its Appl.*, vol. 5, no. 2, pp. 304–313, 2021, doi: 10.29244/ijsa.v5i2p304-313.

[17] T. Krisdiyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 7, no. 1, pp. 32–37, 2021, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/coreit/article/view/12945.

[18] R. M. Pradhana, "Analisis Sentimen Publik Terhadap Kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat Skala Mikro Menggunakan Algoritma Support Vector Machine Studi Kasus Twitter," 2021.

[19] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, no. October, pp. 233–237, 2018, doi: 10.1109/ICACSIS.2017.8355039.

[20] R. Prakoso, "Analisis Sentimen KBBA," 2017. https://github.com/ramaprakoso/analisis-sentimen/blob/master/kamus/kbba.txt (accessed Sep. 15, 2021).

[21] M. N. Y. Utomo, "Python-sentianalysis-ID," 2018. https://github.com/yasirutomo/python-sentianalysis-id (accessed Sep. 15, 2021).

[22] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 4, pp. 42–47, 2012, [Online]. Available: http://www.ijetae.com/files/Volume2Issue4/IJETAE_0412_07.pdf.

[23] Y. Yiming and P. Jan O., "A Comparative Study on Feature Selection in Text Categorization," *Proceeding ICML '97 Proc. Fourteenth Int. Conf. Mach. Learn.*, vol. 53, no. 9, pp. 412–420, 1997.

[24] K. Kirasich, T. Smith, and B. Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," *Data Sci. Rev.*, vol. 1, no. 3, p. 9, 2018, [Online]. Available: https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/9

[25] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML*, 1998, doi: https://doi.org/10.1007/BFb0026666.

[26] A. Sureka and P. Jalote, "Detecting duplicate bug report using character N-gram-based features," *Proc. - Asia-Pacific Softw. Eng. Conf. APSEC*, pp. 366–374, 2010, doi: 10.1109/APSEC.2010.49.

Automatic Text Summarization with Categorization on Online News About Indonesian Public Figures Using Fuzzy Logic Method

Sigi Kemahduta
Informatics Department
Sebelas Maret University
Surakarta, Indonesia
sigaeasu@student.uns.ac.id

Sari Widya Sihwi
Informatics Department
Sebelas Maret University
Surakarta, Indonesia
sariwidya@staff.uns.ac.id

Wisnu Widiarto
Informatics Department
Sebelas Maret University
Surakarta, Indonesia
bethoro_wisnu@staff.uns.ac.id

***Abstract* —Mass media has a main purpose to provide actual information and lifestyle in society. For that purpose, mass media has been developed into many forms with online news as one of them. With so many news articles on the internet, it's been difficult for readers to get information about an event or a person in a short time and efficiently. Text summarization is one solution to save time and improve efficiency by shortening the document into a shorter version but containing important information from the source document. This research's goal is to create an automatic text summarization in several news articles in mass media that have been categorized regarding Indonesian public figures using Fuzzy Logic approach. Automatic text summarization can be used in a variety of news articles which are structured texts with neat grammar to help users find the desired information from the many news articles that are available. The result showed automatic text summarization with categorization could provide summarized news articles that are easy to understand, contain the main information of the article, grouped into matched categories and public figure, and worked efficiency.**

Keywords—automatic text summarization, categorization, Fuzzy Logic, Indonesian public figures, online news

I. INTRODUCTION

The development of mass media both in paper and electronic is developing rapidly and is increasingly advanced accompanied by the demands of the times and advances in science and technology. The mass media has a role in providing information about the events of society and also the development of people's lifestyles [1]. One of the purposes and functions of mass media is to realize the needs of information needs through the media both through paper and electronic, to public in an orderly manner [2]. As one of the mass media that is currently growing rapidly, electronic media provides a lot of information about events or subjects that are spread on the internet.

Information on the internet is increasingly developed from various media sources, makes it increasingly difficult for humans to follow all of the information. News articles are a form of information about an event that is most widely spread and read by users [3]. There are several online platforms in Indonesia such as Detik, Kompas, and CNN Indonesia. The online platform provides various types of news about events both domestic and foreign. With various news articles provided, there are difficulties in getting information quickly and efficiently.

Text Summarization is a process of shortening a document into a shorter version but containing important information from the source document to save time and improve efficiency. One method in text summarization is extraction which works by giving a score for each sentence which depicts the significant meaning[4], selecting several portions of text from the source document b, and then combining them into shorter texts [3].

Automatic text summarization means an automatic summary that is created when input is given in the form of a structured document that goes through several processes to finally produce a summary of the document using approximately 20% of the sentence from the document [5]. Automatic text summarization can be used in a variety of news articles which are structured texts with neat grammar to assist users in finding the desired information efficiently and quickly from many news articles that are available. Even so, summarization will be difficult if the articles do not have the same topic and can produce an unrelated summary so that categorization is needed in these documents.

With articles that have been categorized, summarization can more easily provide the information that the user wants. Text categorization is used in sorting and grouping documents with the same knowledge into new categories to facilitate the data retrieval function. In classifying, documents are categorized by counting how often words that are categorized as parameters appear in a document [6]. It uses broad terms and narrow terms related to the specified topic to increase accuracy in document identification [7].

This research proposes automatic text summarization in several news articles in a mass media that has been categorized regarding Indonesian public figures using the Fuzzy Logic approach. Fuzzy Logic provides a relatively simple way to draw conclusions from unclear, ambiguous or inaccurate information. In another sense, fuzzy logic can resemble human decision making with its ability to work and find the right solution, unlike classical logic which requires a deep understanding of a system, the right equation, and the right numerical value [8]. The Fuzzy Logic method used in this study is Fuzzy Sugeno because it has an output value that

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

can be a constant or a linear equation (Putri & Effendi, 2017). The news articles used in this study were sourced from CNN Indonesia because they have a simple page format and have a neat label arrangement on each news article [9].

Other research in Indonesian text summarization is [10] which utilize four elements, which are title feature, sentence length, sentence position and keyword frequency to know the importance of a sentence. Research [11] uses Mamdani Fuzzy by considering word with capital letter, and also numerical data. In this research, features used as input in text summarization are Title Words, Sentences Length, Sentence to Sentence Similarity**,** Numerical Data, Thematic Words, Term Weight, Proper Nouns.

The purpose of the research is to summarize the various existing news articles to make it easier for readers to obtain the desired information about Indonesian community leaders. It is hoped that this research can help the public to obtain information quickly and efficiently about figures of Indonesian community leaders.

II. METHODOLOGY

The research will be carried out based on the research design as shown in Fig. 1 below:

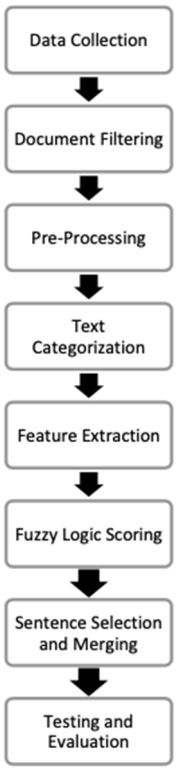


Fig. 1. Research Methodology

A. Data Collection

Data collection will search some news articles about Indonesian's public figure sourced from one of the electronic media namely the website https://www.cnnindonesia.com/. Data collection is done by web crawling method using BeautifulSoup4 in order to collect various articles from related websites automatically and then be processed at a later stage. The articles that will be used as research material are 50 news article documents related to Indonesian public figures.

B. Document filtering

The document filtering process is a selection process on documents in the form of news articles that aims to ensure that documents that are not related to problem boundaries are not processed so they can use time efficiently.

The indicators that will be used are 3 news categories on CNNIndonesia namely Political News, Legal & Criminal News and Financial Economics. In addition to indicators based on news article categories, the document filtering process uses the name of president of Indonesia and his five ministers related to the "Kabinet Indonesia Maju as indicators as follows:

- Joko Widodo (President of Indonesia 2019-2024)
- Prabowo Subianto (Minister of Defense)
- Mohammad Mahfud (Minister of Political, Legal and Security Affairs)
- Tito Karnavian (Minister of Home Affairs)
- Nadiem Makarim (Minister of Education)
- Erick Thohir (Minister of State Owned Enterprises)

C. Pre-Processing

Pre-Processing is the initial stage to prepare the to be more structured. The purpose of the pre- processing process is that the extraction results in text weighting can be maximized. At this stage, the application will make a selection of data that will be processed on the document. The pre-processing process includes 4 steps to prepare and select data, namely:

- Case Folding: The process of converting the entire text in a document into a standard form.
- Filtering: The process of removing words that have no meaning.
- Stemming: The process of dividing documents into sentences by using the punctuation mark "."
- Segmentation: The process of removing affixes in a word so that the word returns to its basic form.

D. Text Categorization

Text categorization will sort and group documents with the same knowledge into predetermined categories. The grouping of documents will use the term frequency feature which is the weighting of the value of a document by observing the ratio of occurrence of words that have been determined in a document as follows as formula (1). The grouping of news articles uses 25 keywords for each category obtained from the crawling of 200 news articles in each category: Politics, Criminal Law and Financial Economics on CNN Indonesia. Keywords are determined based on the number of occurrences of the word using the FreqDist feature in the NLTK library for 200 documents in each category and checked again manually.

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ words\ in\ document\ j} \quad (1)$$

E. Features Extraction

Sentences need to be represented as vectors in order to use a statistical method. Features extraction are attributes that

attempt to represent the data used for their task. We use seven features for each sentence given a value between '0' and '1'. Therefore, we can extract the appropriate number of sentences according to 20% of compression rate. There are seven features as follows:

1) Title Words: The score of this feature is the ratio of the number of words in the sentence that occur in the title over the number of words in the title.

$$F1 = \frac{No\ Title\ word\ in\ Sentence}{No\ Word\ in\ Title} \quad (2)$$

2) Sentences Length: The number of words in a sentence. This feature is useful to filter out short sentences such as datelines and author names commonly found in news articles. We use the calculated length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

$$F2 = \frac{No\ Word\ occuring\ in\ S_i}{No\ Word\ occuring\ in\ longest\ sentence} \quad (3)$$

3) Sentence to Sentence Similarity: Similarity between sentences. For each sentence s, the similarity between s and each other sentence is computed by the cosine similarity measure. The score of this feature for a sentence s is obtained by computing the ratio of the summary of sentence similarity of sentence s with each other sentence over the maximum of summary.

$$F3 = \frac{Sum\ of\ Sentence\ Similarity\ S_i}{Max(Sum\ of\ Sentence\ Similarity)} \quad (4)$$

4) Numerical Data: The number of numerical data in a sentence. A sentence that contains numerical data is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data in a sentence over the sentence length.

$$F4(S_i) = \frac{No\ Numerical\ Data\ in\ S_i}{Length\ (S_i)} \quad (5)$$

5) Thematic Words: The number of thematic words in a sentence. This feature is important because terms that occur frequently in a document are probably related to the topic. The number of thematic words indicates the words with maximum possible relativity. We used the top 10 most frequent content words for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words in the sentence over the maximum summary of thematic words in the sentence.

$$F5(S_i) = \frac{No\ Thematic\ Words\ in\ S_i}{Max\ (No\ Thematic\ Words)} \quad (6)$$

6) Term Weight: The average of the TF-ISF (Term Frequency - Inverse Sentence Frequency). The frequency of term occurrences within a document has often been used for calculating the importance of a sentence.

$$F6 = \frac{\sqrt{TF-ISF_i}}{Max\ (TF-ISF_i)} \quad (7)$$

7) Proper Nouns: The number of proper nouns in a sentence. Usually, the sentence that contains more proper nouns is an important one and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of proper nouns in sentence over the sentence length.

$$F7 = \frac{No.Proper\ Nouns\ in\ S_i}{Length\ S_i} \quad (8)$$

F. Fuzzy Logic Scoring, Sentence Selection and Merging

Fuzzy logic is one of the methods that can be used for text summarization. Fuzzy logic is designed to represent a form of knowledge as one of the developments of fuzzy set theory. Fuzzy logic has a difference with fuzzy in general that controls into the form of yes or no (1,0), fuzzy logic has more specific values such as true, more or less true, rather false, etc (1, 0.7, 0.5, 0.1, 0). Fuzzy logic translates strategies from human experience into formulations that can be accepted by computers. The main purpose of this logic control system is to apply the knowledge that can produce high performance in safe and reliable operations [8]. The membership function used in the research is Triangular consisting of 3 main components, which are: fuzzification, inference mechanism, and defuzzification.

1) Fuzzification: Fuzzification is the process of changing the input given into fuzzy sets that play a role in measuring information on a predetermined rule based so that it can be used by fuzzy systems. The results from features extraction are used as the input for fuzzy inference system after generalized using Triangular membership function in fuzzification (Fig. 2). The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low (L) and very low (VL), median (M) and significant values high (H) and very high (VH).

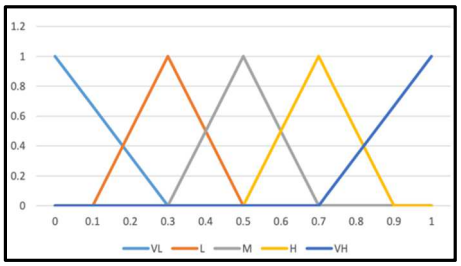


Fig. 2. *Triangular Membership Function Used in The* Research

2) Inference Mechanism: It emulates the expert's decision making in interpreting and applying knowledge about how best to control the plant. In the inference mechanism, rules If-Then is developed to control the process.

3) Defuzzification: Defuzzification is the conversion of a fuzzy quantity to a precise quantity(x). Each member of the set is given a quantity value as the initial weight, namely Important is 3, Average is 2 and Unimportant is 1. The method used is Weight Average (formula 9), which is

calculating the weighted average using to define the importance of a sentence.

$$WA = \frac{\sum_{i=1}^{n} a_i z_i}{\sum_{i=1}^{n} a_i} \quad (9)$$

G. Sentence Selection and Merging

Sentences in the Important category will be considered important and included in the summary. Sentences in the Average group will be included in the summary if all Important sentences have been included in the summary or there are no Important sentences. Meanwhile, sentences in the Unimportant category will be removed because they are deemed not to contain important information in the document. If the final result does not contain sentences with Important or Average groups or the number of sentences is less than 1, then the document is considered not to have summary results.

TABLE I. MEMBERSHIP OUTPUT

Unimportant	*Average*	*Important*
$x \leq 1.0$	$1.0 < x < 2.0$	$2.0 \leq x$

H. Testing and Evaluation

After the training process, 50 news articles sourced from CNNIndonesia were used as testing on the application. The test was carried out by 2 Indonesian language experts from Sebelas Maret University. Experts assessed applications based on three categories namely Suitability, Readability, and Utility. In addition to testing and evaluating applications based on the three predetermined categories

III. EVALUATION AND RESULTS

This research used 50 news articles from https://www.cnnindonesia.com/ as data sets to discover the information about current Indonesia' public figures that related to "Kabinet Indonesia Maju". One sample of the documents can be seen in Fig. 3. All documents collected then go through a filtering and pre-processing process. The results are then processed for text categorization and text summarization.

The Text Categorization process sorts and groups documents with the same knowledge into predetermined categories using the Term Frequency feature. There are three types of grouping used as the categorization result, which are Politics category, Law & Criminal category, and Economy category. The category with the highest score will be the result of categorization. Table 2 is the categorization result of the sample document (Fig. 3). Since Politics has the highest term frequency result, the document is grouped as Politics category.

After a document is categorized, then it will be processed to produce a summary. Each document is split into sentences by separators. As a sample, Fig. 3 has 19 sentences as the split output. All of the sentences will be the input of the Feature Extraction Process.

The Feature Extraction Process generates sentence values for use as input to the Fuzzy Logic method. There are 7 features in each sentence with a value between 0 to 1. The features values of each sentence are obtained from the calculation result of formulas 2, 3, 4, 5, 6, 7, and 8. As an example, table 3 representing sentences 1 and 12, shows how F1 until F2 resulted. The result of overall feature extractions of all sentences in Fig. 3 is shown in table. 4.

Based on the Triangular membership function that has been formed (Fig. 2), the results of the Feature Extraction process will be used as input for the next process, namely the Inference Mechanism. Therefore, the results of the Feature Extraction process in each sentence will be converted into a membership function so that it can be processed. The results of features extraction will be used as input to the fuzzy inference machine that has been defined in the previous process (Fig. 4 After the training process, 50 news articles sourced from CNNIndonesia were used as testing on the application. The test will be carried out by 2 Indonesian language experts from Sebelas Maret University. Experts will assess applications based on 3 categories namely Suitability, Readability, and Utility. In addition to testing and evaluating applications based on the three predetermined categories, suggestions were also asked to these experts for future research improvements. After the training process, 50 news articles sourced from CNNIndonesia were used as testing on the application. The test will be carried out by 2 Indonesian language experts from Sebelas Maret University. Experts will assess applications based on 3 categories namely Suitability, Readability, and Utility. In addition to testing and evaluating applications based on the three predetermined categories, suggestions were also asked to these experts for future research improvements.). It aims to determine the membership output of each sentence, which are unimportant, averages, and important.

The composition of the rules obtained from the fuzzification process will be adjusted to the IF-THEN rules that have been set to obtain results in the form of conclusions, namely Important, Average and Unimportant. The implication function in the form of a MIN (minimum) function is also applied to the value contained in the composition of the rules to take the minimum membership level of the input variable as the output. Fig. 5 is one example of some rules used in the research.

The Defuzzification process plays a role in converting the quantity results from Fuzzy into actual quantity results using the output membership function by converting three members of the set, namely Unimportant, Average and Important to the form of firm numbers, by calculating the Weight Average. Then, the result of the weight average is used to define the importance of a sentence by evaluating it with table 1. Table 6 shows how this process happened.

After the training process, 50 news articles sourced from CNNIndonesia were used as testing on the application. The test will be carried out by 2 Indonesian language experts from Sebelas Maret University. Experts will assess applications based on three categories namely Suitability, Readability, and Utility. In addition to testing and evaluating applications based on the three predetermined categories, suggestions were also asked to these experts for future research improvements.

The result of Fuzzy Inference Machine as seen in table 3 is used to determine which sentences are used in the summary. From 19 sentences of the origin document, we are able to generate four sentences as the summary result. It can be seen in Fig. 5.

Judul : Jokowi Minta Siswa SMA Taruna Nusantara Berani Lawan Hoax
Bogor, CNN Indonesia Presiden RI Joko Widodo (Jokowi) meminta para siswa SMA Taruna Nusantara berani melawan fitnah dan hoaks alias kabar bohong yang beredar di media sosial. Jokowi mengatakan para siswa jangan diam ketika mengetahui ada hoaks yang disebarkan kepada masyarakat. "Saya minta semuanya para siswa harus berani meluruskan, berani merespons kalau ada kabar fitnah, kabar bohong, hoaks, yang banyak beredar di media sosial," kata Jokowi di Ruang Garuda, Istana Kepresidenan Bogor, Kota Bogor, Jawa Barat, Senin (4/3). Jokowi menyampaikan pengarahan kepada 366 siswa kelas 11 SMA Taruna Nusantara yang berasal dari seluruh provinsi di Indonesia. Selain itu hadir pula 20 orang pamong. Para siswa tersebut kompak mengenakan seragam warna biru. "Berani meluruskan yang benar katakan benar, yang salah katakan salah. Jangan dibalik-balik," ujar Jokowi menambahkan. Jokowi tak menyebut contoh fitnah dan hoaks yang kerap muncul di media sosial. Namun, umum diketahui bahwa dalam setiap kesempatan Jokowi selalu meminta masyarakat melawan fitnah dan hoaks terkait keterlibatan PKI, kriminalisasi ulama, maupun antek asing. Kepada para siswa SMA Taruna Nusantara, Jokowi melanjutkan agar urusan politik tak membuat masyarakat terpecah belah. Ia pun menegaskan pemilihan umum, baik pemilihan bupat, wali kota, gubernur maupun presiden kontestasi politik yang rutin dilakukan lima tahun sekali. Jokowi menyebut gesekan di tengah masyarakat biasanya dimulai gara-gara politik elektoral yang digelar lima tahun sekali tersebut. Padahal, kata mantan wali kota Solo itu, pesta demokrasi lima tahunan itu seharusnya disambut dengan gembira. "Sebagai anak bangsa sangat rugi besar kita, gara-gara urusan politik antar-teman enggak saling ngomong. Harus dipakai sebagai pendewasaan politik, pematangan politik kita bagaimana memilih seorang pemimpin," ujarnya. Jokowi pun mengingatkan perbedaan yang di tengah masyarakat merupakan anugerah Tuhan. Mantan gubernur DKI Jakarta itu tak ingin perbedaan suku, agama, adat, budaya, bahasa daerah justru membuat masyarakat terbelah. "Jangan sampai karena perbedaan ini kita menjadi tidak seperti saudara, saudara se-Tanah Air," kata pria yang ikut dalam Pilpres 2019 sebagai capres nomor urut 01 tersebut.

Fig. 3. A sample of document that will be processed

TABLE II. THE CATEGORIZATION RESULT

	Category		
	Politics	Law & Criminal	Economy
Keywords Number	13	0	1
Term Frequency	0,040625	0	0,003125

TABLE III. F1-F2 OF SENTENCES 1 AND 12

Doc. Title:	**Jokowi Minta Siswa SMA Taruna Nusantara Berani Lawan Hoax**			
Num of Words in Title	**9**			
	Sentences(Si)			
i = 1	Bogor, CNN Indonesia Presiden RI Joko Widodo (Jokowi) meminta para siswa SMA Taruna Nusantara berani melawan fitnah dan hoaks alias kabar bohong yang beredar di media sosial.			
F1	**No.Title Words in Si and title**	**Words in Si and title**		**Result**
	8	(jokowi)', 'meminta', 'siswa', 'sma', 'taruna', 'nusantara', 'berani', 'melawan'		0,888888889
F2	**No. Words Occuring in Si**	**No Words of Longest Si**		**Result**
	27	37		0,72972973
F3	**Sum of Sentence Similarity Si**	**Max(Sum of Sentnc Similarity Si)**		**Result**
	2,327390844	2,705594258		0,86021429
F4	**No Numerical Data in Si**	**Numeric Words**	**Length(Si)**	**Result**
	0	-	27	0
F5	**No Thematic Words in Si**	**Thematic Words**	**Max(Num Thematic Words)**	**Result**
	7	minta', 'siswa', 'lawan', 'para', 'hoaks', 'berani', 'fitnah'	7	1
F6	**Max (TF – ISF)**	**TF - ISF**		**Result**
	4,859401736	4,740402844		0,975511617
F7	**No Proper Nouns in Si**	**Proper nouns in Si**	**Length(Si)**	**Result**
	12	Bogor,', 'CNN', 'Indonesia', 'Presiden', 'RI', 'Joko', 'Widodo', '(Jokowi)', 'SMA', 'Taruna', 'Nusantara', 'sosial.'	27	0,444444444

TABLE IV. INPUT VARIABLES

Sentence	F1	F2	F3	F4	F5	F6	F7
1	0.888889	0.72973	0.860214	0	1	0.975512	0.444444
2	0.222222	0.378378	1	0	0.571429	0.601952	0.142857
3	0.555556	1	0.795163	0.052632	1	1	0.297297
4	0.555556	0.486486	0.55229	0.111111	0.142857	0.757452	0.277778
5	0	0.189189	0.099274	0.142857	0	0.459571	0.142857
6	0.111111	0.216216	0.289056	0	0.142857	0.537072	0.125
7	0.111111	0.27027	0.262782	0	0	0.431036	0.1
8	0.111111	0.135135	0.544269	0.166667	0	0.411751	0.4
9	0.111111	0.351351	0.733646	0	0.285714	0.638317	0.153846
10	0.333333	0.621622	0.598272	0.043478	0.714286	0.883822	0.173913
11	0.555556	0.432432	0.909674	0	0.571429	0.756549	0.3125
12	0	0.567568	0.401376	0.047619	0.285714	0.850204	0.047619
13	0.111111	0.459459	0.773992	0.055556	0.285714	0.73133	0.058824
14	0	0.432432	0.350681	0.0625	0	0.736462	0.25
15	0	0.378378	0.177664	0	0.142857	0.723895	0
16	0	0.351351	0.265937	0	0.142857	0.612001	0
17	0.111111	0.297297	0.566568	0	0.142857	0.603928	0.181818
18	0	0.486486	0.332895	0	0.142857	0.817477	0.277778
19	0	0.702703	0.315463	0.074074	0	0.892209	0.076923

TABLE V. FUZZY RULES

IF							THEN
F1	F2	F3	F4	F5	F6	F7	
VH	H	VH	M	VH	VH	M	Important
VH	*Semua*	VH	*Semua*	VH	VH	*Semua*	Important
H	*Semua*	VH	*Semua*	VH	VH	*Semua*	Important
VH	*Semua*	VH	*Semua*	H	VH	*Semua*	Important
VH	*Semua*	H	*Semua*	VH	VH	*Semua*	Important
VH	*Semua*	VH	*Semua*	VH	H	*Semua*	Important
H	*Semua*	VH	*Semua*	H	VH	*Semua*	Important
H	*Semua*	H	*Semua*	H	VH	*Semua*	Important
H	*Semua*	H	*Semua*	H	H	*Semua*	Important
VH	*Semua*	H	*Semua*	VH	H	*Semua*	Important
VH	*Semua*	H	*Semua*	H	H	*Semua*	Important
H	*Semua*	H	*Semua*	VH	H	*Semua*	Important
VH	*Semua*	VH	*Semua*	H	H	*Semua*	Important
H	*Semua*	H	*Semua*	VH	VH	*Semua*	Important
H	M	H	L	H	H	L	Average
M	*Semua*	M	*Semua*	M	M	*Semua*	Average
M	*Semua*	H	*Semua*	H	H	*Semua*	Average
H	*Semua*	H	*Semua*	M	H	*Semua*	Average
H	*Semua*	M	*Semua*	H	H	*Semua*	Average
H	*Semua*	H	*Semua*	H	M	*Semua*	Average
M	*Semua*	H	*Semua*	M	H	*Semua*	Average
M	*Semua*	M	*Semua*	M	H	*Semua*	Average
H	*Semua*	M	*Semua*	H	M	*Semua*	Average
H	*Semua*	M	*Semua*	M	M	*Semua*	Average
M	*Semua*	M	*Semua*	H	H	*Semua*	Average
H	*Semua*	H	*Semua*	M	M	*Semua*	Average
M	*Semua*	M	*Semua*	H	H	*Semua*	Average
Semua	*Semua*	*Semua*	*Semua*	VL	*Semua*	*Semua*	Unimportant
Semua	*Semua*	*Semua*	*Semua*	L	*Semua*	*Semua*	Unimportant
Lainnya							Unimportant

IF ((F1 is VH) and (F2 is H) and
(F3 is VH) and (F4 is M) and
(F5 is VH) and (F6 is VH) and
(F7 is VH))
THEN
(Sentence is Important)

Fig. 4. The rule for categorizing sentence as important

TABLE VI. THE MEMBERSHIP OUTPUT AS THE RESULT OF FUZZY INFERENCE MACHINE

No	Sentences	Weight Average	Actual Result
1	Bogor, CNN Indonesia Presiden RI Joko Widodo (Jokowi) meminta para siswa SMA Taruna Nusantara berani melawan fitnah dan hoaks alias kabar bohong yang beredar di media sosial.	2.99999999 99999996	Important
2	Jokowi mengatakan para siswa jangan diam ketika mengetahui ada hoaks yang disebarkan kepada masyarakat.	1.0	Unimportant
3	"Saya minta semuanya para siswa harus berani meluruskan, berani merespons kalau ada kabar fitnah, kabar bohong, hoaks, yang banyak beredar di media sosial," kata Jokowi di Ruang Garuda, Istana Kepresidenan Bogor, Kota Bogor, Jawa Barat, Senin (4/3).	2.0	Average
4	Jokowi menyampaikan pengarahan kepada 366 siswa kelas 11 SMA Taruna Nusantara yang berasal dari seluruh provinsi di Indonesia.	1.0	Unimportant
5	Selain itu hadir pula 20 orang pamong.	1.0	Unimportant
6	Para siswa tersebut kompak mengenakan seragam warna biru.	1.0	Unimportant
7	"Berani meluruskan yang benar katakan benar, yang salah katakan salah.	1.0	Unimportant
8	Jangan dibalik-balik," ujar Jokowi menambahkan.	1.0	Unimportant
9	Jokowi tak menyebut contoh fitnah dan hoaks yang kerap muncul di media sosial.	1.0	Unimportant
10	Namun, umum diketahui bahwa dalam setiap kesempatan Jokowi selalu meminta masyarakat melawan fitnah dan hoaks terkait keterlibatan PKI, kriminalisasi ulama, maupun antek asing.	1.05088390 0695649	Average
11	Kepada para siswa SMA Taruna Nusantara, Jokowi melanjutkan agar urusan politik tak membuat masyarakat terpecah belah.	1.20627375 63819688	Average
12	Ia pun menegaskan pemilihan umum, baik pemilihan bupat, wali kota, gubernur maupun presiden kontestasi politik yang rutin dilakukan lima tahun sekali.	1.0	Unimportant
13	Jokowi menyebut gesekan di tengah masyarakat biasanya dimulai gara-gara politik elektoral yang digelar lima tahun sekali tersebut.	1.0	Unimportant
14	Padahal, kata mantan wali kota Solo itu, pesta demokrasi lima tahunan itu seharusnya disambut dengan gembira.	1.0	Unimportant
15	"Sebagai anak bangsa sangat rugi besar kita, gara-gara urusan politik antar-teman enggak saling ngomong.	1.0	Unimportant
16	Harus dipakai sebagai pendewasaan politik, pematangan politik kita bagaimana memilih seorang pemimpin," ujarnya.	1.0	Unimportant
17	Jokowi pun mengingatkan perbedaan yang di tengah masyarakat merupakan anugerah Tuhan.	1.0	Unimportant
18	Mantan gubernur DKI Jakarta itu tak ingin perbedaan suku, agama, adat, budaya, bahasa daerah justru membuat masyarakat terbelah.	1.0	Unimportant
19	"Jangan sampai karena perbedaan ini kita menjadi tidak seperti saudara, saudara se-Tanah Air," kata pria yang ikut dalam Pilpres 2019 sebagai capres nomor urut 01 tersebut.	1.0	Unimportant

Jokowi Minta Siswa SMA Taruna Nusantara Berani Lawan Hoax
Politics
Bogor, CNN Indonesia Presiden RI Joko Widodo (Jokowi) meminta para siswa SMA Taruna Nusantara berani melawan fitnah dan hoaks alias kabar bohong yang beredar di media sosial. "Saya minta semuanya para siswa harus berani meluruskan, berani merespons kalau ada kabar fitnah, kabar bohong, hoaks, yang banyak beredar di media sosial," kata Jokowi di Ruang Garuda, Istana Kepresidenan Bogor, Kota Bogor, Jawa Barat, Senin (4/3). Namun, umum diketahui bahwa dalam setiap kesempatan Jokowi selalu meminta masyarakat melawan fitnah dan hoaks terkait keterlibatan PKI, kriminalisasi ulama, maupun antek asing. Kepada para siswa SMA Taruna Nusantara, Jokowi melanjutkan agar urusan politik tak membuat masyarakat terpecah belah.

Fig. 5. The final result of summarization

IV. CONCLUSION

This study applies Automatic Text Summarization which is equipped with categorization on a number of online news articles. The algorithm used is Fuzzy Logic with function extraction to give weight to sentences with several features such as: Title Words, Sentence Length, Sentence To Sentence Similarity, Numerical Data, Thematic Words, Term Weight, and Proper Nouns. The test was carried out on 50 online news articles sourced from https://www.cnnindonesia.com/ to be summarized when the articles contained information about Indonesian public figures associated with the Advanced Indonesia Cabinet.

The results were evaluated by two Indonesian literature experts from the Faculty of Cultural Sciences, Sebelas Maret University. The evaluation results indicate that the results of this research have proven to be able to provide a summary of online news that is understandable, contains important information from related articles, and is considered efficient, and can save time. In addition, the experts considered that the categorization of news and categorization of national figures was appropriate.

For improvement in further research, it can be considered is to add more features in Feature Extractions in order to increase the possibility of sentences having 5W + 1H (What, When, Where, Who, Why, How) are included in the summary.

REFERENCES

[1] Al-Qadri, B., "Persepsi Masyarakat Tentang Pemberitaan Delik." in Jurnal Supremasi," 2016, pp. 163-173.

[2] Setiawan, A., A., "Peran Media Massa Dalam Meningkatkan Kualitas." in Jurnal Ilmu Politik. 2013, pp. 39-48.

[3] Romadhony, A., R., F. Z., Yusliani, N., and Abednego, L., Text Summarization untuk Dokumen Berita Berbahasa Indonesia. ISSN: 2088-8252, 2017, pp. 408-414.

[4] Khotimah, N., Wibowo,A. , Andreas, B., Girsang, A.S., International Journal of Emerging Technology and Advanced Engineering, vol. 11, Issue 08, August 2021

[5] D.Patil, P., N.J.Kulkarni., "Text Summarization using Fuzzy Logic," in International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 1, Issue 3, 42-45, 2014.

[6] Shetty, A., Bajaj, R., "Auto Text Summarization with Categorization and Sentiment Analysis," in International Journal of Computer Applications, Vol. 130, No. 7, 57–60, 2015.

[7] Gupta, V., Lehal, G., S., "A Survey of Text Mining Techniques and Applications," in Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, 60-76, 2009. https://doi.org/10.4304/jetwi.1.1.60-76

[8] Nasution, H., "Design Methodology of Fuzzy Logic Control," Johor Darul Ta'zim, Malaysia: Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 2006.

[9] Putri, A.D., & Effendi, "Fuzzy Logic Untuk Menentukan Lokasi Kios Terbaik Di Kepri Mall," in Jurnal Edik Informatika, vol. 3, 2017, pp. 49-59.

[10] D Gunawan et al., "Automatic Text Summarization for Indonesian Language Using TextTeaser," in IOP Conference Series: Materials Science and Enginerring, vol 190, no. 012048

[11] Ridok, A., Romadhona, T.C.," Peringkas Dokumen Otomatis Menggunakan Metode Fuzzy Model Sistem Inferensi Mamdani," in Seminar Nasional Teknologi Informasi dan Multimedia, 2013

Dynamic Topic Tracking and Visualization using COVID-19 related Tweets in Multiple Languages

Riki Murakami
Dept. of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate 020-0693, JAPAN
g236r005@s.iwate-pu.ac.jp

Basabi Chakraborty
Dept. of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate 020-0693, JAPAN
basabi@iwate-pu.ac.jp

Yukari Shirota
Dept. of Management
Gakushuin University
Tokyo 171-8588, JAPAN
yukari.shirota@gakushuin.ac.jp

***Abstract*—With the increasing proliferation of mobile phone, internet and communication technologies, social network sites (SNS) are gaining importance worldwide. People express and exchange their opinion on various social network sites like twitter, facebook, blogs over different local and global issues. Efficient analysis of such vast text data from SNS provides a good way of understanding insights of public opinion, government policy and social condition of different countries. Topic modeling is a popular tool for extracting information from text data. Dynamic topic tracking and its visualization provides a means for capturing the change of topics over time which is important for visualization of changing needs of the society and keeping updated with the current situation. In this work, COVID-19 related twitter data in two different languages are collected and analyzed by dynamic topic model to track the spread of the evolved topics during the pandemic in two different countries in order to visualize the differences and commonness of the effect of pandemic. Here we mainly focused on the tweet data related to Japan and India in Japanese and English respectively. It is found that the country specific characteristics are prominent in some topics while some topics express the general concerns during the pandemic. This study seems to be effective to provide a technique for capturing the opinion and needs of people during a pandemic by analysis of tweet data.**

***Index Terms*—Neural topic model, dynamic topic tracking, short text, word embedding**

I. Introduction

Topic modeling is an important technique for extracting significant semantically structured contents from text data in the area of natural language processing. Latent Dirichlet Allocation (LDA) is the most popular topic model proposed in [1]. Topic models assume that a document is generated by probabilistic distribution of multiple latent topics consisting of semantically similar words. In the area of social data mining applications, topic models are widely used for extraction of information on social events from text data on social networking sites like twitter, facebook or blogs [2]. Topic models based on LDA or Probabilistic Latent Semantic Analysis (PLSA) have been used in several real world applications like summarization of reviews, information retrieval or trend detection from short or long texts [3]. Dynamic topic models are useful to capture the change or movement of topics over a time period [4]. Recently deep neural network based topic models are also proposed for text data analysis [5].

For the past several months, the new coronavirus (COVID-19) has spread worldwide, and there is still no sign of the infection being contained. Governments around the world are implementing various measures to cope with COVID-19 according to the characteristics of each country, the public opinion and government strategy on this pandemic differs from country to country. The opinions of people of different nationalities, religions, genders, can be observed in documents posted on social networking sites such as Twitter, Blogs, Facebook etc. If such documents can be analyzed to gather the opinions of the public, it may be possible to come up with a response to COVID-19 that is more in line with actual public sentiment and public demand. The analysis of COVID-19 related tweet data is gradually becoming an active research area and some studies have been published recently [6] [7] [8].

In this work, we collected Tweets about COVID-19 for a certain period of time and analyzed them using several text mining techniques. We used Dynamic Topic Modeling to visualize how the topics changed through the passage of time and what topics dominated people's attention on Twitter. Here we focus on tweets about India and Japan in particular, and collect and analyze tweets related to these two countries in two different languages to see how the topics differ from country to country. We also tried to separate the universal issues of this disaster from the local issues of each country. The next section describes some related works on topic models and the analysis of tweeter messages about COVID-19.

II. Related Works on COVID-19 Tweet Analysis and Topic Models

With the unabated spread of the COVID-19 pandemic, scientific research on various aspects of the pandemic has been accelerated. In the area of natural language processing and text data mining, researchers have started analysing text data on SNS about COVID-19 to extract knowledge regarding the nature and effects of the pandemic.

In [6], analysis of Covid-19 related twitter data has been done by topic model in order to extract the setiment dynamics of general public on health status and other concerns related to the pandemic. The study in [7] reveals that the increase of infection leads to spread of some misinformation related to

the pandemic over general public through SNS messages. The government should be aware of this spread of fake information and take appropriate action to deliver correct information to its citizen. A country specific analysis of public sentiment has been done in [8] by sentiment analysis and topic modeling by Gibbs Sampling Dirichlet multinomial mixture (GSDMM). Recently multilingual COVID-19 tweet data are collected by researchers in different countries and are analyzed with the help of natural language processing (NLP) and machine learning tools in [9] to explore popular discourse about the pandemic and policies implemented to mitigate it. The study on COVID-19 related tweet data in [10] used LDA topic model and Long Short-term Memory (LSTM),a deep learning based NLP method, to detect sentiment trends in a number of topics like social justice, mental health, vaccines. misinformation etc.

In most of the analysis of COVID-19 based tweets, LDA based topic models are used which does not take into account the temporal information. But as the pandemic continues, people's focus changes and the trend of different topics gradually changes. So tracking the trending topics over different time windows is essential to understand general public's needs and interests. Dynamic topic model(DTM) [4] is a representative method for capturing temporal variation of topics. Text data with time-series information is generally huge in size, and their processing requires high computational cost. On the other hand, in recent years, several software packages have been released to efficiently train machine learning models such as neural networks using GPUs, and research has been very active in this area. In response to this, methods for training topic models using Variational Autoencoders (VAE) [11] have begun to appear in the field of topic models. For example, Neural Variational Document Model (NVDM) [12], ProdLDA [13], and Gaussian Softmax Model (GSM) [14] are well known.

The Dynamic Embedded Topic Model (D-ETM) [15] was proposed as an extension of this VAE-based method to use time series information. D-ETM is a method that has both the advantages of DTM and VAE. It is capable of constructing topics that take into account the temporal evolution of topics like DTM. Also, D-ETM shows improved efficiency of training with GPUs like VAE-based topic models. In our study, we used GSM and D-ETM.

III. Experimental Methods and Data sets

In this section, the experimental methods for analysis of tweet data and the COVID-19 data sets have been described.

A. Data

In our experimental analysis, we used two COVID-19 tweet data sets : Japanese language twitter data from Japan and English language tweet data from India. The details are in the following subsections.

1) Tweets about Japan: The COVID-19 Japanese Twitter Data set [1] is a collection of labelled tweets with IDs containing the word "COVID" or corona from January to June of 2020. We extracted tweets from twitter using Hydrator[2]. The original data set contains 53,640 tweet IDs, but the number of available tweets was reduced. After processing, only Noun, Adjectives and Adverbs are extracted using SpaCy[3]. We also excluded words that appeared in fewer than 10 documents. Finally, each document was only retained if it had a total token count of at least 2 after preprocessing. We also removed user IDs, URLs, and numeric-only tokens. This resulted in 34558 tweets and 5450 unique tokens.

2) Tweets about India: For English language India based COVID-19 related tweet data, we collected from Twitter using Twitter API[4] containing keywords "India pandemic" or "India COVID" during April 2, 2021 to June 22, 2021, mainly during second wave of corona infection in India. For these tweets also, we performed the same pre-processing as we did for Japanese tweets. As a result, 424447 tweets are available for our study which includes 17328 unique tokens.

B. Experimental Analysis

The analysis is done in two steps. In the first step we performed topic modeling using Gaussian Softmax Model (GSM), a neural topic model, on the entire corpus ignoring the timestamps. This allows us to check the overall topics during the whole period of data collection and we have also checked how the keywords related to COVID-19 are used in the corpus.

In the second step, as mentioned in the previous section, D-ETM is used for dynamic topic tracking and visualization of changes of topics over time and Jax [16]/Flax [17] was used to implement this. The timestamp in this paper is constructed based on the collected data. 6 months of tweets about Japan were collected, so the month is used as the time unit. 3 months of tweets about India were collected, so the week is used as the time unit. D-ETM requires that the number of latent topics has to be given by the user in advance. In this case, the number of latent topics was set to 50 after some initial experiments, since we are interested in the readability of the visualization of the inference results.

C. Topic Evaluation Metric

For evaluation of the constructed topics, the following metrics for assessing topic coherence are used.

- Normalized Point-wise Mutual Information (NPMI) [18]: NPMI is a measure of the semantic coherence of a group of words and is defined by the following equation:

$$NPMI(w) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \quad (1)$$

  where w is the list of top-N words for a topic. N is usually set to 10. For K topics, average of NPMI over all topics are used for evaluation.

[1] http://www.db.info.gifu-u.ac.jp/covid-19-twitter-dataset/

[2] Hydrator: https://github.com/docnow/hydrator

[3] spaCy: Industrial-strength NLP https://github.com/explosion/spaCy

[4] Twitter API: https://developer.twitter.com/en

- Word Embeddings Topic Coherence (WETC) [19]: WETC represents word embedding based topic coherence and pairwise WETC for a particular topic is defined as

$$\mathrm{WETC}_{PW}(E) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \langle E_{i,:}, E_{j,:} \rangle \quad (2)$$

where $\langle ., . \rangle$ denotes inner product. For calculation of the WETC score for English corpus, the pre-trained weights of GloVe[5] [21] has been used. For the Japanese pre-trained word embedding vector, we used chiVe[6]. Since we could not find any GloVE like pre-trained vectors for Japanese, we employed Skip-gram trained on a large corpus of Japanese as an alternate method.
- Topic Diversity (TD) [20]: Topic diversity is defined here as the percentage of unique words in the top 25 words of all topics according to [20]. Diversity close to 0 represents redundant topic and close to 1 indicates more varied topics.

IV. Results and Analysis

The results of topic modeling with tweet data is described in the section.

A. Quantitative evaluation

The results of topic coherence metrices are shown in the Table I. All of these metrics are better when they are larger. WETC evaluates topic quality with a large external corpus, and NPMI evaluates the same with the corpus used for training, TD is Topic Diversity.

TABLE I
TOPIC EVALUATION SCORES

Language of Tweets	With GSM			With D-ETM		
	WETC_{pw}	NPMI	TD	WETC_{pw}	NPMI	TD
English	0.28	-0.87	0.84	0.09	-0.33	0.61
Japanese	0.171	-0.87	0.48	0.19	-0.82	0.92

B. Overview of Topics

Table II and III present the COVID-19 related topics from the 50 topics generated by Gaussian Softmax Model (GSM) for Japanese tweet data and Indian tweet data respectively. Each latent topic has a probability distribution of occurrence for all words that appear in the corpus. The top 10 words in the order of high probability of occurence are shown in the tables.

Considering India related tweets in English in Table II, it seems that there are more COVID-19 related topics for India compared to the Japanese one. Also, there are more hashtags than the Japanese Tweets, appearing as top probability words. Topic 20 is a topic related to media like BBC, podcasts, etc. Topic 43, as represented by '#wearmask', seems to be a collection of things that are generally considered to be required as anti-corona measures.

Similarly from the results for Japanese tweets in Table III, Topic15 was strongly influenced by those tweets that talked about COVID-19 reading as "new pneumonia". Also, some conspiracy theory related remarks were found. Topic16 is related to the WHO announcement. Discussions about the relationship between obesity and smoking also influenced the construction of the topic. Topic 44 is about the explosive increase of infections in Japan and the United States. Topic 46 is not clear if it is a corona topic, but it correlates well with other corona related topics. The tweets in Japanese are from the year 2020. Perhaps that is why there seems to be more talk about foreign countries like the U.S. and about prevention, etc., than about corona infection in Japan.

Comparing the analysis results of the two languages, we find that English has more names of famous people and hashtags. English hashtags are often sentences or phrases, and the use of hashtags in English is relatively more positive than in Japanese. It is important to note that there is a difference in the way the corpus has been collected, but it is clear that English has a higher importance in the analysis of hashtag trends than Japanese.

C. How keywords are used in the corpus

Each word in the trained GSM corresponds to a topic embedding vector. It can be used as an analogy for how a word has been used in this twitter corpus. Some keywords related to COVID-19 and a list of similar words are presented in Table IV and Table V.

The keyword "pandemic" is a good description of the time when the tweet was posted. It can be seen that the tweet was posted in the context of a discussion about the deterioration of business performances of the large Indian companies. Words similar to "vaccine" include the names of vaccine manufacturing companies and many words related to vaccine production, such as "authorization" and "patent". This was around the time when vaccines had started to spread around the world and the vaccine related topic was a big part of the tweets about India. In this table, the keyword "home" was used in place of going out, but unlike expectations, it is not often used in the context of "stay home.

We used Japanese words meaning "pandemic," "vaccine," "going out," "children," "elderly," and "medical care" as keywords. Similar words for "pandemic" include "overshoot", "simulation", and "American newspaper". The timing of the tweet may be indicative of the transitional period of the COVID-19 outbreak in Japan. Similar words for "go out" include "refrain," "unnecessary," and "stay." These make sense, but it is interesting to see words like "coupon" and "ticketing". It seems to be referring to the coupons that the Japanese government and Japanese local governments announced and introduced as an economic recovery measure at the beginning of the coronavirus disaster. As for "children" and "old people," the results show that they are not used in a way that is specific

[5]pre-trained weights obtained from https://nlp.stanford.edu/projects/glove/

[6]chiVe: Japanese Word Embedding with Sudachi & NWJC https://github.com/WorksApplications/chiVe

TABLE II
TOPIC EXTRACTION USING GSM (INDIA)

Topic2	Topic17	Topic20	Topic21	Topic22	Topic37	Topic43
stress.	negotiation	thankyou	take.	estimate	colossal	turning
boarding	caste	#empoweringthepoor	genome	cases/day	sero	#wearmask
profession	colour	fact-checker	expenditure	rea	sma	#covidupdates
virus'	infact	patahi	wion	band	objection	saint
down.	joke	pandemics.	#policy	vigil	objective	isolation
fungus'	pandemic':	#bbc	sensex	hon	#ayaansh	haridwar.
centr	#pandemic:	#podcast	pt	hospitalisation	#pandemic_in_india	nrc
delta	#ramnavami	grea	screening	minor	desi	doctor's
si	guidance	conduct	pe	india"	#end	#state
subramanian	adherence	cur	concern.	factor	safe.	thumping

TABLE III
TOPIC EXTRACTION USING GSM (JAPAN)

Topic15	Topic16	Topic44	Topic46
(Virus)	(Death)	(Infection)	(Rubella)
(New)	(Meifu)	(Expansion)	(Get)
(Pneumonia)	(mistakes)	(in-hospital)	(consultation)
(Spraying)	Who (Who)	(Prevention)	(Chemistry)
(Snake)	(Obesity)	(Epidemiology)	(Well-known)
(Specialist)	(CauseofDeath)	(Perspective)	(HighFever)
(Epidemic)	(Tobacco)	(Explosion)	(Free)
(Behind)	(Caucasian)	(DomesticandForeign)	(Inoculation)
(Dashboard)	(Preferential)	(USOpen)	(Measles)
(Toxicity)	(Cancer)	(Splashes)	(Jurisdiction)

to the corona disaster. For "medical care," relatively good words such as "donation," "gratitude," and "respect" are listed.

Comparing the two languages, there are significant differences in topics related to "children" and the "elderly". These words do not show any relationship to words about coronas in the Japanese corpus, but they clearly appear in similar words in English. It is interesting to note that the word "overshoot," which was used by politicians in Japan to describe "explosive infections," has come to bear a large degree of similarity to the "pandemic" in Japanese corpus.

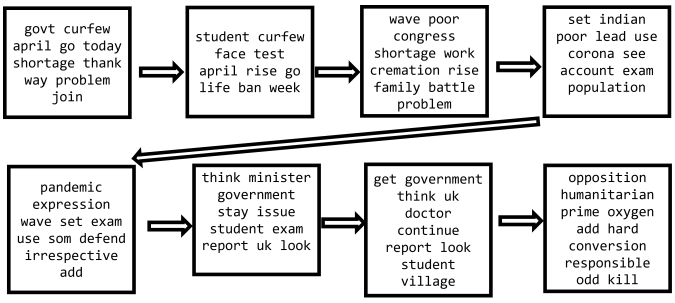


Fig. 1. Top10 token based on term-score (India)

D. Changes in Topics with Time

Here we describe the analysis of the results of our second step of experiment where we considered the changes of topics with time. We considered the topics for each time period (one month unit for Japanese tweets and one week unit for English tweets) We constructed topics by respective timestamp using D-ETM. If the topic model behaved as we expect, every topic would be a set of words with different meanings, and we would be able to understand the topic of the whole document at a glance. However, in this experiment, several topics were filled with generic words that had no meaning in the current analysis. But we obtained some topics which carry meaning in the context of corona. One such example of the topic that is somewhat easier to understand is shown in Fig. 1.

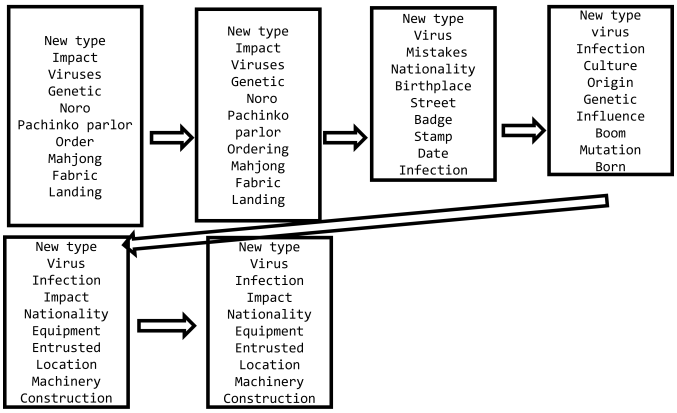


Fig. 2. Top10 token based on term-score (Japan)

In English corpus, this topic contains words about the government's policies towards schools and students, and at the beginning one could find those issues and concerns raised, but gradually it turns to more specific topics such as liability and oxygen injectors. In Japanese corpus, the topic has not changed much regarding the context with time as shown in Fig. 2.

Overall, we were not able to visualize major changes in topics or shifts in discussions in this analysis according to the results shown in Figure 3. This figure is the result of D-ETM that assigns 10 as the number of topics with English corpus. Each document can be embedded into a potential semantic vector (topic vector) with a number of topic elements by using Topic Model. Here, we converted all sentences appearing at each time into a topic vector and then averaged them. By doing so, we can find in Fig 3 the topics that are well discussed at

TABLE IV
WORDS THAT ARE HIGHLY SIMILAR TO THE KEYWORD (INDIA)

	pandemic	vaccine	home	child	elder	medical
0	#revival_of_bsnl	producer	suvs	resultant	populati	moment
1	#survival_of_bsnl_families.	maker	buy	pediatrician	protection	history
2	nomad	dose	dexamethasone	panelist	justice.	junior
3	rehabilitation	covishield	luxury	way!	one	thur
4	forever.	biotechnology	#nation_with_modi	reason:	#covid	healthcare
5	families.	pfizer	success.	#webinar	solution	#indiafightsback
6	downfall	crunch	isolation.	bride	sister	show.
7	convalescent	authorisation	#bhfyp	parent	experience	faculty
8	date.	tracker	cashless	most.	accountability	#pehlejaanphirexam
9	communities.	patent	exec	adolescent	#tata	world's

TABLE V
WORDS THAT ARE HIGHLY SIMILAR TO THE KEYWORD (JAPAN)

	(Pandemic)	(Vaccine)	(Going out)	(Children)	(Elderly)	(Medical)
0	(Overshoot)	(Development)	(Self-restraint)	(Adults)	(Elderly)	(Engagement)
1	(Explosion)	(ClinicalTrial)	(Unurgent)	(Son)	(Security)	(Fundraising)
2	(Ricepaper)	(Inoculation)	(Rock-paper-scissors)	(Childcare)	(Hate)	(Institution)
3	(Times)	(Astra)	(Living)	(School)	(Nursing)	(Hikakin)
4	(Public)	(Patent)	(Redemption)	(Marriage)	(Family)	(Impulse)
5	(Cases)	(Pharmaceuticals)	(Ticketing)	(Children)	(Development)	(Weakness)
6	(Simulation)	(Practical)	(Necessary)	(ElementarySchool)	(Clearance)	(Thanks)
7	(Behind)	(Uterus)	(Coupons)	(EntranceExams)	(Theater)	(Support)
8	(Ethnicity)	(Treatment)	(Stay)	(Vegetables)	(Gender)	(Respect)
9	(Intention)	(Rubella)	(Instructions)	(Husband)	(Toku)	(Site)

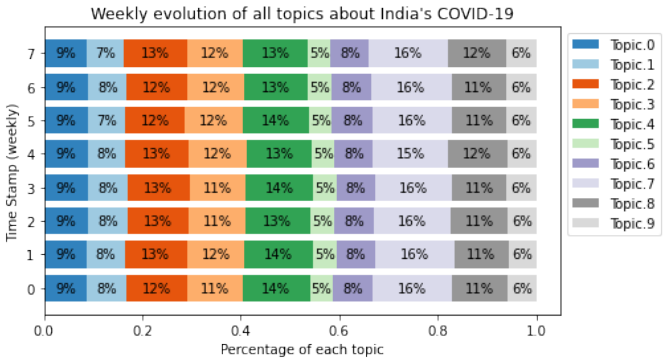


Fig. 3. Development of Topic proportion by times (India)

each time unit.

Every word (token) that appears in the corpus has its own probability of occurrence from each topic. We assume that the topic with the highest probability of occurrence is the topic to which the word belongs. Fig.4 shows the number of words belonging to each topic and the topics to which the words are spreading at the consecutive time units. This means that we can find the range of words covered by each topic.

It can be seen from these figures that the average percentage of topic allocation for documents per time has not changed much and that the cluster size of topics has changed only in a very small number of very large topics, with no overall movement. In other words, these figures show that the topic of COVID-19 has not changed significantly during the period of data collection.

V. CONCLUSION

In this paper, important and dynamic topics from twitter data related to COVID-19 disaster in two languages has been extracted with the help of several text mining tools. The comparative analysis of tweet data in two languages mainly related to two countries has been presented. The differences in the findings in two countries are highlighted. Though the data collected in the two languages correspond to two different time slots, we could able to find out the characteristic topics during the whole span of the time period considered. The interesting country specific topics are also found.

In the second part of the analysis for the dynamic topics, not much overall change of topics has been noted in the collected tweets. The reason may be that the time period of data collection is shorter compared to the duration of the pandemic which is still ongoing. We need to use tweet data for a longer time period to analyse the changes in future. This study seems to be effective to find out the opinion and needs of general public during a disaster by analysis of tweet data.

REFERENCES

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] E. S. Negara, D. Triadi and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019, pp. 386-390, doi: 10.1109/ICECOS47637.2019.8984523.

[3] J. Boyd-Graber, Y. Hu and D. Minmo, *Applications of Topic Models* , now, 2017.

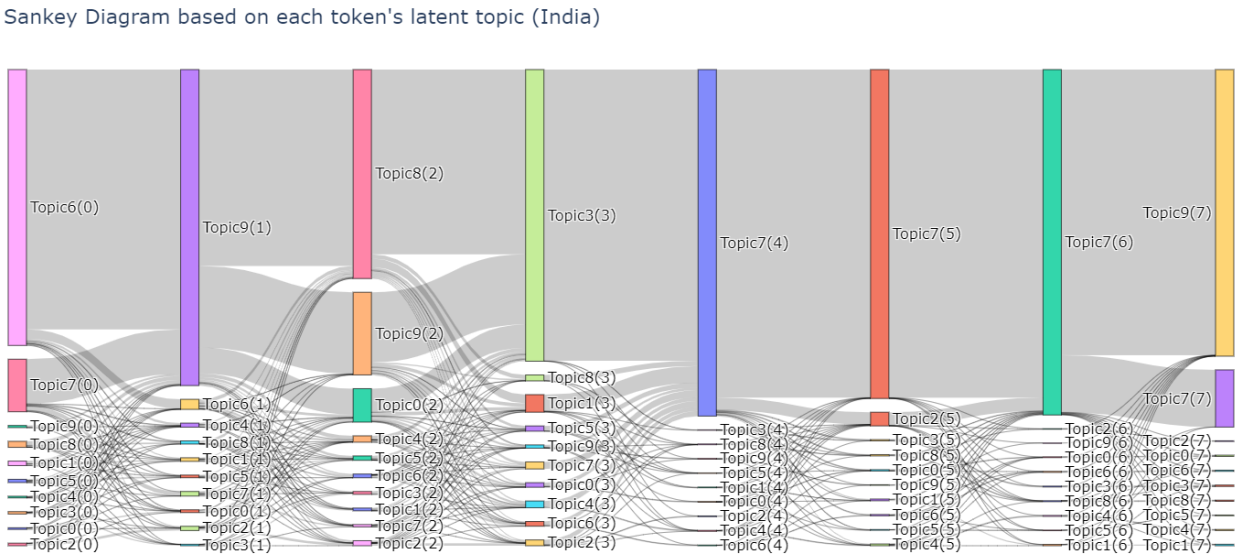


Fig. 4. Development of size of Topic cluster(India)

[4] D. M. Blei and J. D. Lafferty, " Dynamic Topic Models", In *Proceedings of the 23 rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[5] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du and W. Buntine," Topic Modelling Meets Deep Neural Networks: A Survey" , arXiv:2103.00498v1, 2021.

[6] M. S. Ahmed, T. T. Aurpa, M. M Anwar, "Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic", *PLOS ONE* 16(8), 2021.

[7] H. Wang, Y. Li, M. Hutch, A. Naidech, Y. Luo, "Using Tweets to Understand How COVID-19–Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study", *J Med Internet Res*, 23(2), 2021.

[8] K. M. Ridhwan, C. A. Hargreaves, "Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore", *International Journal of Information Management Data Insights*, Volume 1, Issue 2, 2021.

[9] C. E. Lopez, M. Vasu and C. Gallemore, "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset" arXiv:cs.SI/2003.10359,2020.

[10] Y. Song, X. Wang and Y. Jia, "Deep Learning-Based COVID-19 Twitter Analysis, *Proc. of 6th International Conference on Big Data and Computing*, 2021.

[11] D. P. Kingma and M. Welling,"Auto Encoding Variational Bayes", *Proc. 2nd International Conference on Learning Representations, ICLR*, 2014.

[12] Y. Miao, L. Yu and P. Blunsom, "Neural variational inference for text processing", *In Proc. of 33rd International Conference on Machine Learning*, 2016.

[13] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models", *In Proc. of 5th International Conference on Learning Representations, ICLR*, 2017.

[14] Y. Miao, E. Grefenstette and P. Blunsom, "Discovering discrete latent topics with neural variational inference", *In Proc. of 34th International Conference on Machine Learning*, 2017.

[15] A. B. Dieng, F. J.R. Ruiz and D. M Blei, "The dynamic embedded topic model". arXiv preprint arXiv:1907.05545, 2019.

[16] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs", 2018.

[17] J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Ronde- pierre, A. Steiner, and M. van Zee, "Flax: A neural network library and ecosystem for JAX", 2020.

[18] J. H. Lau, D. Newman, and T. Baldwin, " Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality", *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, Association for Computational Linguistics, pp. 530 – 539, 2014.

[19] R. Ding, R. Nallapati, and B. Xiang, "Coherence-Aware neural topic modeling, *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 830 – 836, 2018.

[20] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces", *Trans. Assoc. Comput. Linguist.*, Vol. 8, 439 – 453, 2020.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation", *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA, 2014.

Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier

Khansa Afifah[ad1], Intan Nurma Yulita[bd2], Indra Sarathan[cd3]
[a]*Department of Mathematics*
[b]*Department of Informatics Engineering*
[c]*Department of Indonesian Literature*
[d]*Research Center for Artificial Intelligence and Big Data*
Universitas Padjadjaran
Sumedang, Indonesia
e-mail: [1]khansa18001@mail.unpad.ac.id, [2]intan.nurma@unpad.ac.id, [3]sarathan@unpad.ac.id

***Abstract*— In recent years, companies have widely used sentiment analysis with machine learning classification algorithms to help business decision-making. Sentiment analysis helps evaluate customer opinions on a product in goods or services. Companies need this opinion or sentiment to improve the performance, quality of their products, and customer satisfaction. Machine learning algorithms widely used for sentiment analysis are Naive Bayes Classifier, Maximum Entropy, Decision Tree, and Support Vector Machine. In this study, we propose an approach of sentiment analysis using a very popular method, Extreme Gradient Boosting or XGBoost. XGBoost combines weak learners into an ensemble classifier to build a strong learner. This study will focus on the reviews data of the most popular telemedicine application in Indonesia, Halodoc. This study aims to examine the people'ssentiment towards telemedicine applications in Indonesia, especially during the COVID-19 pandemic. We also showed a fishbone diagram to analyze the most factors the users complained about. The data we have are imbalanced; however, XGBoost can perform well with 96.24% accuracy without performing techniques for imbalanced data.**

Keywords—sentiment analysis, machine learning, telemedicine, reviews, xgboost.

I. INTRODUCTION

Amid the COVID-19 pandemic that has hit Indonesia since early 2020, the government has taken several steps to prevent the spread of COVID-19 community activities. The government is also actively involved in ensuring that the public adheres to health protocols. Even though so many people do not comply with the rules set by the government, daily COVID-19 cases in Indonesia also reached a record high on July 14, 2021, which increased by 54,517 points. As a result, the number of patients admitted to the hospital grew, but the limited capacity of the hospital was not able to accommodate all patients. This impacts medical personnel who feel overwhelmed by the increasing number of COVID-19 patients. Finally, the Ministry of Health and the government urged the public to use telemedicine. Telemedicine applications aim to reduce the number of patients who are not seriously ill in hospitals to increase the capacity of patients who need treatment. COVID-19 patients who are self-isolating at home can use the telemedicine application to see a doctor at any time, either looking for prescription drugs or buying drugs from pharmacies and then sending them to their homes. Halodoc is a telemedicine application that is very popular in Indonesia and has been very successful at the age of 5 years. On August 13, 2020, CB Insights named Halodoc one of the startups on the Digital Health 150 list, including 150 of the most promising digital health companies. This is the second year in a row that Halodoc has entered the Virtual Care Delivery category at this event. Halodoc has various health services, namely Pharmacy Delivery, Contact Doctor, Appointment, and others. During the pandemic of COVID-19, Halodoc also provides vaccination programs and COVID-19 tests to support the government in breaking the chains of transmission of Coronavirus. Halodoc still needs to improve the quality of their product and evaluate their performance by looking at the data of their product sentiment. With reviews on the google play store, companies can use them to see public sentiment regarding the version of their application.

In this study, an analysis of public sentiment will be carried out on the usage of Halodoc telemedicine application in Indonesia. Sentiment analysis, also known as opinion mining, is one of the most critical tasks in natural language processing. Sentiment analysis works by analyzing a text's mood, emotion, or feeling towards products, services, individuals, organizations, and events [8]. The sentiment is then classified as positive, negative, or neutral. Sentiment analysis is often used to improve decision-making and customer satisfaction in a company's business process. Sentiment analysis has become an exciting research topic in various fields such as products reviews [6], [17], services [5], [17], politics [1], [2], and even in gaming chat applications to reveal the existing of cyberbullying among online gamers [11]. Data for sentiment analysis can mainly be obtained from social media such as Twitter and Facebook, application reviews in google play store and app store, blogs, and websites. Sentiment analysis has been widely applied to various languages. In [5], the authors conducted a sentiment analysis on an e-payment service in Jordan which the data were in Arabic taken from social media Facebook and Twitter. The authors revealed that one of the most complex challenges when they had to handle Arabic was Dialectical Arabic, which is an understanding that can help identify the context. Another study [17] has conducted sentiment analysis on online product reviews in Chinese.

In recent years, sentiment analysis has been used with machine learning dan deep learning models. In [7], the authors conducted sentiment analysis research using the proposed method, namely SentiXGBoost. SentiXGBoost is an XGBoost model that combines several models such as Decision Tree, Naïve Bayes, Random Forest, KNN, LR, and SGD to be trained as base classifiers and got an accuracy of 90.8%. In [5], the authors analyzed sentiment using a proposed methodology that combines a neutrality detector model with XGBoost and genetic algorithms. Inspired by [5] and [7], we used XGBoost for this study and applied it in the

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

field of healthcare service reviews in Indonesia. We used the Sastrawi library to handle preprocessing sentiment analysis steps in the Indonesian language.

II. METHODOLOGY

In this section, we present the methodology used in this study. There are five stages carried out, as shown in Fig 1. The first stage is data collection. In this stage, we collected the reviews data of Halodoc applications using the google-play-scraper library in Python. After collecting the data, we labeled the sentiment of each review. In the second stage, we conducted data preprocessing to the dataset, consisting of case folding, punctuation removal, stopwords removal, tokenization, normalization, and stemming. Data preprocessing aims to remove noise, so the texts are cleaner and more understandable. In the third stage, we performed feature extraction to transform text data into numeric or vector data. Next, we divided the dataset into training data and test data with a proportion of 75:25. We built a machine learning classification model in the next stage and trained our model to predict positive and negative sentiment. We evaluated the model using k-fold cross-validation and confusion matrix in the last stage.

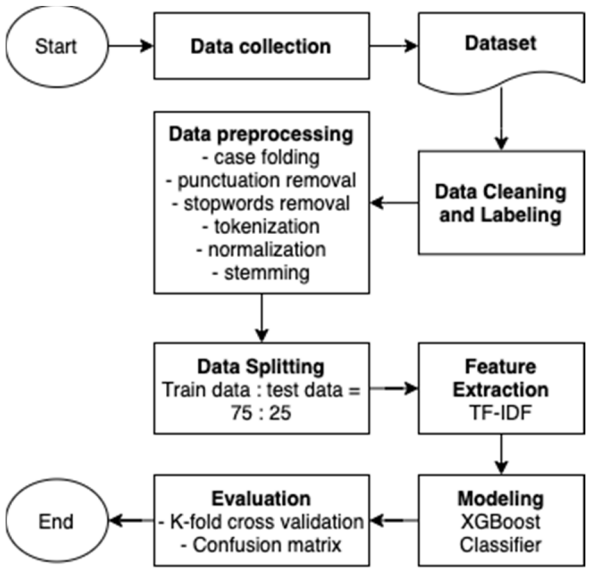


Fig. 1. Proposed Methodology

A. Data Collection

Data were collected from Google Play Store using the google play scraper library in Python. We got 12,969 data reviews from January 1st – September 30th, 2021. The data reviews contain users' names, content, rating scores, dates, and replies. However, we only used the content, rating score columns.

B. Data Cleaning and Labelling

After we collected the data, it was saved into a .csv format file. The reviews contain emojis, which we don't need for analysis, so we removed the emojis using the emoji library in Python. After removing emojis, we continued to label our data. We mapped the rating score using the definition determined by google play store to label the data. Score 1 for negative, 2 for somewhat negative, 3 for neutral, 4 for reasonably positive, and 5 for positive. We only take the negative and positive labels; it yields 11,550 data. We label negative as 0 and positive as 1, as illustrated in Fig. 2.

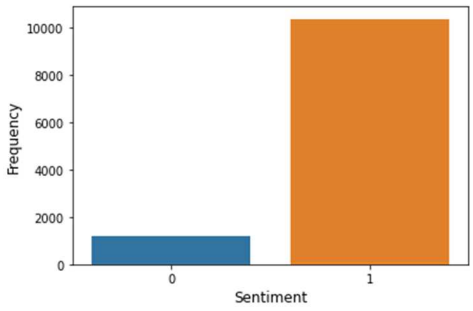


Fig. 2. Number of reviews based on sentiment

More detailed numbers for the figure above can be seen in Table 1.

TABLE I. NUMBER OF REVIEWS BASED ON SENTIMENT

Sentiment	Number of Reviews
Positive	10,353
Negative	1,197

The data we have is imbalanced. However, we did not perform any technique to handle imbalanced data in this study. We wanted to determine how good the XGBoost model is on an imbalanced dataset.

C. Data Preprocessing

In natural language processing (NLP), the information used contains unstructured data and a lot of noise. Therefore, it is necessary to convert the form into structured data before further processing. The data used in this study is reviews data in Bahasa Indonesia, therefore in this preprocessing stage, we used the Python Sastrawi library that can be accessed on github.com/har07/PySastrawi, which is a simple library to help text preprocessing such as stopword removal and stemming.

- Case Folding

  Case folding is a preprocessing stage to make all letters lowercase or uppercase. In most NLP cases, it is to convert all letters lowercase. Case folding aims to avoid two or more words with the same meaning but are treated differently by the machine due to writing in different forms; lowercase and uppercase.

- Punctuation Removal

  Punctuation removal aims to remove all the punctuation marks from sentences. A punctuation mark doesn't add extra information to the ruling. By removing punctuation marks, the dimension of our dataset can be reduced.

- Stopword Removal

  Stopword removal is one of the most common preprocessing steps used in NLP applications. The idea is to remove the most common words across all the documents. Stopword does not add much information to the text. Articles, prepositions, pronouns, and conjunction can be classified as stopwords.

- Tokenization

  Tokenization breaks the raw text into small units called tokens. These tokens help to understand the context in developing the model for NLP. Tokenization allows us to interpret the meaning of the text by analyzing the order of the words.

- Normalization

  Normalization is the process of converting a token to its basic form. The normalization process removes the inflected form of a word so that the basic form can be preserved. Normalization also transforms the short words or abbreviations into their complete form. For example, the term "tdk" is changed to its complete form, which is "tidak," and the word "baguuus" is transformed to its base, which is "bagus." For normalization, we used the abbreviation dictionary by meisaputri21 [13].

- Stemming

  Stemming is a step to remove affixes in a word, both affixes that appear before and after the word. Stemming converts each word to its root word without affixes.

D. Feature Extraction

Machines or algorithms cannot understand characters/words, it can only accept numbers as input. However, the inherent nature of textual data is unstructured and noisy, making it impossible to interact with machines. The process of converting raw text data into machine-readable formats (numbers) or features is called feature extraction from text data. There are certain techniques we can use for feature extraction such as Count Vectorizer and TF-IDF. TF-IDF stands for Term Frequency Inverse Document Frequency.

TF is simply a method to calculate the frequency of the occurrence of certain words in a document. The more often the word appears, the greater the TF value. While the IDF calculates the weight of a word against its appearance in the entire document. The more these words appear throughout the document, the smaller the IDF value. TF-IDF is calculated as Eq. 1.

$$IDF = \log\left(\frac{N}{DF}\right) \quad (1)$$

TF(k,d) denotes the number of the word shown in document d, while IDF(k) denotes the inverse document frequency, as shown in Eq. 2.

$$TF - IDF(d,k) = TF(d,k) \times IDF(k) \quad (2)$$

E. XGBoost Classifier

The XGBoost or Extreme Gradient Boosting algorithm was first developed as a research project at the University of Washington by Tianqi Chen and Carlos Guestrin. XGBoost is an implementation of gradient boosted decision trees designed to improve speed and performance. XGBoost is known for the ability to optimize the consumption of time, memory resources, and handle imbalanced data. The XGBoost or Extreme Gradient Boosting algorithm is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Ensemble learning offers a solution to combine the predictive power of multiple learners. In boosting, trees are built sequentially so that each next tree aims to reduce errors from the previous tree. Each tree learns from its predecessors and updates residual errors. Therefore, the tree that grows next in the sequence learns from the updated residuals. The base learners in boosting are weak learners in which the bias is high. Each of these weak learners contributes to give some information for prediction, enabling the boosting technique to produce a strong learner by effectively combining these weak learners. Suppose we have a training data x_i and their labels y_i, XGBoost utilize classifier to predict the final prediction $\hat{y}_i^t$

$$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i) \quad (3)$$

Where $\hat{y}_i^{t-1}$ is previous prediction and $f_t(x_i)$ is new prediction. To get a good model, in XGBoost we need to minimize the following objective function.

$$\mathcal{L}^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f_t) \quad (4)$$

The objective function contains loss function $l(y_i, \hat{y}_i)$ and regularization term $\Omega(f_t)$. With the existing of (3), we can rewrite the objective function as follow.

$$\mathcal{L}^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) \quad (5)$$

The loss function measures how well the model fits on the training data, while regularization measures the complexity of trees. Optimizing loss function encourages predictive models for higher accuracy while optimizing regularization encourages generalized simpler models. Regularization is also utilized to avoid the model from overfitting. To enhance the XGBoost model, we can tune the hyper-parameters which are similar to decision tree hyper-parameters such as learning rate, max depth, n_estimators, and sub-sample. Learning rate and n_estimators are two critical hyper-parameters for gradient boosting algorithms. The learning rate parameter has a role to control the step weight. In other words, it tells us how fast the model learns. While n_estimators is the number of decision trees in XGBoost. If we set it to 1 then it makes the algorithm generate only a single tree. To yield the best performance of XGBoost, the model needs careful tuning of its parameters

.

F. Cross-Validation

Cross-validation, also known as K-Fold Cross Validation, is a statistical method to estimate the quality of machine learning models to predict unseen data. It is a popular evaluation method in machine learning applications because it is easy to understand and gives beneficial results. The general procedure of K-Fold Cross Validation can be written as follow.

1) shuffle the dataset randomly
2) split the dataset into k folds
3) for k iteration, each fold will become validation data, and the rest will become training data, fit the model on training data, evaluate the model on the validation data
4) summarize the quality of the model using the mean of scores from each iteration.

This allows us to see if our model has stable performance overall folds. There might be other problems if there are spikes in high scores or low scores. In this study, we set k=10 because it is widespread to use in machine learning applications.

III. Result and Discussion

This section explains how we performed the preprocessing, feature extraction, modeling, and evaluation.

A. Data Preprocessing

We have applied the preprocessing steps already mentioned in the previous section to the dataset. We can see the result of each preprocessing step in Table 2. After the data were preprocessed, we explored the data by visualizing the most frequent words in Fig. 3-6. The comments later will be helpful to analyze what factors the users complain about the most.

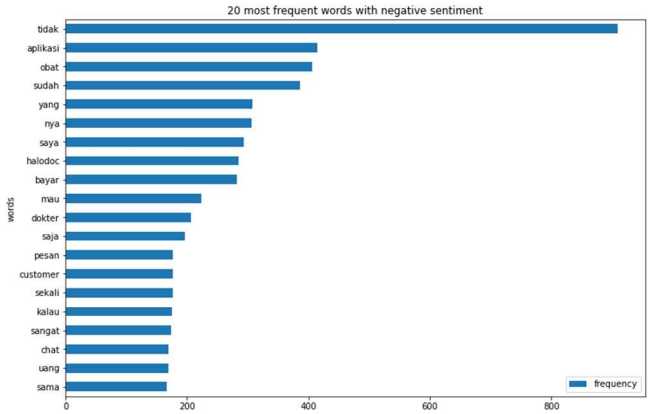


Fig. 3. Most frequent words with negative sentiment

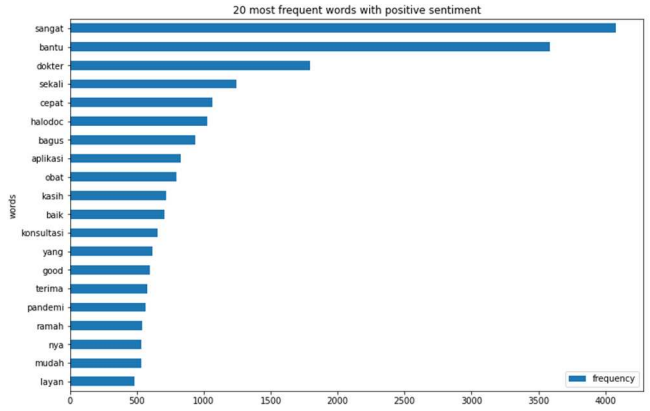


Fig. 4. Most frequent words with positive sentiment

Fig. 5. Wordcloud of negative sentiment words

TABLE II. Sample Result of Data Preprocessing

Data	Result
Case Folding	
Saya kecewa sekali, ya. Konsul gak bisa, diminta cari dokter pengganti, tapi saldo uang terpotong. Padahal saya lagi khawatir kondisi anak saat itu dan butuh penanganan cepat.	saya kecewa sekali, ya. konsul gak bisa, diminta cari dokter pengganti, tapi saldo uang terpotong. padahal saya lagi khawatir kondisi anak saat itu dan butuh penanganan cepat.
Punctuation Removal	
saya kecewa sekali, ya. konsul gak bisa, diminta cari dokter pengganti, tapi saldo uang terpotong. padahal saya lagi khawatir kondisi anak saat itu dan butuh penanganan cepat.	saya kecewa sekali ya konsul gak bisa diminta cari dokter pengganti tapi saldo uang terpotong padahal saya lagi khawatir kondisi anak saat itu dan butuh penanganan cepat
Stopword Removal	
saya kecewa sekali ya konsul gak bisa diminta cari dokter pengganti tapi saldo uang terpotong padahal saya lagi khawatir kondisi anak saat itu dan butuh penanganan cepat	kecewa sekali konsul gak diminta cari dokter pengganti saldo uang terpotong padahal lagi khawatir kondisi anak itu butuh penanganan cepat
Tokenization	
kecewa sekali konsul gak diminta cari dokter pengganti saldo uang terpotong padahal lagi khawatir kondisi anak itu butuh penanganan cepat	kecewa, sekali, konsul, gak, diminta, cari, dokter, pengganti, saldo, uang, terpotong, padahal, lagi, khawatir, kondisi, anak, itu, butuh, penanganan, cepat
Normalization and Stemming	
kecewa, sekali, konsul, gak, diminta, cari, dokter, pengganti, saldo, uang, terpotong, padahal, lagi, khawatir, kondisi, anak, itu, butuh, penanganan, cepat	kecewa, sekali, konsul, tidak, minta, cari, dokter, ganti, saldo, uang, potong, padahal, lagi, khawatir, kondisi, anak, itu, butuh, tangan, cepat

Fig. 6. Wordcloud of positive sentiment words

Based on the visualization above, the most frequent words with negative sentiment are 'tidak,' 'aplikasi,' and 'obat.' The most frequent words with a positive view are 'sangat,' 'Bantu,' and 'dokter.' We created a fishbone diagram to identify the possible cause of the problem, as shown in Fig 7.

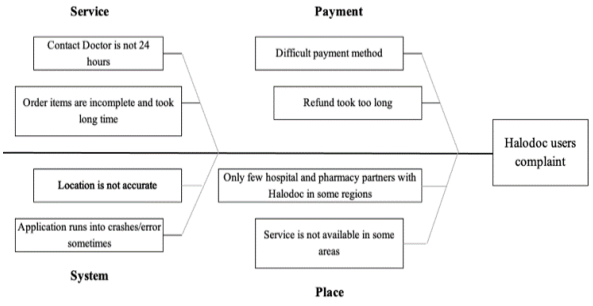


Fig. 7. Fishbone diagram of possible causes of the problem

We found out four main factors the users most complain about; payment, place, service, and system. Users complain about difficult payment methods and refunds that took too long. Users also complain about the service that has not available in some regions, the incomplete order items, and the crashes or errors that sometimes happen on apps.

B. Training

After preprocessing the data, we split our dataset into train data and test data with a proportion of 75:25. We got 8383 rows of train data and 2795 rows of test data. Then, we performed feature extraction using TF-IDF to convert text into vectors. We used the TfidfVectorizer function, which is available in the sklearn library in Python. After extracting features, we defined our model and performed training using our data train. This experiment was conducted in Python using an open-source machine learning library, Scikit-Learn. To enhance our model, we did hyperparameter tuning with the following parameters in Table 3.

TABLE III. LIST OF PARAMETERS AND ITS BEST VALUES

Parameters	Definition	Values	Best Values
learning_rate	Step size	[0.05, 0.1, 0.3, 0.5]	0.1
n_estimators	The number of trees built in the model	[500,1000]	500
max_depth	Maximum number of tree depth	[4, 6, 8]	4

C. Model Evaluation

We evaluated the model using K-Fold Cross-Validation and confusion matrix. The results of each fold of cross-validation can be seen in Table 4. A confusion matrix is a performance measurement for machine learning classification problems, is shown in Table 5. The confusion matrix for our model can be seen in Fig 8. With the confusion matrix, we can get the value of accuracy, precision, recall, and f1-score.

TABLE IV. XGBOOST CLASSIFIER EVALUATION METRICS

Sentiment	Precision	Recall	F1-Score
Negative	0.85	0.77	0.81
Positive	0.97	0.98	0.98

TABLE V. K-FOLD CROSS VALIDATION RESULTS

Iteration	F1-Score
1	97.8%
2	98.5%
3	98.2%
4	97.8%
5	97.8%
6	98.5%
7	97.7%
8	97.5%
9	98.6%
10	97.6%

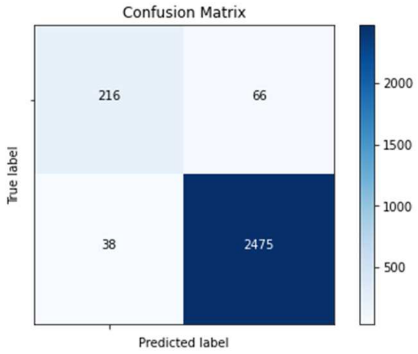


Fig. 8. Confusion Matrix

The equations 4-6 can calculate the three evaluation measures above:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision\ \times\ Recall}{Precision + Recall} \quad (6)$$

IV. CONCLUSION

This study aims to examine public sentiment towards using a telemedicine application in Indonesia, especially during the pandemic of COVID-19. We choose Halodoc since it is the most popular telemedicine application in Indonesia. In this study, we use the XGBoost classifier, which is known as the best algorithm in terms of speed and performance. XGBoost combines weak learners to build strong learners. To enhance the result of XGBoost, we performed hyperparameter tuning using the grid search method by setting the value of the parameter of learning_rate, n_estimators, and max_depth into specific values that have been shown in the previous section. This study proves another assumption about XGBoost, which says XGBoost is quite good at handling imbalanced data. We got 96.24% of accuracy. Overall, the public sentiment towards Halodoc is quite good. People appreciate Halodoc's services, especially during the pandemic of COVID-19. However, we analyzed the negative reviews and found the four main factors Halodoc can improve: payment, place, service, and system. Users complain about the problematic payment method, refund, unavailability of service in some areas, unsatisfying services, and crashes on apps. We suggest using one of the techniques for handling imbalanced data and compare to this study.

REFERENCES

[1] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P Singh, "Analysis of political sentiment orientations on Twitter", Procedia Computer Science, 167, 1821-1828, 2020.

[2] W. Budiharto, and M. Meiliana, M, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis", Journal of Big data, 5(1), 1-10, 2018.

[3] J. Brownlee, A Gentle Introduction to k-fold Cross-Validation, 2018, Accessed: Sept 30, 2021, https://machinelearningmastery.com/k-fold-cross-validation/.

[4] J. Brownlee, Extreme Gradient Boosting (XGBoost) Ensemble in Python, 2020,. Accessed: Sept 30, 2021.

https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/.

[5] D. A. Al-Qudah, A. M. Al-Zoubi, P. A. Castillo-Valdivieso and H. Faris. *Sentiment Analysis for e-Payment Service Providers Using Evolutionary eXtreme Gradient Boosting*. IEEE Access, vol. 8, pp. 189930-189944, 2020, doi: 10.1109/ACCESS.2020.3032216.

[6] B. Gaye, and A Wulamu, "Sentimental analysis for online reviews using machine learning algorithms", International Research Journal of Engineering and Technology (IRJET), *6*(08), 2395-0056, 2018.

[7] R. Hikmat, and N. Dimililer, "SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier", Journal of the Chinese Institute of Engineers, 44:6, 562-572, DOI: 10.1080/02533839.2021.1933598, 2021.

[8] A. Kulkarni, and A. Shivananda, Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. 10.1007/978-1-4842-4267-4, 2021.

[9] G. Kundi, G. How to Scrape Google Play Reviews in 4 simple steps using Python. Accessed: Sept 30, 2021. https://www.linkedin.com/pulse/how-scrape-google-play-reviews-4-simple-steps-using-python-kundi/.

[10] B. Liu, "Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies", 2018, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

[11] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell, "Machine learning and semantic analysis of in-game chat for cyberbullying". ArXiv, abs/1907.10855.

[12] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers", Future Generation Computer Systems, 106, 92-104, 2016.

[13] M. S. Saputri, R. Mahendra, and M. Andriani, M. "Emotion Classification on Indonesian Twitter Dataset", Proceeding of International Conference on Asian Language Processing 2018.

[14] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," in IEEE Access, vol. 7, pp. 163677-163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[15] J. Singh, G. Singh, and R. Singh, "Optimization of sentiment analysis using machine learning classifiers", Hum. Cent. Comput. Inf. Sci. **7,** 32, 2017

[16] B. N. Supriya, and C. B. Akki, "Sentiment Prediction using Enhanced XGBoost and Tailored Random Forest", International Journal of Computing and Digital Systems, 2021.

[17] L. Yang Y. Li, J.Wang, J., and R. S. Sherratt, "Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning, IEEE Access, 8, 23522-23530, 2020.

Sentiment Analysis of YouTube Video Comments with the Topic of Starlink Mission Using Long Short Term Memory

Ayu Masyita Putri
Research Center for Artificial Intelligence and Big Data
Department of Chemistry
Padjadjaran University
Sumedang, Indonesia
ayu18004@mail.unpad.ac.id

Daniel Ananda Putra Basya
Department of Computer Engineer
Telkom University
Bandung, Indonesia
danielapb@student.telkomuniversity.ac.id

Muchammad Teguh Ardiyanto
Department Robotics and Artificial Intelligence Engineering
Airlangga University
Surabaya, Indonesia
muchammad.teguh.ardiyanto-2020@stmm.unair.ac.id

Indra Sarathan
Faculty of Cultural Studies
Padjadjaran University
Sumedang, Indonesia
sarathan@unpad.ac.id

***Abstract*— The launch of the Starlink Satellite by a private company from the United States (SpaceX) has become a hot topic for conversation. SpaceX is a project that aims to provide satellite-based internet services worldwide with high performance and affordable prices even in remote places. The satellite launch video was shared on the YouTube platform. Many comments were on the satellite launch video, and various reactions were written on the video's comments column. Therefore, in this study, sentiment analysis will be carried out to analyze the responses of internet users through comments on YouTube to the launch of the Starlink satellite using deep learning. The YouTube comment data used in this study amounted to 22,000 comments. The model used is Long Short Term Memory (LSTM). This study will produce an accuracy value of the LSTM model by applying different activation and optimization functions. This study shows that the highest accuracy is 86% using the LSTM model, Softmax activation function, and Adam's optimization.**

Keywords*— *Deep Learning, LSTM, Sentiment Analysis, Starlink Mission, YouTube

I. INTRODUCTION

The internet has become an inseparable part of human life in this digital era. Young people and adults use it to help carry out various activities effectively and efficiently in times of pandemic like now, which requires people to minimize direct contact with other people, making the existence of the internet even more helpful to support various activities ranging from work, education, communicating with the closest people without having to meet in person. Reporting to *We Are Social* [1], internet users worldwide have reached 4.66 billion. That number is up to about 316 million (7.3) percent from the previous year. That number covers about 59.5 percent of the earth's total population, reaching 7.83 billion people. One example of the use of the internet is the YouTube platform. All of us can upload any and as many videos on this YouTube platform if we already have a registered account. The whole world can also view the uploaded videos.

There is a comment feature; it allows users to express opinions or comments on a video subject online. However, due to differences in audience psychology, educational level, and theme preferences, there is uncertainty in demand for online videos. This uncertainty is often found in video comments due to differences in opinions of YouTube viewers. In addition, video comments also leave a digital communication trail, which other users can read as a reference for video content. Therefore, the reputation of a YouTube video can be measured by the comments on the video [2].

The Starlink company is the name of a satellite network developed by the private aerospace company SpaceX to provide high-performance, affordable satellite-based internet to remote locations. And this time SpaceX is using the YouTube platform to broadcast a video of the launch of the Starlink satellite. The first satellite launch was conducted on May 23, 2019, using a Falcon 9 rocket containing 60 satellites. The launch mission was broadcast through the YouTube channel, which attracted much attention and various reactions from internet users worldwide.

Starlink is a revolutionary project like never before. A breakthrough project needs feedback and comments from the community for project development. By analyzing the input and opinions of internet users worldwide through YouTube comments, it can be a reference and material for the company's consideration to develop and improve the project further. Community feedback is essential for creating a project in the future. Consequently, it is necessary to conduct sentiment analysis to study YouTube video comments.

To analyze the reactions of internet users to the launch of the Starlink satellite on videos with the keyword Starlink Mission, sentiment analysis can be used. Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [3].

The algorithm used in this sentiment analysis is LSTM (Long Short Term Memory). The LSTM deep learning method was chosen for several considerations, including the LSTM has good accuracy for data in the form of text, and the LSTM is a development of the RNN (Recurrent Neural Network) deep learning method, which has the advantage of

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

being able to process relatively long data (long-term dependencies) [4].

Previously, several studies have been carried out on LSTM, including by Hasan & Mahmood (2017), regarding the classification of sentence sentiment using the Recurrent Neural Network (RNN) with a single layer LSTM and the word2vec model generate satisfactory results with an error rate of 14.3% from the SSTb dataset and error of 11.32% on IMDB dataset. Superior to some of the other methods used [5].

another study, Ciftci & Apaydin (2018) conducted a Turkish-language sentiment analysis using the Recurrent Neural Network with LSTM. They considered that the more commonly used naive Bayes and logistic regression methods were not optimal. The experimental results show that the LSTM method has better accuracy than other methods with a validation value of 83% accuracy, testing accuracy of 82.9%, the precision of 0.86%, and recall of 0.83 [6]. Research related to sentiment analysis has also been investigated, starting from Twitter comments [7], movie reviews [8], hotel reviews [9] and also stock closing pricing prediction [10].

II. LITERATURE STUDY

A. Sentiment Analysis

Sentiment analysis is the process of managing and also understanding sentiment information related to a particular topic where the topic involves grouping comments into categories such as "positive" or "negative" [10]. Sentiment analysis can be a handy tool for checking affinity for a brand, product, or domain [10]. Currently, the implementation of sentiment analysis covers the fields of product marketing, health care, finance, elections, politics, sports, film reviews, and tourism and hospitality which aims to extract opinions and comments on specific issues or explore the line of thought of certain groups based on various textual data online [11].

B. Long Short Term Memory (LSTM)

All LSTMs are one type of RNN where modifications are made to the RNN by adding a memory cell that can store information for an extended period [11]. The LSTM network contains an iterative processing model that signals from one step to another. Based on the statement of the LSTM-RNN study by the new method for semantic consistency verification of aviation radiotelephony communication, that LSTM has been successfully applied to various sequential tasks and modeling languages [12]. LSTM has three types of gates, including forget gates, input gates, and output gates. The forget gate is the gate that decides which information to delete from the cell. The input gate is the gate that determines the value of the input to be updated into state memory. The output gate is a gate that determines what output will be produced according to the feedback and memory in the cell [13].

The LSTM model can overcome the vanishing gradient problem, common in the relatively long data processing. A cell in LSTM stores a value or state (cell state) for a long or short period. LSTM has a memory block that will determine which value will be selected as the relevant output for the given input. This is an advantage possessed by LSTM [14].

C. Activation Function

The activation function is used in neural networks to activate or deactivate neurons [15]. The activation process is to decide whether the artificial neuron should be activated or not. This helps the neural network learn complex patterns in the data and helps normalize each neuron's output to a range between 1 and 0 or between -1 and 1 [15]. Examples of activation functions are Rectified Linear Unit (ReLU), Softmax, Sigmoid, and Tanh.

1) Sigmoid

The sigmoid function has a typical S shape and is a real function that can be derived for any real input value [11]. Sigmoid will accept a single number and convert the value of x into a deal range from 0 to 1. Eq. 1 for the sigmoid function:

$$(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

2) Tanh

The output of this Tanh function is centered at zero with a range of -1 to 1[15]. Therefore, optimization is more accessible and preferable to the sigmoid activation function. However, still have the missing gradient issue. Eq. 2 for the Tanh function:

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2)$$

3) ReLU

ReLU basically only creates a limit on the number zero, meaning that if $x \leq 0$ then $x = 0$ and if $x > 0$ then $x = x$. Compared with sigmoid and tanh, the calculation is simpler, and ReLU can only be achieved by limiting activation to zero [15]. Another important advantage of ReLU is that it avoids or fixes the missing gradient problem. Eq. 3 is the formula for the ReLU function:

$$\sigma(x) = \{(0, x) \ , x \geq 0 \ \ 0 \qquad , x < 0 \quad (3)$$

4) Softmax

The Softmax function calculates the probability of each target class over all possible target classes and will help determine the target class for the given input [15]. The main advantage of using Softmax is that the output probabilities range from 0 to 1, and the sum of all probabilities will equal one. If the softmax function is used for a multi-classification model, it will return the possibility of each class, and the target class will have a high probability. Eq. 4 for the Softmax function:

$$\phi(z_j) = \frac{e^{z_j}}{\sum_j^k \quad z_j} \quad (4)$$

D. Optimization

Optimization is the key to how networks learn. Learning is an optimization process [15]. This optimization algorithm can be defined as an algorithm that aims to find the value of x. This value of x can produce f(x), which is as small or as large as possible for a given function f, accompanied by some limitations on the value of x, where the value of x can be a scalar or a vector of continuous or discrete values—several examples of optimization algorithms such as Adam, AdaGrad, and RMSProp.

E. Accuracy

Accuracy describes how accurate the model is in classifying correctly [16]. It can be described in the confusion matrix shown in Fig 1.

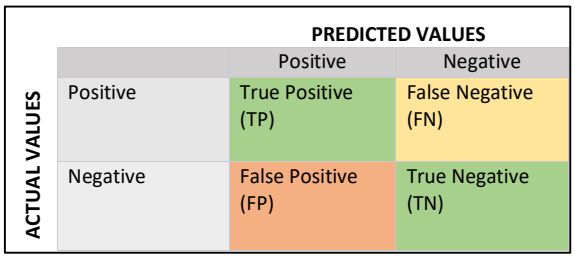


Fig. 1. Confusion Matrix

Accuracy is also the ratio of True (positive and negative) predictions to the overall data. Eq. 5 is the accuracy.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (5)$$

III. METHODOLOGY

The research steps are as follows in Fig.2 shows the workflow carried out in this analysis.

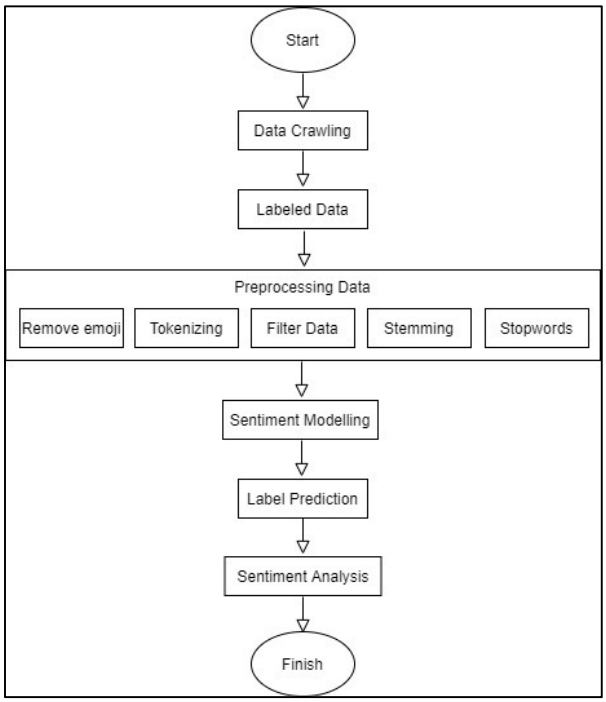


Fig. 2. Workflow

A. Collecting Data

The dataset used is obtained from www.youtube.com, with the keyword Starlink Mission with data coverage obtained from videos in the first three pages of the search with the comment field activated. The dataset contains comments from each of these videos.

The crawling takes video comment data from the YouTube platform with the keyword Starlink Mission. The data taken are only comments in English. The crawled data can be seen in Table 2. Comments are taken from videos on the first three pages of searches with the keyword Starlink Mission. Words fetched only the main comment ignore the chain of reply comments. The features on the YouTube comment data can be seen in Table 1 and the examples in Table 2.

TABLE I. DATASET FEATURES

No	Features	Description
1	Video_id	Video ID uploaded to YouTube
2	Title	The title of the video uploaded
3	Comment	Contains comments
4	Class	Which class category

TABLE II. SAMPLE DATASET COMMENT YOUTUBE

Video_id	Title	Comment	Class
4372QYiPZB4	Starlink Mission	Wait, starlink couple year the download speed suck get ani faster due telecom company to want cover hous middle road line stop half a mile away hous sign Starlink ago wait give anoth send equip pleas hurri spacex	-1
4372QYiPZB4	Starlink Mission	kamu baik sengat ita fafilah	-1
4372QYiPZB4	Starlink Mission	kamu tungu fadilah aku bg kamu rasa duduk penjara sebatan sebab sendudukkn dgn sky warga cina laki org sambil isap dadah depan ank se ka	-1
4372QYiPZB4	Starlink Mission	tungu laka fadilah	-1
4372QYiPZB4	Starlink Mission	aku nk bg sku rasa duduk penjara kena sebatan kerana fadilah bersedudukkn dengan laki org warga cina sky ank nya kerajaan ambil	-1
4372QYiPZB4	Starlink Mission	doe spacex like public spotifi playlist pre stream sound track mental sooth	1
4372QYiPZB4	Starlink Mission	whi signal alway loss stage near land platform	1
4372QYiPZB4	Starlink Mission	love liftoff land everyth	1
4372QYiPZB4	Starlink Mission	took deposit month ago get us starlink ye wizard tech come save us se oklahoma thi deadzon live	1
…..	…..	……	……

B. Preprocessing Data

After the comment data is obtained, the data is labeled automatically using the TextBlob library. TextBlob is a Python library for processing textual data which contain simple API for Natural Language Processing (NLP) [17]. Once labeled, comments need to be cleaned as video comment data contains irrelevant letters, numbers, emojis and symbols. The LSTM model has been shown to perform well in learning context-free and context-sensitive languages [18].

The data that is still not clean will be preprocessed in several stages, including:

1) *Remove Emoji*
   Comments containing emoji (emoticons) will be removed by the program.
2) *Tokenizing*
   Comment sentences are broken down into tokens based on punctuation. The tokens used in this study are: tok=WordPunctTokenizer()
3) *Filter data*
   Data words and sentences comments '@', 'link(https?://)', '#', repeat words, symbols and lowercase are selected and then removed.
4) *Stemming*
   Data commentary in each sentence is changed to the basic form or deletion of affixes and suffixes.
5) *Stopwords*
   Words in comments that are too general and less important will be removed. The characteristics of the word to be deleted are the words that appear quite a lot compared to the other words.

C. Model Deep Learning

After going through the preprocessing stage, the data is trained with a deep learning model using Long Short Term Memory (LSTM). In this research, the LSTM model uses the input sequence form. Sequence Analysis is used to find patterns in a series of events called Sequence.

Then the next layer is the first layer, namely the LSTM embedding layer, which functions to train text data into a numeric vector which represents the closeness of the meaning of each word. By adjusting the size of the embedded vector and also in this layer, feature extraction will be carried out where each word in the dataset will be searched for its vector weight which will then be classified as negative and positive in sentiment.

The LSTM layer in this study uses 100 LSTM units and a dropout of 0.2. The dropout parameter is added to reduce the risk of overfit. The dropout value ranges from 0 to 1. The commonly used range is 0.2 to 0.5. The closer to 0 it tends to be overfit, while getting closer to 1 has the risk of underfitting. The algorithm also applied several optimization algorithms such as Adam, AdamGrad, and RMSProp. In addition, this LSTM uses activation functions such as ReLU, Sigmoid, Tanh, and also Softmax.

The output layer with the activation function aims to transform the input value for each neuron. The activation function will give a signal to each neuron in the output layer to actively transform the input value (the result of calculating the weight operation in the previous layer) into a value with a smaller range. Loss function to measure the effectiveness of the model in making predictions on each epoch (iteration).

IV. ANALYSIS AND RESULT

This research begins with the process of collecting data by performing a crawling process on YouTube video comments where the results of this crawling process get data as much as 20,000 comments. The data is labeled with a negative labels and a positive labels. After getting the data, the data preprocessing process is carried out where the destination is carried out and after that the data is processed using the Long Short Term Memory (LSTM) model.

The LSTM model used the input sequence form. The LSTM layer in this research used 100 LSTM units and a dropout of 0.2 using 10 epoch. The number of epochs is a hyperparameter that determines how many times the learning algorithm will work to process the entire training dataset. Epoch is one of the hyperparameters that affect model's performance. When the amount of epoch increased, model generalization towards training dataset will increase. Choosing the right amount of epoch is important because if the epoch number is too big, the probability to overfitting is getting higher and lessening the model generalization. To find out the results of the test, in this research the evaluate model was used to find out the results of the test in the form of accuracy values and loss values.

TABLE III. MODEL TEST RESULT

NO	ACTIVATION	OPTIMIZER	EPOCH	ACCURACY (%)
1.	RELU	ADAM	20	79,85%
2.	TANH	ADAM	20	80,35%
3.	SOFTMAX	ADAM	20	86 %
4.	SOFTMAX	RMSPROP	20	84%
5	SOFTMAX	ADAGRAD	20	54,7%
6	SIGMOID	ADAGRAD	20	54,45%
7	SOFTMAX	RMSPROP	10	80,9%
8	SOFTMAX	ADAGRAD	10	55,4%
9	SIGMOID	ADAGARD	10	54,7%
10	RELU	ADAM	10	69,3%
11	SOFTMAX	ADAM	10	80,2%

LSTM able to overcome long term dependence on the input. Based on Table 3, the results of the LSTM model test are found and it is known that the best accuracy is obtained when training data using the Softmax activation function with the Adam optimizer which reaches 86% compared to other activations and optimizers. The main advantage of using Softmax is that the output probabilities range from 0 to 1, and the sum of all probabilities will equal one. If the softmax function is used for a multi-classification model, it will return to the probability of each class and the target class will have a high probability. While the Adam optimization algorithm is a replacement optimization algorithm from stochastic gradient descent for training models in deep learning that combines the best properties of the Adagrad and RMSProp algorithms to provide a more optimal algorithm, which can handle gradients that spread and have noise. Adam is relatively easy to configure, with the default configuration parameters working well for most problems generated during training. LSTM has a memory block that will determine which value to choose as output relevant to the input given. It causes this method to be suitable to be applied in this research.

V. CONCLUSION

From the experiments conducted, it can be concluded that sentiment analysis using the LSTM method, Softmax activation function and Adam's optimization algorithm obtains an accuracy of 86%. Softmax activation function has the advantage of classification model data by returning the probability of each class and the target class will have a high probability and the Adam optimization algorithm has the advantage of being able to configure well, where the default configuration parameters work well in most problems generated during training. The high accuracy results using LSTM, Softmax activation function and Adam's optimization algorithm, indicate if the method is suitable for use in this study.

YouTube is one of the most popular social media in the world. This can be an effective tool to find out the responses and reactions of internet users around the world. By analyzing the responses of people from all over the world, their comments and reactions to videos on YouTube relating to these projects, it can be a basis for companies to find out what they like and what should be improved and improved in the future. Effectiveness and accuracy can be increased and expanded with greater data processing and development of LSTM methods.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] "Digital 2021: the latest insights into the 'state of digital' - We Are Social UK." https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/ (accessed Nov. 10, 2021).

[2] Z. Li, R. Li, and G. Jin, "Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary," *IEEE Access*, vol. 8, pp. 75073–75084, 2020, doi: 10.1109/ACCESS.2020.2986582.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–184, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[4] N. K. Manaswi, "Deep Learning with Applications Using Python," *Deep Learn. with Appl. Using Python*, 2018, doi: 10.1007/978-1-4842-3516-4.

[5] A. Hassan and A. Mahmood, "Deep learning for sentence classification," *2017 IEEE Long Isl. Syst. Appl. Technol. Conf. LISAT 2017*, Aug. 2017, doi: 10.1109/LISAT.2017.8001979.

[6] B. Ciftci and M. S. Apaydin, "A Deep Learning Approach to Sentiment Analysis in Turkish," *2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018*, Jan. 2019, doi: 10.1109/IDAP.2018.8620751.

[7] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A Combined CNN and LSTM Model for Arabic Sentiment Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11015 LNCS, pp. 179–191, Aug. 2018, doi: 10.1007/978-3-319-99740-7_12.

[8] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis," *Multimed. Tools Appl.*, vol. 78, no. 18, pp. 26597–26613, Sep. 2019, doi: 10.1007/S11042-019-07788-7.

[9] R. A. Priyantina and R. Sarno, "Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity and LSTM," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 142–155, 2019, doi: 10.22266/IJIES2019.0831.14.

[10] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Comput. Appl.*, vol. 32, no. 13, pp. 9713–9729, Jul. 2020, doi: 10.1007/S00521-019-04504-2.

[11] I. N Yulita, S. Purwani, R. Rosadi, and R. M Awangga, "A quantization of deep belief networks for long short-term memory in sleep stage detection", 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA) (pp. 1-5), 2017..

[12] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 3, pp. 517–529, Mar. 2015, doi: 10.1109/TASLP.2015.2400218.

[13] Y. Lu, Y. Shi, G. Jia, and J. Yang, "A new method for semantic consistency verification of aviation radiotelephony communication based on LSTM-RNN," *Int. Conf. Digit. Signal Process. DSP*, vol. 0, pp. 422–426, Jul. 2016, doi: 10.1109/ICDSP.2016.7868592.

[14] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," *undefined*, 2015.

[15] W. Di, A. Bhardwaj, and J. Wei, "Deep learning essentials : your hands-on guide to the fundamentals of deep learning and neural network modeling," Accessed: Nov. 10, 2021. [Online]. Available: https://www.perlego.com/book/578845/deep-learning-essentials-pdf.

[16] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," *AAAI Work. - Tech. Rep.*, vol. WS-06-06, pp. 1015–1021, 2006, doi: 10.1007/11941439_114.

[17] S. Lorla, "TextBlob Documentation," *TextBlob*, p. 69, 2020.

[18] P. Goyal, S. Pandey, and K. Jain, "Deep Learning for Natural Language Processing," *Deep Learn. Nat. Lang. Process.*, 2018, doi: 10.1007/978-1-4842-3685-7.

Indonesian Food Price Prediction with Adaptive Neuro Fuzzy Inference System

Sheila Azhar Almufarida[ab1], Rahma Batari[ab2], Akik Hidayat [b3],Anindya Apriliyanti Pravitasari[ac4]
[a]*Research Center for Artificial Intelligence and Big Data,* Universitas Padjadjaran
[b]*Department of Computer Science, Faculty of Mathematics and Natural Science*, Universitas Padjadjaran
[c]*Department of Statistics, Faculty of Mathematics and Natural Science,* Universitas Padjadjaran
Sumedang, Indonesia
e-mail: [1]sheila18001@mail.unpad.ac.id, [2]rahma18002@mail.unpad.ac.id, [3]akik@unpad.ac.id, [4]anindya.apriliyanti@unpad.ac.id

***Abstract*— Rising food price in Indonesia is an uncertain issue that significantly affects the lives of the people and the country's economy. In addition, the demand for Indonesian staple foods, namely rice, is very high. Of course, the fluctuations in the price of rice will have a significant impact on the Indonesian population. So, we need a system that functions to predict food prices. Therefore, this study aims to predict Indonesian food prices. Rice price data is taken from the Central Bureau of Statistics (BPS) website and the prediction was built using the *Adaptive Neuro-Fuzzy Inference System* (ANFIS) method. Then the model will be evaluated using *Mean Squared Error* (MSE) and *Mean Absolute Percentage Error* (MAPE). In this study, the ANFIS model is made with the number of membership functions (MFs) as much as 2, the number of iterations (epochs) is 300, and the type of membership function (MF type) is gaussmf. The training and testing results using ANFIS show that the MSE and MAPE values are 0.9176 and 0.70059%. These results prove that the ANFIS method can predict Indonesian food prices well.**

Keywords— Prediction, Food, Rice, Indonesia, ANFIS.

I. INTRODUCTION

Indonesia's economic growth is very influential on the need for food which is also accompanied by the development of the population. Based on Law Number 18 of 2012, food is anything that comes from biological sources of the agricultural, plantation, forestry, fishery, animal husbandry, water, both processed and unprocessed, which are intended as food or drinks for human consumption, including food additives, food raw materials, and other materials used in the process of preparing, processing, and making food or beverages.

One of the leading foods in Indonesia is rice. Rice is the primary source of calories for most Indonesians. The portion of rice in total calorie consumption is 54.3 percent, so half of the calorie intake comes from rice—no wonder the demand for rice in Indonesia is enormous. Rice consumption in Indonesia is very high compared to other Asian countries, with an average of fewer than 100 kilograms per capita per year [1]. The rice consumption per capita in Indonesia is increasing rapidly in line with high population growth [2].

Fulfilling food needs, especially rice, is very important to maintain the sovereignty of a nation, so we need to ensure that all Indonesian people can meet their food needs. So, by predicting the price of Indonesian food, especially rice, Indonesian people can prepare themselves and other things when there is a problem of rising and falling rice prices in the future.

The previous study that made predictions using the ANFIS method was Betul and Teoman [3]. They applied the ANFIS method to predict building energy needs in the early design stage, which results in a reasonable degree of accuracy of 96.5% and 83.9% for heating and cooling, respectively. Then there was Yordanos and Zheng [4], which results in forecast errors for model testing being lower than 10%, and the average MAPE model test is 6.88%. Then there was Raafat et al. [5], which resulted in the ANFIS model being better than Mamdani and Sugeno. The researchers believed the ANFIS model is perfect for forecasting the fund's price in periods of severe economic fluctuations. Thus, some of this previous research has inspired this study to predict Indonesian food prices using the ANFIS method, which is expected to produce accurate results with a small error value.

By utilizing technology, we can make predictions using the *Adaptive Neuro Fuzzy Inference System* (ANFIS) method with existing data samples. The application of the ANFIS is quite broad, and it ranges from commercial applications to medical imaging [6]- [7]. In forecasting time series methods, data availability is a must for better results and good modeling. The use of the ANFIS method is due to several advantages, such as the ability to capture the non-linear structure of a process, adaptability, and fast learning capacity [8]. The main objective is to formulate an efficient method for selecting optimally available data from the existing dataset to build an ANFIS model with the precise prediction of future values.

II. RESEARCH METHODS

Our research used a dataset on Indonesian food prices, especially rice, sourced from the site *https://www.bps.go.id/indicator/20/295/12/rata-harga-rice-di-level -the big-wholesale-indonesia.html* year 2010 - 2020 every month as many as 132 data. The dataset contains 12 columns (January - December) and 11 rows (2010 - 2020). Then the dataset is used as training data, 80% of which is 108 data, and 20% of the test data is 24 data.

This research was conducted using Matlab R2018a software, with the following computer specifications:

- Laptop ASUS A409FJ-EK702T
- Processor Intel Core i7-8565U
- RAM 8 GB DDR4
- Memory 1 TB HDD

From the diagram in Fig 1 can be seen the flowchart of prediction using ANFIS method that the early stages of prophecy ANFIS in this research defines input parameters such as membership function type selection (MF *type*), number of iterations (*epochs*), and number of membership functions (MFs). Then carry out the process *training* and *testing* until the best data model is found, then the model will be used as a reference in the prediction process. Then validate the performance of ANFIS using the calculation *Mean*

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

Squared Error (MSE) and *Mean Absolute Percentage Error* (MAPE).

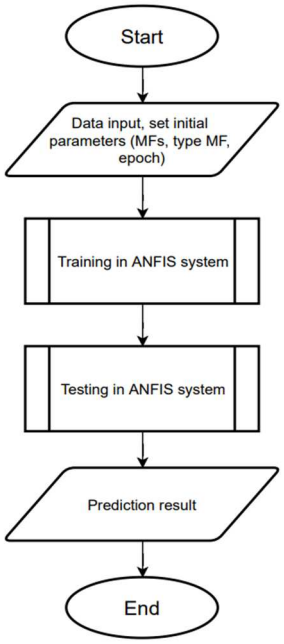


Fig. 1. Flowchart of ANFIS Prediction

A. Determining Initial Input Parameter

To produce a good ANFIS model, it is necessary to use the correct input parameters. Therefore, three input parameters were tested initially: MF type, epochs, and MFs value. In this research, the dataset will be tested with 4 MF types (trapmf, trimf, gaussmf, gbellmf), the number of iterations (*epochs*) is 100, 200, 300, 400, 500, and the number of MFs is 2, and 3. To determine the appropriate initial ANFIS parameters to produce the smallest MAPE value. After obtaining the correct initial input parameters values, the ANFIS model is created.

B. Adaptive Neuro-Fuzzy Inference System

This study used the Adaptive Neuro Fuzzy Inference System method to make a prediction system. *Adaptive Neuro-Fuzzy Inference System* or ANFIS is an adaptive network-based on fuzzy inference system. ANFIS is divided into two parameters, namely premise parameters and algorithms *hybrid*. JSR Jang [9] stated that the algorithm hybrid is a combination of learning methods Least Square Estimation (LSE) and backpropagation, which are also carried out in two steps, namely forward and backward steps.

The first stage of fuzzy logic modeling determines the membership function of input-output variables, the second is the construction of fuzzy rules, and the last is the determination of output characteristics, output membership functions, and system results. To determine membership input-output variable function, referred to as backward propagation algorithm, and hybrid-learning algorithm providing ANFIS learning and rule construction is used [10]. Functionally, ANFIS has the same architecture as *fuzzy rules based* on Sugeno's model [4]. So ANFIS is a method in which a learning algorithm is used when setting rules on a set of data that allows the authorities to adapt. Fig 2 is a form of ANFIS architecture consisting of five layers with different functions, namely:

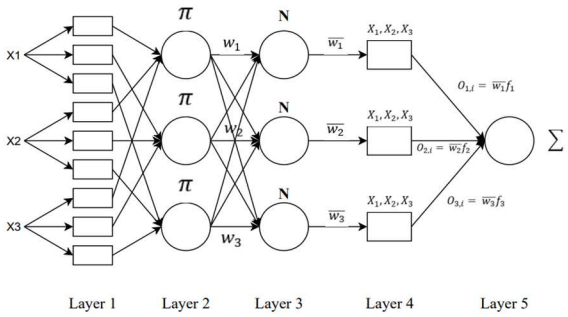


Fig. 2. ANFIS Architecture

1) Fuzzification Layer (Layer 1)

In the fuzzification layer, each node i in the fuzzification layer is an adaptive node with the following node functions.

$$O_{1,i} = \mu_{Ai}(x), i = 1,2 \quad (1)$$

Where x and y is the i signal input, A_i is the language label, and $O_{1,i}$ is the membership level of the fuzzy set.

Then, suppose we use the Generalized-Bell function to calculate the degree of membership in order to obtain the premise parameters.

$$\mu A_i(x) = \frac{1}{1 + \left|\frac{xc}{a}\right|^{2b}} \quad (2)$$

Where a and c are the set of attributes and $\mu_{Ai}(x)$ is the degree of membership.

2) Rule layer (Layer 2)

The nodes in the rule layer are labeled π non-adaptive to generate the ignition strength of each node in this layer.

$$O_{2,i} = w_i = \mu_{Ai}(x)\Delta\mu_{Bi}(y), i = 1,2 \quad (3)$$

Where $\mu_{Ai}(x)$ is the degree of membership of the set and $\mu_{Bi}(x)$ is the degree of membership of the set.

3) Normalization Layer (Layer 3)

The nodes in this layer are labeled "N" are non-adaptive to normalize the ignition strength of each node displaying the normalized activation degree of the form:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2 \quad (4)$$

Where $\overline{w_i}$ is the normalized firing strength and $w_1 + w_2$ is the output.

4) Defuzzification Layer (Layer 4)

The defuzzification layer functions to calculate the output of existing rules based on variables *consequent*. Each node is an adaptive node with the function:

$$O_{4,i} = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \quad (5)$$

Where $\overline{w_i}$ is and $p_i x, q_i y, r_i$ is the set parameters on fuzzy with Sugeno's first-order model.

5) Output Layer (Layer 5)

Each node in this layer is a fixed node which is the sum of all inputs.

$$O_{5,i} = \sum_i \overline{w_i}\, f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (6)$$

C. Metric of Evaluation

A number of criteria can be used to evaluate the performance of Indonesian food prices prediction. In this study, to calculate the accuracy of the ANFIS model is use MSE (*Mean Squared Error*) and MAPE *(Mean Absolute Percentage Error*).

1) Mean Squared Error (MSE)

This is a criterion that minimizes the variance of the error distribution. MSE of proposed ANFIS model is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2 \quad (7)$$

Where n is the number of data, Y_i is predicted values, and $\overline{Y_i}$ is target values [11].

2) Mean Absolute Percentage Error (MAPE)

This criterion evaluates the prediction error by calculating the actual data with the data obtained from the prediction results. The following statistical performance with MAPE calculation according to the equation:

$$MAPE = \frac{\frac{\Sigma|Y_i - \overline{Y_i}|}{Y_i}}{n} x\ 100\% \quad (8)$$

Where Y_i is the actual data, $\overline{Y_i}$ is the prediction result, and n is the number of data.

III. Results And Discussion

In the process of predicting food prices, especially rice in Indonesia, it considers the input data that has been obtained previously. The input dataset is divided into 80% training data, and 20% test data. The training data contains the price of rice in 2010 - 2018 as many as 108 pairs of data. And the test dataset contains the price of rice in 2019 - 2020 as many as 24 pairs of data. The input features of the ANFIS neural network in this study are the dataset of Indonesian rice prices for 2010-2020, and several other input parameters such as MFs, type MF, and epochs. So, to get the best prediction results, we determine initial parameters by testing them one by one.

A. Determining Initial Parameters

1) Comparison of Membership Function (MF) types

To determine the appropriate initial parameters of ANFIS in this study, the researcher first compares the 4 types of MF that will be used namely trapmf, trimf, gaussmf, and gbellmf.

Table I shows that the MF type gaussmf has the smallest test MAPE value of 0.700593% compared to other MF types. This can happen because the gaussmf curve shape is not rigid, more flexible, and has a smooth curve [12], so the MF type gaussmf will be a reference in the next test.

TABLE I. RESULT OF COMPARISON OF MF TYPES

MF Type	MAPE
trapmf	0.753456 %
trimf	3.986296 %
gaussmf	0.700593 %
gbellmf	1.360458%

2) Comparison of Number of Iterations (epochs)

Next, compare the number of iterations (*epochs*). In this study, we will compare the number of iterations (*epochs*) of 100, 200, 300, 400, and 500.

TABLE II. COMPARISON TEST RESULTS OF THE NUMBER OF EPOCHS

Epoch	MAPE
100	1.111551 %
200	1.273034 %
300	0.700593 %
400	0.848106 %
500	1.228445 %

Table II shows that the number of iterations (*epochs*) 300 produces the smallest MAPE value compared to the number of other iterations tested. This means that the model will be optimal if it uses the number of iterations of 300. So the MF type gaussmf and the number of iterations (*epochs*) 300 will be the reference in the next test.

3) Comparison of Number of Membership Functions (MFs)

The next test compares the number of membership functions (MFs). The MFs values that will be reached in this study are 2 and 3.

From Table III it can be seen that the number of *Membership Function* (MFs) 2 produces the smallest MAPE value compared to the value of MFs 3. So the value of MFs 2 will be used in this study. From the results of the comparison above, it is found the initial parameters that will be used in predicting Indonesian food prices with the ANFIS method are the MF type gaussmf, the number of epochs 300, and the number of MFs 2.

TABLE III. COMPARISON TEST RESULTS VALUE OF MFs

Total MFs	MAPE
2	0.700593 %
3	1.842617 %

B. Training Results

The input parameters used are the MFs value or membership function 2, type MF gaussmf, and the number of iterations (*epochs*) 300. In Fig 3 is the result of training from price data rice for 2010 - 2018. It can be seen that the resulting graph of training results is very similar to the original target graph. From this training, the MSE and MAPE values were 0.4054 and 0.316 %, respectively. So that the results of the training using the ANFIS algorithm can be continued to the testing stage.

Fig. 3. Graph of Rice Price Training Results with ANFIS

C. *Test Results*

After the data is trained, then the test data containing the 2019 - 2020 rice prices are tested using the previously trained ANFIS model. The test results can be seen in Fig 4. The graph in Fig 4 can be seen that the test results are almost the same in shape as the original target and the results of this test produce MSE and MAPE values are 0.9176 and 0.70059%, respectively. This result is quite good because it has a smaller MAPE value than previous studies regarding the prediction of water quality parameters using ANFIS by Armin, et al [13], which shows a MAPE value of 13,43%. So from the results of this test it can be said that it is optimal in predicting food prices, especially rice using the ANFIS method.

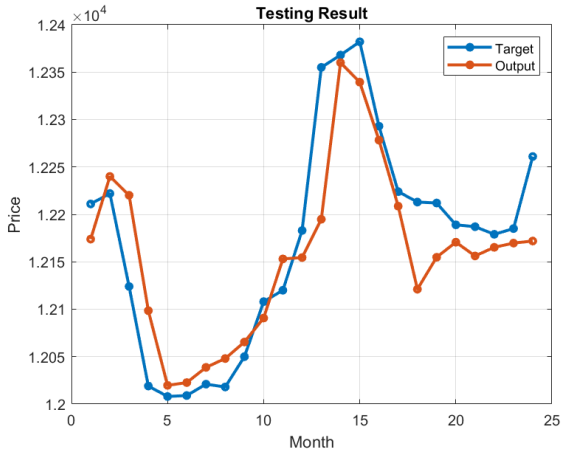


Fig. 4. Graph of Rice Price Test Results with ANFIS

D. *Indonesia Food Price Prediction 2021*

After obtaining optimal test results, the ANFIS model created can predict rice prices in Indonesia quite well. Then the next researcher tries to predict food prices, especially rice in Indonesia, in 2021. In Fig 5, the results of predicting food prices, especially rice in Indonesia in 2021 using the ANFIS method that has been made, produce a price graph that tends to decrease from January to December.

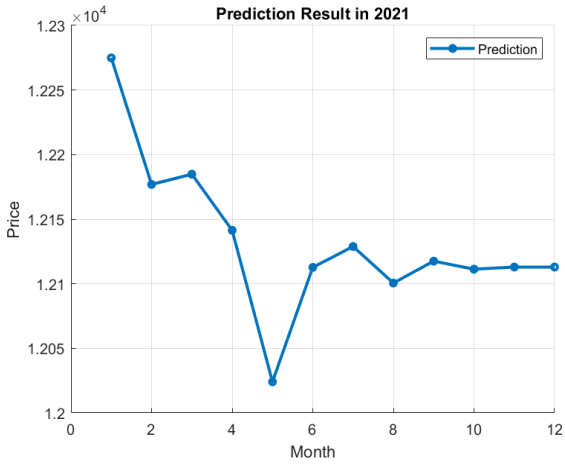


Fig. 5. Graph of Rice Price Prediction Results in Indonesia in 2021

IV. Conclusion

This study aimed to predict Indonesian food prices using *Adaptive Neuro Fuzzy Inference System* (ANFIS) method. And the following is a summary of conclusions:

1. Initial input parameter that produces the smallest MAPE value in this study is the value of MFs is 2, the number of iterations (*epoch*) is 300, and MF type is gaussmf.
2. The training data results using the ANFIS method resulted in the MSE and MAPE values being 0.4054 and 0.316 %, respectively.
3. The ANFIS method can predict the price of rice in Indonesia by producing the MSE and MAPE values of 0.9176 and 0.70059%, respectively.
4. Prediction of food prices, especially rice in Indonesia in 2021 using the ANFIS method, produces a price graph that tends to decrease from January to December.

Acknowledgment

The Author thanks the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

References

[1] B. Arifina, N. A. Achsanib, D. Martiantoc, L. K. Sarid and A. H. Firdause, "The Future of Indonesian Food Consumption," *Jurnal Ekonomi Indonesia,* vol. 8, no. 1, pp. 71-102, 2019.

[2] T. N. Maraseni, R. C. Deo, J. Qu, P. Gentle, and P. R. Neupane, "An international comparison of rice consumption behaviours and," *Journal of Cleaner Production,* pp. 1-13, 2017.

[3] B. B. Ekici and U. T. Aksoy, "Prediction of building energy needs in early stage of design by using ANFIS," *Expert Systems with Applications,* vol. 38, no. 5, pp. 5352-5358, 2011.

[4] Y. Kassa, J. H. Zhang, D. . H. Zheng, and D. Wei, "Short Term Wind Power Prediction Using ANFIS," *IEEE International Conference on Power and Renewable Energy,* p. 388–393, 2016.

[5] R. Fahmy, H. Zaher and A. E. Kandil, "A Comparison between Fuzzy Inference Systems for Prediction (with Application to Prices of Fund in Egypt)," *International Journal of Computer Applications,* vol. 109, no. 13, pp. 6-11, 2015.

[6] R. K. Meleppat, E. B. Miller, S. K. Manna, P. Zhang, E. N. P. Jr. and R. J. Zawadzki, "Multiscale Hessian filtering for enhancement of OCT

angiography images," in *SPIE 10858, Ophthalmic Technologies XXIX*, San Francisco, California, United States, 2019.

[7] R. K. Meleppat, M. V. Matham and L. K. Seah, "Optical frequency domain imaging with a rapidly swept laser in the 1300nm bio-imaging window," in *International Conference on Optical and Photonic Engineering (icOPEN2015)*, Singapore, 2015.

[8] M. Sahin and R. Erol, "A Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games," *Mathematical and Computational Applications,* vol. 22, no. 43, 2017.

[9] J.-S. R. Jang, "ANFIS : Adap tive-Ne twork-Based Fuzzy Inference System," *IEEE Transactions On Systems, Man, And Cybernetics,* vol. 23, no. 3, pp. 665-685, 1993.

[10] M. Firat and M. Gungor, "River flow estimation using adaptive neuro fuzzy inference system," *Mathematics and Computers in Simulation,* vol. 75, pp. 87-96, 2007.

[11] G. Ozkana and M. Inal, "Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems," *Applied Soft Computing,* vol. 24, pp. 232-238, 2014.

[12] N. Talpur, M. N. M. Salleh and K. Hussain, "An investigation of membership functions on performance of ANFIS for solving classification problems," in IOP Conf. Series: Materials Science and Engineering, Melaka, 2017.

[13] A. Azad, H. Karam, S. Farzin, A. Saeedian, H. Kashi and F. Sayyahi, "Prediction of Water Quality Parameters Using ANFIS Optimized by," KSCE Journal of Civil Engineering, pp. 1-8, 2017.

Sentiment Analysis of Indonesia's National Health Insurance Mobile Application using Naïve Bayes Algorithm

Natalia Syafitri Kustanto[ad1], Intan Nurma Yulita[bd2], Indra Sarathan[cd3]
[a]*Department of Mathematics, Universitas Padjadjaran*
[b]*Department of Computer Science, Universitas Padjadjaran*
[c]*Faculty of Cultural Studies, Universitas Padjadjaran*
[d]*Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]natalia18001@mail.unpad.ac.id, [2]intan.nurma@unpad.ac.id, [3]indra.sarathan@unpad.ac.id

Abstract— **Indonesia is developing one of the biggest single-payer social health insurance systems in the world, aiming to cover its entire population of over 260 million. To further expand membership and improve services, the National Health Insurance Administrator (BPJS Kesehatan) launched Mobile JKN, a mobile application that allows people to register, pay monthly contributions, set appointments with healthcare providers, and more. In this paper, sentiment orientation is explored considering the positive and negative sentiments using Indonesian user reviews for Mobile JKN application to discover user perception against service quality of Mobile JKN. Naive Bayes classifier method is used. The accuracy obtained from the model is 93%. Then the model is applied to analyze the sentiment of user review for Mobile JKN application version 3 and above. The results indicate that Mobile JKN version 3 and above tend to receive negative reviews rather than positive from users. Positive reviews do not refer to the specific features of the application. Whereas in negative reviews, many users complain about application updates that are too frequent and problems when logging in and signing in related to phone number verification.**

Keywords—Naïve Bayes, user reviews, Mobile JKN, Sentiment Analysis

I. INTRODUCTION

Indonesia is developing one of the biggest single-payer social health insurance systems in the world, aiming to cover its entire population of over 260 million. Indonesia's national social health insurance scheme (*Jaminan Kesehatan Nasional*, JKN) unified the previously segmented health insurance schemes into one organization under BPJS Kesehatan [1]. Since its implementation in 2014, JKN has extended health insurance coverage from less than 50 percent to more than 80 percent of the population. In an effort to further expand membership and improve services, the National Health Insurance Administrator (BPJS Kesehatan) launched Mobile JKN, a mobile application that allows people to register, view billing information, pay monthly contributions, select or change the primary healthcare provider, and set appointments with healthcare providers, all from their cellular devices [2]. Since its launch in November 2017, Mobile JKN has been downloaded by nearly 10 million users on the Android version.

Mobile JKN is constantly exploring ways to improve user engagement through the application by providing updates of improvement; including adding new features, fixing issues, and improving the user interface. The updates have a subjective point associated from the user. The subjective points of view are review, proposals, remarks, appraisals and individual experience shared by various users. Several studies have been used to look at user perceptions of the Mobile JKN application. That is by conducting a survey to users of the Mobile JKN application [3] [4] and analyses the success factors of the application [5]. The results of these studies provide good input for the development of the Mobile JKN application, but takes time and resources to do it.

Sentiment classification is useful to extract the sentiment or opinion from the review. For example, online product reviews are usually analyzed to decide what products are the best to produce in the future to reduce risk. Machine learning technologies, such as Naïve Bayes are widely used in sentiment classification. Because it doesn't require as much training data, is fast, and can be used to make real-time predictions. Some previous works using Naïve Bayes method are on analyzing e-commerce applications [6] [7], movie reviews [8], and Twitter data [9] [10]. Based on that, this paper uses Naïve Bayes Classifier in aim to discover user perception against service quality of Mobile JKN version 3 and above. The result is intended to be taken into consideration or recommendation for future development and optimization on application.

II. METHODOLOGY

Fig. 1 shows the workflow that will be carried out in this analysis.

A. Data Crawling

The data used in this research is secondary data. Data was taken using the scraping process from the Mobile JKN review column on the Google Play Store site. The sample used is the review from 2020 to October, 3rd 2021. In crawling, the google-play-scraper library is used to retrieve reviews on applications with parameters only in Indonesian language reviews.

B. Pre-processing

The data gathered is formatted but not structured and cannot be used in the next step as it is. The data must have a preprocessing step to clean up the data from the noise to

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

become smaller and structured [11]. The preprocessing process to clean the reviews carried out in this sentiment analysis is as follows:

- Case folding aims to convert all letters in the document to lowercase. Only letters 'a' to 'z' are accepted. Characters other than letters are omitted and are considered delimiters.
- Tokenization stage is the stage of cutting sentences into a list of words that make up the sentence separated by commas and spaces so that the results are single words that are collected in the form of array data which will later be used in the weighting process.
- Filtering is the stage of taking important words from the token results by using a stoplist algorithm (discarding less important words) or wordlist (saving important words). Stopwords are common words that usually appear in large numbers and are considered meaningless. Examples of stopwords in Indonesian are "yang", "dan", "di", etc. For filtering in the sentiment analysis of this project, colloquial-indonesian-lexicon [12] and stopwords-id-satya [13] is used.
- Stemming is the process of removing the inflection of a word to its basic form (there is no affix to the word). For example the words "listened"/"*mendengar*", will be transformed into the word "hear"/ "*dengar*". In this sentiment analysis, Python Sastrawi library is used.

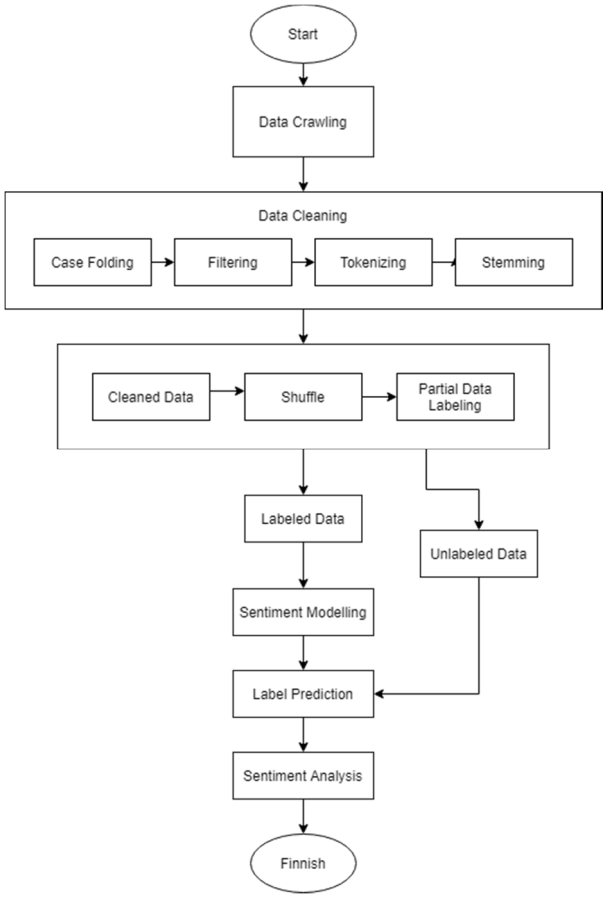


Fig. 1. Workflow

C. Data Labeling

There are two labels used, namely "Negative" and "Positive" denoted with -1 and 1 respectively. A "Negative" label is a review that contains bad words, ridicule, or contraries. The "Positive" label is review data containing words of kindness, praise, agreement, or support. Labeling was done by the author.

D. Naïve Bayes Classifier

Naïve Bayes Classifier is a classification using statistics based on the Bayes theorem. General principle of Naïve Bayes Classifier assumes that value from an attribute does not depend on and influences other attributes [6].

Suppose there are m possible classes $C = \{c_1, c_2, \dots, c_m\}$ for documents $D = \{d_1, d_2, \dots, d_n\}$. Let $W = \{w_1, w_2, \dots, w_s\}$ be the set of unique words, each of which appears at least once in D. The probability of a document d being in class c can be computed using Bayes rule:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (1)$$

Word occurrence in a document is not noticed in Naïve Bayes. Therefore, the probability calculation can be calculated as:

$$P(c|d) = \frac{\prod_k P(w_k|c)P(c)}{P(w_1,w_2,w_3,\dots,w_k,\dots,w_n)} \quad (2)$$

By making a probabilistic model from the training document, the classification process can be carried with calculating $p(w_k|c)$ value. Thus, it can be found the probability for all values using (1) and (2) model [14].

$$P(w_k|c) = \frac{D_b(w_k|c)}{D_b(c)+|V|} \quad (3)$$

with $|V|$ is the amount of w_{kj} value possibility and $D_b(c)$ a function that restores the amount of b document which has c category.

Giving category in a text document can be done by choosing the maximum of c value in $P(c|\ d)$ value, such as

$$c^* = argmax_{c \in C} \prod_k P(w_k|c)P(c) \quad (4)$$

In classification, Naïve Bayes method is used and divided into two stages of the process, namely the training process and the testing process. Sentiment is determined by calculating the opportunity value or probability of document data testing by looking at the opportunities or probabilities of the training data.

Classification with Naïve Bayes is done using the sklearn package in python, conceptually this function compares the weight of each word in the test data based on the training data. The data will be added up between the positive and negative word weights and then seen if the weight of a document is greater than the positive probability then the sentiment is positive, and vice versa.

III. RESULT

A. Data

The data taken is approximately 20 thousand data. A total of 7139 reviews for the application version 3 and above is separated for later analysis. Then the other review data (13,417) are labeled with a positive label of 6337 and a negative label of 7080. The implementation examples are in Table 1.

TABLE I. SAMPLE OF RAW DATA

Sample Reviews	English Translation
Aplikasi yang sangat membantu, mempermudah transaksi tanpa harus mengantri, mobile jkn semudah telapak tangan memberi solusi ketika diperlukan	A very helpful application, simplifying transactions without having to queue, mobile JKN is as easy as the palm of the hand to provide solutions when needed
Mantap sangat membantu untuk pengecekan kartu BPJS kesehatan nya sehingga kita bisa tahu aktif atau tidak nya	Great, very helpful for checking the BPJS health card so we can know whether it is active or not
Entah jaringan saya atau emang aplikasi ini yang tidak bisa dibuka, sering update tapi susah untuk buka aplikasi	I don't know if it's my network or this application can't be opened, it updates frequently but it's hard to open the application

B. Pre-Processed Data

Table 2 is a comparison of example reviews that have gone through the text pre-processing stage.

TABLE II. SAMPLE OF PRE-PROCESSED DATA

Raw Reviews	Pre-processed Reviews
aplikasi yang bgus membantu masyarakat	*bagus bantu masyarakat*
Tiap buka minta update terus. Mohon untuk di perbaiki	*buka minta update terus mohon untuk perbaiki*
Capeeeeee update mulu,, sekalinya dibuka malah tidak bisa kebuka itu aplikasi dh kaya jam eror muter mulu ... gimana siiiiiii ???	*lelah update terus sekalinya dibuka tidak kebuka deh kayak jam eror muter terus sih*

These sentences have undergone changes in the pre-processing stage of the text. Where the case folding, tokenization, filtering, and stemming processes have been carried out.

C. Naïve Bayes Classification Model

The total data used for training and testing were 13,417 data, with a positive label of 6337 and a negative label of 7080. The training and test data used were split and shuffled, respectively 7: 3 (9391 and 4026). The following is a summary of the model using the Naive Bayes Classifier.

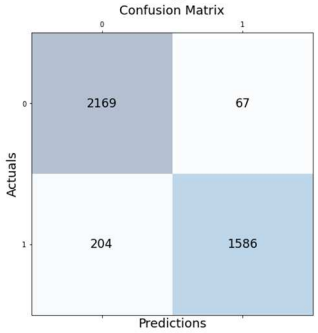


Fig. 2. Confusion Matrix of classification model

A confusion matrix is generated with 2169 true positive data, 1586 true negative data, 204 false negative data (type 2 error), and 67 false positive (1st error type) in Fig. 2.

TABLE III. MODEL EVALUATION SUMMARY

Model Summary	
Accuracy	0.93
Precision	0.94
Recall	0.93
f1-score	0.93

The accuracy of model is 0.93 or 93% in Table 3, which means that by using the model created, it can predict the sentiment classification of the Mobile JKN application review with an accuracy of 93%. The precision value obtained (positive predictive value) 0.94 or 94%, meaning with the model above the level of accuracy between the data requested and the prediction results provided by the model is 94%. The recall (true positive rate) is 0.93 or 93%, which means that the success of the model in finding information is 93%. Also, the comparison value of the average precision with the weighted recall value or also known as the F1-Score in the model above is 0.93 or 93%.

D. Label Prediction

The data used as analysis comes from Mobile JKN reviews on the Google Play Store specifically version 3 and above. By using the prediction of the Naïve Bayes Classifier model, the following results are obtained in Table 4. From the label prediction result, there are two predictive values, namely negative (-1) or positive (1) which will be used to analyze the results of public sentiment.

TABLE IV. SAMPLE OF MODEL PREDICTION RESULTS

Reviews	English Translation	Label
lebih mudah dan cepat	easier and faster	1
update harus 1 minggu sekali, giliran mau di gunakan di buka aja tidak bisa	updates must be once a week, but when I want to use it, I can't even open the app	-1

IV. Discussion

Based on the prediction results using the model, it produces results that can be visualized as follows in Fig. 3.

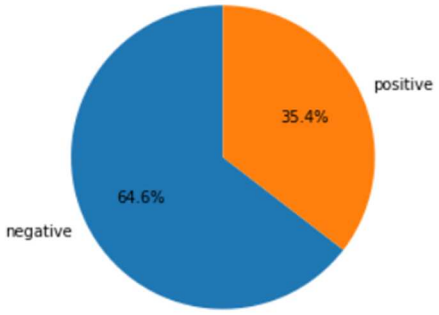


Fig. 3. Comparison of Sentiment Reviews for Mobile JKN Application Version 3 and above

From the percentage and data visualization on Fig. 3, Mobile JKN Application for Version 3 and above has 35.4% positive reviews and 64.6% negative reviews out of a total of 7139 data reviews. From the results of this percentage, Mobile JKN for Version 3 and above has a negative impression according to users.

A. Positive Reviews Data

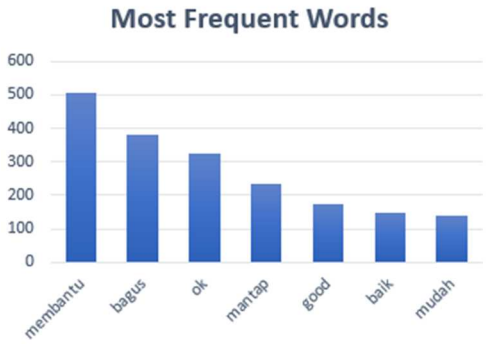


Fig. 4. Most Frequent Words on Positive Labeled Reviews

Based on Fig. 4 the term often used by Mobile JKN version 3 and above users in positive class is the term "*membantu*"/"help" 506 times, "*bagus*"/"good" 381 times, "ok" 326 times, "*mantap*"/"great" 233 times, et cetera. Overall, positive reviews don't really refer to special aspects of the features in the Mobile JKN application.

B. Negative Reviews Data

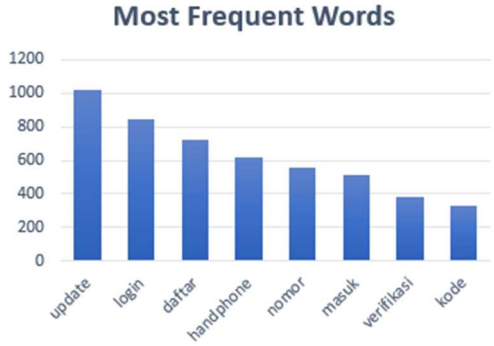


Fig. 5. Most Frequent Words on Negative Labeled Reviews

Based on Fig. 5 the term often used by Mobile JKN version 3 and above users in negative class is "update" 1015 times, "login" 847 times, "*daftar*"/"register" 808 times, "handphone" 617 times et cetera. Unlike positive reviews, in negative reviews users refer to certain features that concern them. What users often complain about is the mobile application updates that are done too frequently, problems when signing in or logging in to the application that possibly relates to verification via phone number.

V. Conclusion

In this study, sentiment classification was carried out using the Naive Bayes Classifier method on data reviewing the Indonesian national health insurance application, Mobile JKN. The mobile application review data is obtained from Google Play Store. The raw data is then cleaned and separated to be analyzed and used for modeling. The data used for modeling is labeled positive or negative, then training is carried out using the Naive Bayes method. From the training, we obtained a model with high accuracy. The model is then used to classify reviews for Mobile JKN version 3 and above. The result is that Mobile JKN version 3 and above tend to receive negative reviews rather than positive from users. Positive reviews do not refer to the specific features of the application. Whereas in negative reviews, many users complain about application updates that are too frequent and problems when logging in and signing in related to phone number verification. This result is intended to be taken into consideration or recommendation for future application development and optimization.

Acknowledgment

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

References

[1] S. Satriana, "Easing Access to the National Health Insurance through a Mobile Application Indonesia," Social Protection Floors in Action, pp. 1-4, 2018.

[2] BPJS Kesehatan, "Access Service At Your Palms, BPJS Kesehatan Launches JKN Mobile Application Which Has A Lot of Benefits and Facilitates JKN-KIS Participants," BPJS Kesehatan, 15 November 2017. [Online]. Available: https://bpjs-kesehatan.go.id/bpjs/index.php/post/read/2017/596/Akses-Pelayanan-Dalam-Genggaman-BPJS-Kesehatan-Luncurkan-Aplikasi-Mobile-JKN-Banyak-Manfaat-dan-Mudahkan-Peserta-JKN-KIS. [Accessed 4 October 2021].

[3] F. Ichsani and S. Hartono, "Analysis of satisfaction of national health insurance jkn participants through quality of mobile services and confidence mediated by decision to choose mobile jkn service on bpjs kesehatan of pekanbaru," Dinasti International Journal of Education Management And Social Science, vol. 1, no. 3, pp. 262-269, 2020.

[4] O. Hapsara, Gupron and A. Yandi, "Improving the Satisfaction of National Health Guarantee Participants through JKN Mobile Services That Educed by Brand Image (Study on Jambi City Health BPJS)," Saudi Journal of Economics and Finance, vol. 3, no. 5, pp. 230-236, 2019.

[5] P. W. Handayani, D. A. Meigasari, A. A. Pinem, A. N. Hidayanto and D. Ayuningtyas, "Critical success factors for mobile health implementation in Indonesia," Heliyon, vol. 4, no. 11, 2018.

[6] V. Oktaviani, B. Warsito, H. Yasin, R. Santoso and Suparti, "Sentiment analysis of e-commerce application in Traveloka data review on Google Play site using Naïve Bayes classifier and association method," Journal of Physics: Conference Series, 2021.

[7] S. W. Handani, D. I. S. Saputra, Hasirun, R. M. Arino and G. F. A. Ramadhan, "Sentiment Analysis for Go-Jek on Google Play Store," Journal of Physics: Conference Series, 2019.

[8] R. Novendri, A. S. Callista, D. N. Pratama and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," Bulletin of Computer Science and Electrical Engineering, vol. 1, no. 1, pp. 26-32, 2020.

[9] J. Song, K. T. Kim, B. Lee, S. Kim and H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," KSII Transactions on Internet and Information Systems (TIIS), vol. 11, no. 6, pp. 2996-3011, 2017.

[10] H. Parveen and S. Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016.

[11] R. Feldman and J. Sanger, The Mining HandBook : Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[12] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri and A. Jamal, "Colloquial Indonesian Lexicon," International Conference on Asian Language Processing (IALP), pp. 226-229, 2018.

[13] S. Nugraha and A. Chandra, "github stopwords-bahasa-indonesia," 10 December 2017. [Online]. Available: https://github.com/datascienceid/stopwords-bahasa-indonesia. [Accessed 1 October 2021].

[14] T. Mitchell and M. Hill, Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression, Helena: Morgan Kauffman Publisher, 2010.

Forecasting Inflation in Indonesia using Long Short Term Memory

Farah Fauziah Savitri[ab1], Reski Febrianti Siregar[bc2], Feevrinna Yohannes Harianto[bd3], Herlina Napitupulu[be4]
[a]*Department of Statistics, Universitas Padjadjaran*
Sumedang, Indonesia
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Sumedang, Indonesia
[c]*Department of Information System, Universitas Dinamika Bangsa*
Jambi, Indonesia
[d]*Department of Statistics, Universitas Airlangga*
Surabaya, Indonesia
[e]*Department of Mathematics, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]farah18014@unpad.mail.com, [2]reskifebrianti62@gmail.com, [3]feevrinna.yohannes.harianto-2018@fst.unair.ac.id, [4]herlina@unpad.ac.id

***Abstract*— The problem of inflation has always been a macro problem that occurs in developing countries, especially Indonesia. Observation of the inflation rate cannot be done by observing only certain years because the problem of inflation is a long-term problem. The case of inflation in Indonesia has become quite important to note since Indonesia adopted the inflation target system. Modeling and forecasting the inflation rate is needed and considered important because it is related to poverty alleviation where people with low incomes or still are required to be able to meet their needs with high prices of goods. One of the forecasting models that can be used is the Long Short Term Memory (LSTM). This model is a development of the previous model Recurrent Neural Network (RNN). The results showed that the best model is the model with 5 nodes in hidden layer, Adam optimizer and 0.01 learning rate. Forecasting results show that until September 2022 Indonesia's inflation rate condition will increase but not significantly increase so that it remains stable below 2% until September 2022 and still classified as mild inflation.**

Keywords*— *Inflation Rate, LSTM, RNN

I. INTRODUCTION

The problem of inflation has always been a macro problem that routinely occurs in developing countries, especially Indonesia. According to the Indonesian Central Statistics Agency or Badan Pusat Statistika (BPS), Inflation is a tendency to increase the prices of goods and services in general, which continues from one period to the next period. Inflation is caused by an increasing volume of the medium of exchange or liquidity in the market resulting in high demand. It can trigger changes in the price level. If the demand for goods and services increases, the demand for production also increases, so prices rise [1]. In Indonesia, Bank Indonesia (BI) is the institution that has responsibility for the inflation rate. BI has the authority to issue monetary policy to regulate business activities in the country and maintain stability in the value of the rupiah. Inflation is one of the macroeconomic variables that the government and economic actors pay the most attention. Inflation has a significant and positive effect on Human Development Index and poverty for the long term. Inflation also has a significant and positive relationship with BI rate, exchange rate, world oil price, and gold price [1]. In addition, the impact that appears such as economic instability, rising interest rates, and inflation can make people's income unequal so that it can increase unemployment. When inflation increases, the rupiah exchange rate will depreciate against the US dollar, where the value of the rupiah needed to earn one US dollar will increase.

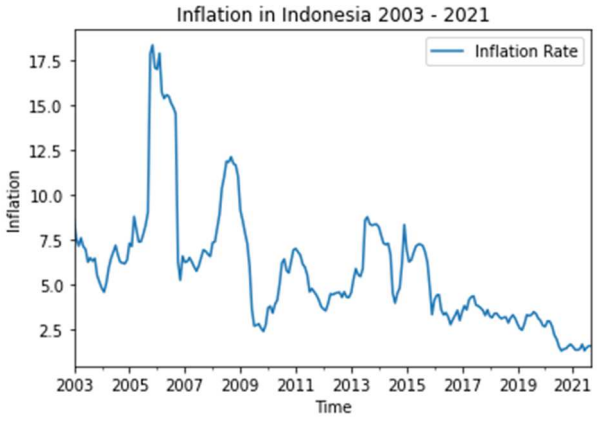


Fig. 1. Plot inflation rate in Indonesia Januari 2003 – September 2021

Fig. 1 is graph of inflation rate condition in Indonesia. The data on the graph is obtained from Bank Indonesia. Based on these data, the inflation rate in Indonesia reached the highest level in November 2005 because there was an increase in fuel prices. Inflation rates in Indonesia have decreased since the COVID-19 pandemic began to spread to Southeast Asia in early 2020.

COVID-19 affects the economic condition in Indonesia and globally. Demand and supply in commodity markets experienced global shocks due to the COVID-19 pandemic [2]. The World Bank predicts that the COVID-19 pandemic will disrupt the growth of developing countries in East Asia, Asia Pacific, and China [3]. The world bank also predicts that the economic growth of developing countries in East Asia and Asia Pacific region will slow down to 2.1% in the baseline scenario and 0.1% in the lower-case scenario. The Indonesian

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

Central Statistics Agency or Badan Pusat Statistik (BPS) stated that Indonesia is one of the countries that experienced a contraction in the economy in several trading partners due to the spread of COVID-19 around the world.

The inflation rate in Indonesia has not shown a significant increase and tends to decline to below 2% from June 2020 to September 2021. The COVID-19 pandemic situation has an alarming impact with sharp changes in demand and the closure of several industries. Companies are also less able to increase demand by lowering prices because consumers are less able to shop. The COVID-19 pandemic also entered a second wave with a peak in July 2021 and inflation has decreased as a result of restrictions on public activities.

Research on forecasting inflation in a country has received positive attention from macroeconomic researchers. Most central banks use inflation as one of the considerations for taking the monetary policy. Monetary policy is taken into consideration the value of future inflation. For the government, inflation forecasting is a connecting to determine the value of future inflation. This research is the development of inflation forecasting in Indonesia that can provide input for Bank Indonesia as consideration for policymaking.

Observation of the inflation rate cannot be done by observing only certain years because the problem of inflation is a long-term problem. The case of inflation in Indonesia has become quite important to pay attention to since Indonesia adopted the inflation target system. The system is carried out by announcing the target numbers to be achieved in the next few years to the public so that stable inflation becomes a common and main goal in monetary policy.

Forecasting time series data with modern methods has continued to grow rapidly in the last few decades. The modern method continues to be applied and developed in various methods that can offer high accuracy, good validity, and ease of calculation [4]. One of the forecasting models that can be used in time series data is the Long Short Term Memory (LSTM). Inflation data forecasting using the LSTM method in this study is expected to provide good accuracy values because the LSTM method is considered suitable for modeling sequential data and complex information.

Previous study using LSTM shows that the Deep Learning methodology with LSTM can help in pricing to determine the fair value of derivatives related to the inflation index. LSTM supports identifying seasonal effects better than traditional standard methodologies [5]. Also, LSTM can predict the rupiah exchange rate for the next five days in the previous study [6] with a model using 50 epochs using five hidden neurons. This model can produce a prediction model that is closest to the existing pattern data. Other studies in forecasting daily air temperature for five years in Bandung can be modeled with LSTM and Facebook Prophet. At the minimum air temperature the LSTM model works better than Facebook Prophet model [7]. In this study we proposed LSTM method to predict monthly inflation rate data using early stopping so that epoch in each LSTM model can be different. We also compare LSTM method with classical method ARIMA in inflation rate prediction since November 2020 to August 2021 to prove that prediction result using LSTM tend to reach smaller error.

II. METHODOLOGY

The data used in this study is monthly historical data regarding year to year inflation rate in Indonesia from January 2003 to September 2021. This historical data has 225 periods collected from the Bank Indonesia website. (https://www.bi.go.id/id/statistik/indikator/data-inflasi.aspx).

All analysis is done in open-source Python 3.7 using Google Collaboratory also the help of TensorFlow and several other packages. The device used has a 64-bit operating system type, x64-based processor using windows 10 operating system. Based on previous considerations, the artificial network method used in this study is LSTM. The artificial network in the process is not based on a perfect stochastic model and can be done for time series data that has a nonlinear relationship between lag times, therefore LSTM does not require any assumption, unlike the classical time series method.

An artificial neural network (ANN) is a computer program designed to simulate the way the human brain processes information biologically [8]. Artificial Neural Network collects information by detecting patterns and relationships in the data through a learning process and based on experience. Deep Learning is a derivative of Artificial Neural Network methods that can study data patterns with more complex construction.

Neural networks can generalize information in data conditions that have noise and uncertainty. Neural Network can generalize to statistical data components because Neural Network can make its representation through the learning process. The basic architecture of Neural networks is Feed-Forward Neural Network and Back Propagation Neural Network. Feed-Forward Neural Network is a cycle of information in one direction through the nodes and connections between neurons in a layer do not form cycles, implying that information only propagates from the input level to the output level [9]. Back Propagation Neural Network (BPNN) is a learning algorithm that refers to the Deepest-Descent technique. BPNN will minimize errors if provided with a suitable number of hidden units in complex cases. The backpropagation method involves a feedforward process from the input training pattern, then error calculations with the backpropagation process, and weight adjustments in synapses [10].

A recurrent Neural Network is a new kind of Neural Network architecture considered a network with memory because it obtains information from input in the previous time period that affects the current input and output. The RNN architecture has at least one feedback loop so that it can store data memory in the network structure. RNN is an artificial network architecture that is used to process sequential data and requires to storage of sufficient dependencies on each other's inputs over the long term.

RNN has drawbacks related to the emergence of vanishing gradient and gradient exploding problems that occur during the data training process. Vanishing Gradient is a situation where the gradient value used to update the weight value is zero, which is one of the shortcomings of the RNN method. This can happen when the number of training data periods increases. The development of an RNN that can overcome these problems includes the Gated Recurrent Unit (GRU) developed by Cho et al. in 2014 and Long Short Term

Memory (LSTM) developed by Hochreiter and Schmidhuber in 1997.

Long Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that can store important information over time using memory cells. LSTM cells consist of a combination of multiple gates which are more complicated and have a network of memory blocks capable of managing data with long dependencies. LSTM solves the problem of vanishing gradients and exploding this gradient through three gates. The first gate is the forget gate which regulates the amount of information that must be deleted, the second gate is the input gate which regulates the nominal of the cell states that should be stored, the third gate is the exit gate which regulates the amount of cell state should be sent to the next cell [11]. Besides being able to combine current information with previous information, LSTM is efficient at recording long information.

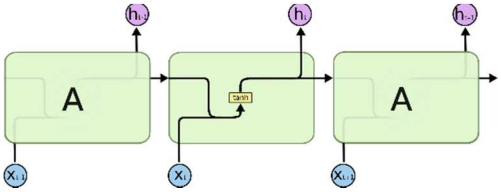


Fig. 2. Network Architecture of RNN [12]

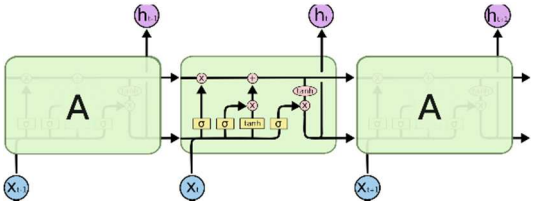
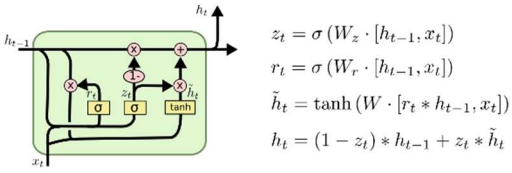


Fig. 3. Network Architecture of LSTM [12]



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Fig. 4. Network Architecture of GRU [12]

Fig. 2, Fig. 3 and Fig. 4 shows the differences between RNN, GRU, and LSTM. GRU and LSTM as the development of RNN have several differences that can be a weakness or strength of each model. LSTM solves problems in RNN through three gates. Those are the forgotten gate, input gate, and output gate while the GRU only uses two gates, which are the reset gate and the update gate [12]. Reset gate sits between the previous activation function and the next candidate to clear out old data, and an update gate that determines how frequently the candidate activation function is utilized to update the cell state GRU uses fewer training parameters so the training process uses less memory than LSTM [13]. This results in a faster training process than LSTMs. Despite this simpler process, the LSTM can outperform the GRU by producing a more accurate model than the GRU when using long data periods [14].

The flow of information shows in Fig. 3 from the forget gate to the output gate in LSTM. Information as input will go through the forget gate as the first gate (1). The input for the forget gate, which is the hidden state at time $t-1$ and the input at time t, is entered through the sigmoid activation function. After going through the forget gate then proceed with the input gate (2). The input gate has the same input and output as the forget gate because it has the same activation function. The result of the input gate is multiplied by the result of the intermediate cell state in (3). The cell state at $t-1$ after the multiplication operation, using the forget gate will be updated with the output gate. This is done by adding operations to obtain a new cell state (4). The new cell state will come out as input for $t+1$. The multiplication between the new cell state and the gate output is done using the activation function tanh (5) to get the hidden state in (6) output.

$$f_t = \sigma\left(W_f.[h_{t-1}, x_t] + b_f\right) \quad (1)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c.[h_{t-1} + x_t] + b_C) \quad (3)$$

$$c_t = (i_t * \tilde{C}_t + f_t * c_{t-1}) \quad (4)$$

$$O_t = \sigma(W_o.[h_{t-1}, \mathrm{xt}] + b_o) \quad (5)$$

$$h_t = O_t * tanh(c_t) \quad (6)$$

The methodology used in this research is given in detail as follows.

A. Normalization

Data preprocessing process in this study by normalizing the dataset using the minimum-maximum scaler technique from sklearn library. Normalization with minimum-maximum scaler is by changing the real value or actual value to a value with a range interval as in (7) . The range interval value in this study is between 0 and 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

B. Split Data

Before we forecast historical data using LSTM, the data will be separated into training data and testing data. The training data serves to find the optimal model through the learning process from the training data, while the testing data functions to evaluate the model that has been obtained at the training stage. In this study, the proportion of 70: 30 will be used for training data and testing data.

Before entering the model, the training data and testing data are reshaped into X and Y or one time period after t because predictions using LSTM are supervised learning. Furthermore, the input in the form of X data is reshaped into the three-dimensional form [samples, time steps, features]. This study uses twelve time steps or lag variables because we found seasonal pattern on lag variables and one feature because it wants to predict one time period.

C. Training Process

The LSTM model used consists of memory units and a dense layer that often follows the LSTM layer. The dense layer can be used to generate predictions. The LSTM layer in this study consists of the number of neurons used in the hidden

layer and the input form. The smaller number of neurons in the hidden layer, the fewer calculations will be performed in each unit and fewer features will be generated. It can be applied to simple cases. Furthermore, in the dense layer, there is one neuron with an activation function.

After we define our network then we have to compile it. The compilation is the way to efficiency by converting a defined sequence into a simple assessment in an efficient format for execution. At the compilation stage, we specifically use optimizers that are used to train the network and loss functions that are used to develop networks and reduced by optimization. In the case of regression, the loss function which usually used is the mean squared error.

After compiling the network, a fit model is obtained that can adjust the training on the dataset. The network uses a backpropagation algorithm that can be optimized according to the specified optimization. The loss function is defined when constructing the model. This study uses early stops with 100 epochs so that the model can stop training at epochs less than 100. Each epoch can be divided into groups of pairs of input-output patterns called batches. In this study, ten batch size was used. This determines the amount of pattern exposed to the network before weight is made in an epoch.

D. Prediction

After the LSTM model is formed, predictions will be made with testing data. Prediction results are denormalized first before we use them to calculate evaluation model.

$$x = x'(max - min) + min \tag{8}$$

E. Prediction

After getting the prediction model from the training process, the test data is entered into the model to get the accuracy value of the model prediction. In this study, we use RMSE to evaluate the model. The resulting value from the RMSE is the average value of the square of the amount of error in the prediction model. Root Mean Square Error (RMSE) is a model evaluation technique that is easy to apply and has been used in research related to forecasting or prediction [15].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{9}$$

In RMSE, n is the number of observations, y_t is the observed value, and $\hat{y}_t$ is the predicted value. The stages of analysis using LSTM can be described in the flowchart as in Fig. 5

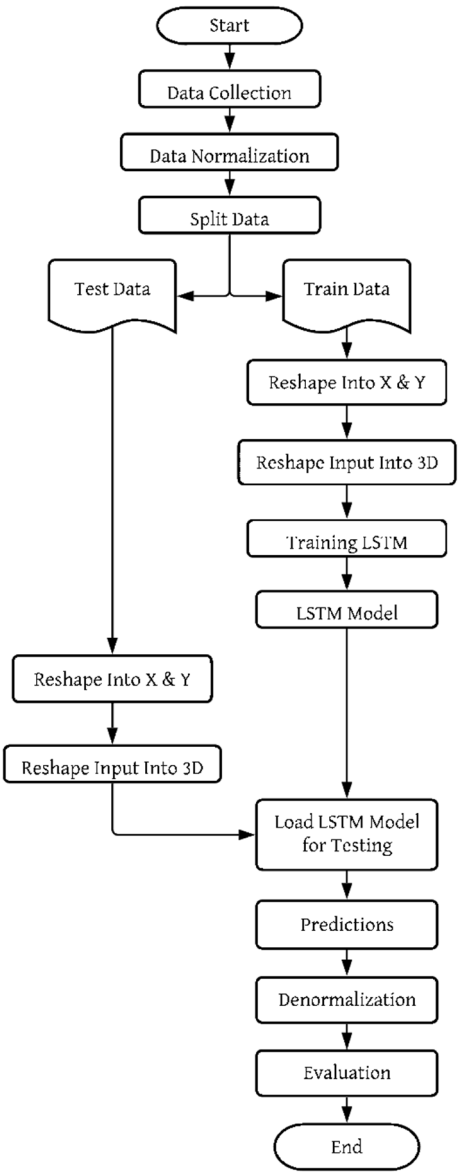


Fig. 5. Flowchart Long Short Term Memory

III. RESULT AND DISCUSSION

The initial step in forecasting time series data is done by selecting the right time forecasting step. From the monetary side, forecasting takes a long time so that preventive measures can be taken to prevent excessive transaction costs. Meanwhile, in terms of forecasting, time estimates tend to be better if using a fairly short number of time forecasts due to the limited time-series data set. In this study, inflation rate forecasts will be generated for the next 12 months. Before conducting the analysis, we identified a model to determine the characteristics of the data. Model identification is done by plotting the Autocorrelation function (ACF) and Partial autocorrelation function (PACF) to determine whether there is a long-term dependency on the data set.

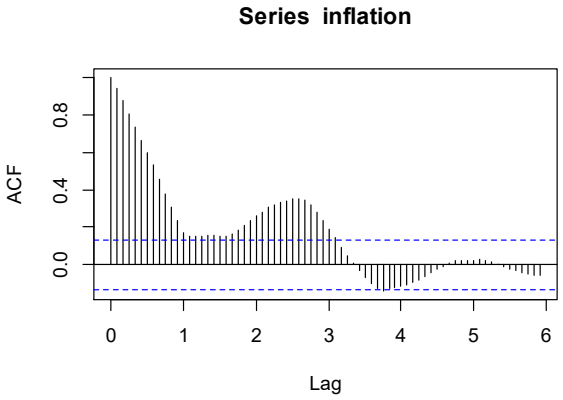


Fig. 6. ACF Plot Inflation in Indonesia 2003 - 2021

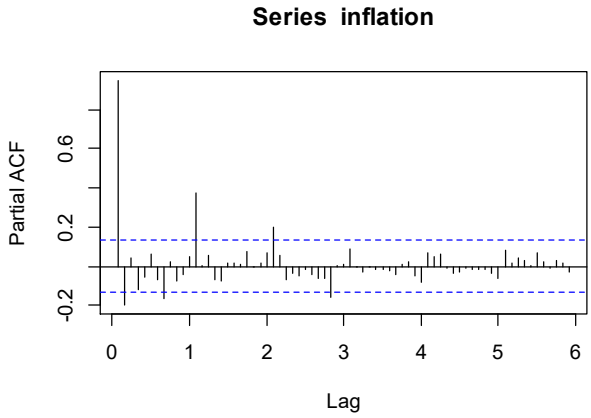


Fig. 7. Partial ACF Plot Inflation in Indonesia 2003 – 2021

Fig. 7 shows that in the PACF diagram, the first and second lags are significant because the lag crosses the line boundary in the diagram. However, the ACF graph in Fig. 6 shows a die down, this means that the inflation data set has a long-term dependence. Based on this, the LSTM model which is included in the neural network can be used as a method of forecasting inflation data.

Experiments in this study will try different hyperparameters on neuron unit in hidden layer, optimizer, and learning rate. The loss function called mean squared error and 10 batch size are used as an initialization parameter. By initializing this parameter, the training process is executed so that the loss function is close to zero which indicates the model has been optimal. Optimizers were Adam, RMSprop, and Stochastic Gradient Descent. We use early stopping with 100 epochs so that the training process makes it possible to stop before reaching 100 epochs.

Table I shows that using Adam optimizer, 0.01 learning rate, and 5 nodes in hidden layer or we can say LSTM_11 model tends to produce a smaller RMSE than another model.

From the illustration in Fig. 8, we can see that the model LSTM_11 is terminated using 55 epoch, which means this model has reached the minimum loss before maximum 100 epoch. If the loss function has reached the minimum, it shows the smallest loss from the model so that the model is properly used.

TABLE I. TABLE EVALUATION OF THE LSTM MODEL

Model	Hyperparameter			Evaluation
	Neuron Units	*optimizer*	*learning rate*	*RMSE*
LSTM_1	1	Adam	0,001	1,21
LSTM_2			0,01	0,25
LSTM_3			0,1	0,44
LSTM_4		RMSprop	0,001	1,79
LSTM_5			0,01	0,35
LSTM_6			0,1	0,71
LSTM_7		SGD	0,001	0,65
LSTM_8			0,01	2,13
LSTM_9			0,1	0,41
LSTM_10	5	Adam	0,001	0,87
LSTM_11			0,01	0,24
LSTM_12			0,1	0,41
LSTM_13		RMSprop	0,001	1,9
LSTM_14			0,01	0,33
LSTM_15			0,1	0,62
LSTM_16		SGD	0,001	0,7
LSTM_17			0,01	2,16
LSTM_18			0,1	0,3
LSTM_19	10	Adam	0,001	1,5
LSTM_20			0,01	0,27
LSTM_21			0,1	0,37
LSTM_22		RMSprop	0,001	1,14
LSTM_23			0,01	0,32
LSTM_24			0,1	0,6
LSTM_25		SGD	0,001	1,89
LSTM_26			0,01	2,57
LSTM_27			0,1	0,29

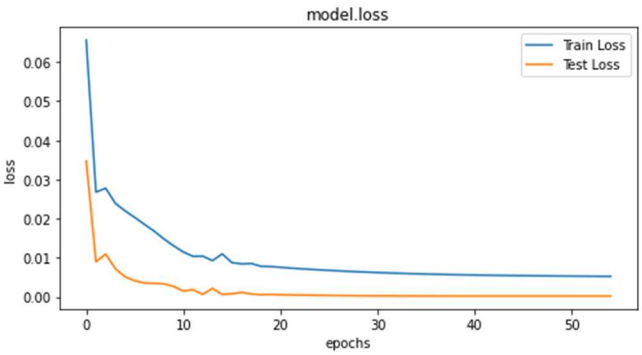


Fig. 8. The Plot of Loss Function using MSE and LSTM_11 Model

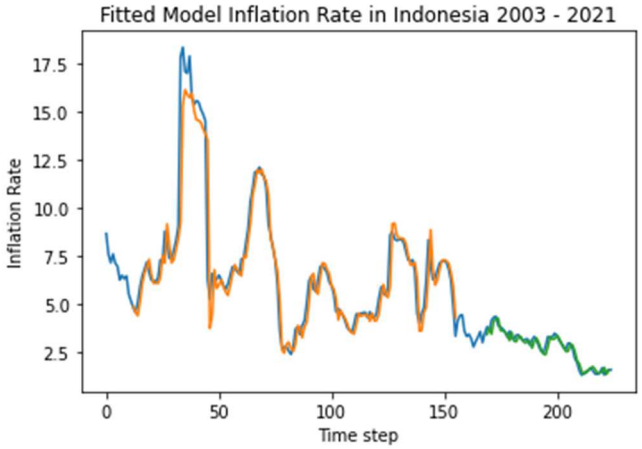


Fig. 9. The Plot of Actual and Prediction Values for Training and Testing Data using LSTM_11 Model

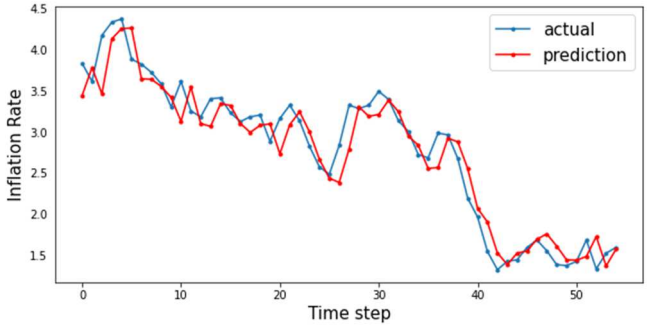


Fig. 10. The Plot of Actual and Prediction Values for Testing Data using LSTM_11 Model

As shown in Table I, the evaluation value of the test data with Adam optimizer, 5 nodes on hidden layer and the 0.01 learning rate has the smallest RMSE of 0.24. Fig. 10 shows the actual and predicted value plots for testing the LSTM_11 model. Fig. 9 shows that the LSTM model can be used to predict according to the actual value pattern in the long-term data. In addition, this model can follow patterns of extreme data values but there is still a gap in predicting extreme data at the highest inflation rate in November 2005.

We try to ensure that the predicted results with LSTM_11 are not significantly different from the actual data. We used a paired t-test with a significance level of 0.05 and obtained a p-value of 0.293. It concludes that the prediction results are not significantly different from the actual data. Therefore, the LSTM_11 model will be used to predict the future value of inflation in Indonesia over the next 12 months.

Before we build the LSTM model in this study, we have tried to apply classical methods in forecasting such as ARIMA (0,1,1)(2,1,1)[12] using inflation rate data until November 2020.

LSTM_11 in this study and ARIMA is compared using Absolute Percentage Error (APE) as we show in Table II.

$$APE = \frac{|y_t - \hat{y}_t|}{y_t} \quad (10)$$

APE is used to evaluate every period from predicting results in December 2020 until August 2021. It can be obtained that the APE of LSTM is smaller than APE of ARIMA in the predicted inflation rate in December 2020 - August 2021. It proves that prediction result using LSTM tend to reach smaller error than ARIMA.

TABLE II. ARIMA AND LSTM COMPARISON

Period	Actual Inflation	Result using ARIMA	Result using LSTM_11	Absolute Percentage Error ARIMA	Absolute Percentage Error LSTM
Dec-20	1,68	1,71	1,69	2,07	0,59
Jan-21	1,55	1,79	1,76	15,22	13,23
Feb-21	1,38	1,73	1,61	25,21	16,41
Mar-21	1,37	1,77	1,44	29,34	5,11
Apr-21	1,42	1,84	1,44	29,86	1,06
May-21	1,68	1,93	1,48	14,87	11,93
Jun-21	1,33	1,96	1,72	47,39	29,45
Jul-21	1,52	2,08	1,36	36,78	10,23
Aug-21	1,59	2,18	1,57	36,85	1,34

TABLE III. RESULTS OF THE LSTM MODEL FORECASTING FOR THE INFLATION RATE IN INDONESIA FOR THE NEXT 12 MONTHS

Period	Forecasting Results	Standard Deviation
Oct-21	1,644237032	0,181826524
Nov-21	1,700332486	0,258130082
Dec-21	1,752111615	0,329089691
Jan-22	1,798494278	0,39395064
Feb-22	1,839217626	0,452227833
Mar-22	1,874266315	0,503564624
Apr-22	1,903835104	0,547862843
May-22	1,928270592	0,585377335
Jun-22	1,948006132	0,616549027
Jul-22	1,963568616	0,642063664
Aug-22	1,975328826	0,662142619
Sep-22	1,983825842	0,677512557

Table III indicates that the inflation rate in Indonesia based on the forecast results from the LSTM model will increase until September 2022. Besides using the LSTM model, we also try to use another deep learning method for forecasting inflation rate data. We use Multi-Layer Perceptron 12–27–1 with one hidden layer containing 27 nodes, 500 epochs, Adam optimizer with 0.01 learning rate and relu activation function. We get RMSE of 0.36. We compare the testing data from LSTM_11 and MLP 12-27-1 using paired t-test to find out whether the predicted results of testing data on the LSTM model are significantly different from the predictions for testing data on the MLP model. The results of the paired t-test with a significance level of 0.05 indicate that the prediction of testing data with LSTM_11 and MLP is not significantly different with p-value of 0.1339.

According to Boediono 1998, inflation is classified into 4 types: first mild inflation where the inflation rate is <10% per year, second moderate inflation where the inflation rate is 10-30% per year, third severe inflation where the inflation rate is 30-100% per year, fourth hyperinflation where the inflation

rate is 100% per year. Inflation in Indonesia until September 2022 is predicted to still categorize as mild inflation.

IV. Conclusion

Based on the observations of the ACF and PACF graph plots, it can be seen that the data set has a long-term dependence so that the selection of the LSTM model has been appropriately used as a neural network method for modeling and forecasting inflation data. LSTM neural network also can overcome the lack of RNN in vanishing gradient and exploding gradient problems. By trying various hyperparameters on the number of nodes, learning rate, and the optimizer will select the best model based on the smallest RMSE value. In addition, in this study, the limit for stopping the training process was 100 epochs.

The results showed that among 27 models that were tested, the best model was using Adam optimizer with 0.01 learning rate, and 5 nodes in the hidden layer (Model LSTM 11) because it has the smallest RMSE value of 0.24 in the testing process. The calculation results for the model are obtained after training for 55 epochs. Forecasting data with the LSTM Model also obtained quite good results, this is evidenced by the model that can follow the data pattern even if the pattern is extreme values even though it cannot predict precisely in extreme data such as inflation that occurred in November 2005.

Since the pandemic conditions began in early 2020, from Fig.1 the inflation rate in Indonesia has not shown a significant increase and tends to decrease to below 2% and stable from June 2020 to September 2021. The forecasting results in this study using the LSTM_11 model show that until September 2022, Indonesia's inflation rate will increase but still below 2% and is classified as mild inflation. The government must still be aware of the occurrence of the third wave of COVID-19, especially at the end of the year when people tend to travel a lot also maintain price affordability and minimize constraints on supply and distribution during the Covid-19 pandemic.

Acknowledgment

The author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service Contract, Universitas Padjadjaran No: 2574/UN6.3.1/TU.00/2021.

References

[1] Y. Yolanda, "Analysis of factors affecting inflation and its impact on Human Development Index and poverty in Indonesia," European Research Studies Journal, pp. 38 - 56, 2017.

[2] W. Bank, Commodity Markets Outlook, April 2020, Washington DC: World Bank Group, 2020.

[3] W. Bank, East Asia and Pacific in the time of COVID-19, Washington DC: World Bank Group, 2020.

[4] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[5] B. Jatim, "Monthly Economic & Market Outlook (Issue March)," Bank Jatim, Surabaya, 2020.

[6] R. Pontoh, S. Zahroh and N. Sunengsih, "New normal policy on the Rupiah exchange rate using Long Short-Term Memory," in Journal of Physics: Conference Series 1863. 012063. 10.1088/1742-6596/1863/1/012063., 2021.

[7] P. Giribone, "Seasonality Modeling through LSTM Network in Inflation-Indexed Swaps.," in DATA ANALYTICS 2020 FTRM Special Track: FinTech Risk Management, 2020.

[8] Y. Hasbi, W. Budi and S. Rukun, "Feed Forward Neural Network Modeling for Rainfall Prediction," E3S Web of Conferences, no. doi: 10.1051/e3sconf/20187305017, 2018.

[9] T. Toharudin, R. Pontoh, S. Zahroh, R. Caraka, Y. Lee and R. C. Chen, "Employing long short-term memory and Facebook prophet model in air temperature forecasting," Communications in Statistics: Simulation and Computation, no. doi: 10.1080/03610918.2020.1854302., 2020.

[10] J. Tarigan, Nadia, R. Diedan dan Y. Suryana, "Plate recognition using Backpropagation Neural Network and Genetic Algorithm," Procedia Computer Science, pp. 365 - 372, 2017.

[11] I. N Yulita, S. Purwani, R. Rosadi, and R. M Awangga, "A quantization of deep belief networks for long short-term memory in sleep stage detection", 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA) (pp. 1-5), 2017.

[12] C. Olah, "Understanding LSTM Networks," 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 24 Oktober 2021].

[13] R. E. Caraka, R. C. Chen, B. D. Supatmanto, Arnita, M. Tahmid dan T. Toharudin, "Employing Moving Average Long Short Term Memory for predicting rainfall," in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2019.

[14] F. Karim, S. Majumdar, H. Darabi dan S. Chen, "LSTM Fully Convolutional Networks for time series classification," *IEEE Access,* 2017.

[15] N. R. Sari, W. F. Mahmudy, A. P. Wibawa dan E. Sonalitha, "Enabling external factors for inflation rate forecasting using fuzzy neural system," International Journal Electrical Computing, pp. 2746 - 2756, 2017.

Face Recognition Using Fisherface and Support Vector Machine Method

Syachrul Qolbi Nur Septi[ab1], Intan Nurma Yulita[ac2], Herlina Napitupulu[ab2]
[a]*Department of Mathematics, Universitas Padjadjaran*
Sumedang, Indonesia
[1]syachrul17001@mail.unpad.ac.id
[2]herlina@unpad.ac.id
[b]R*esearch Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Sumedang, Indonesia
[c]*Department of Computer Science, Universitas Padjadjaran*
Sumedang, Indonesia
[3]intan.nurma@unpad.ac.id

***Abstract*— COVID-19 is declared as a pandemic by WHO and until now COVID-19 pandemic remains a problem in 2021. Many efforts have been made to reduce the spreading virus, one way to reduce its spread is by wearing a mask but most people often ignore it. Monitoring large groups of people becomes difficult by the government or the authorities. Face recognition, a biometric technology, is based on the identification of a face features of a person. This paper describes a face recognition using Fisherface and Support Vector Machine method to classify face mask dataset. Face recognition using Fisherface method is based on Principal Component Analysis (PCA) and Fisher's Linear Discriminant (FLD) method or also known as Linear Discriminant Analysis (LDA). The algorithm used in the process for feature extraction is Fisherface algorithm while classification using Support Vector Machine method. The results show that for face recognition on face mask dataset using cross validation with 10 fold, the average percentage accuracy is 99.76%.**

Keywords*— *PCA, LDA, Fisherface, Support Vector Machine, Face Recognition

I. INTRODUCTION

COVID-19 is declared as a pandemic by WHO and until now COVID-19 pandemic remains a problem in 2021 [1]. Many efforts have been made to reduce its spread virus, one way to reduce its spread is by wearing a mask [2]. People are forced by law to wearing masks in public places and wherever they interact with others since the new normal but many of them ignore it [3], so it is more difficult to monitor large groups of people by the government or the authorities [4].

Face is the easiest part of the human body and is often used to distinguish individual identities. From face, humans can be recognized and recognized easily and quickly [5]. Therefore, the face is used as a person's tool or commonly referred to as Facial Recognition (Face Recognition) [6]. Face Recognition is a computer application that is capable of detecting, tracking, identifying, or verifying human faces from an image or video captured using a digital camera. Currently, face recognition is an attraction for computer vision observers and researchers to continue developing this system.

In general, image recognition is divided into two types, namely: feature-based and image-based. In a feature-based, the extracted feature components such as eyes, mouth, nose, ears, etc. are modeled to determine the relationship between the feature components. While the image-based uses image pixels which represented by certain methods such as Principal Component Analysis, Wavelet Transformation, etc. which is used to train and classify images [7].

Feature extraction is a process to obtain characteristics that distinguish image samples from other image samples. The feature extraction technique is the key in solving pattern recognition problems. An example of a feature extraction method is the Principal Component Analysis (PCA) which is used for face recognition and introduced by Turk and Pentland (1991). Although Principal Component Analysis (PCA) is a fairly well-known technique in image recognition, in reality, Principal Component Analysis (PCA) has problems in handling very large data so that processing time from recognition becomes long and accuracy decreases rapidly as the amount of data increases [8].

In 1997, Belheumeur introduced the Fisherface method for face recognition. This method is a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods. Principal Component Analysis (PCA) is used to solve singular problems by reducing the dimensions of the data before it is used to carry out the Linear Discriminant Analysis (LDA) process [6]. Belhumeur et al's research using the Fisherface method showed better results than the PCA or LDA method itself because the greater the ratio, the more feature vectors produced the less sensitive to changes in expression and changes in light so that it can produce a good classification.

One of the statistical methods that can be applied to perform classification is the Support Vector Machine (SVM). SVM is a technique for finding a hyperplane that can separate two data sets of two different classes. SVM has advantages including in determining the distance using a support vector so that the computing process becomes fast [9].

This paper will describe the research on face recognition using face mask dataset. Finally, we introduce the general evaluation criteria of face recognition.

II. MATERIALS AND METHOD

A. System Design

Face recognition using the Fisherface and SVM method is designed to recognize face images by classifying the feature extraction results with SVM method. This process is expected to determine whether the image to be tested is classified correctly or not. In this research, 5000 face images with and without a mask are used in *.png format which is generated using style GAN-2 [10].

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

B. Process Design

1) Data Preprocessing

The face image must be preprocessed first. This stage converts RBG image to grayscale. Conversion of face image from RGB to grayscale with size 128×128 pixels. Furthermore, the image is divided into training image (training dataset) and test image (testing dataset).

2) Feature Extraction

Features extraction produce feature image of the people face with and without wearing a mask. Fisherface method is chosen which is a merger between PCA and LDA methods.

III. Results

In this section, we discussed about the algorithm and results of face recognition using Fisherface and SVM method. In general, the stage of the face recognition process in this study can be seen in Fig. 1.

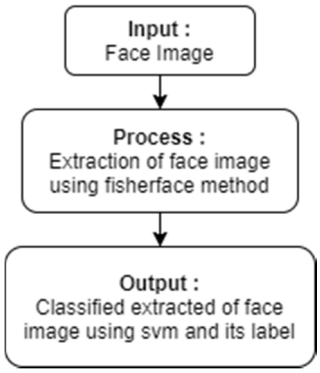


Fig. 1. Stage of process

A. Image Data

Here is a sample of images of people with and without wearing a mask which is generated using style GAN-2. The images can be seen in Fig. 2.

Fig. 2. An example of image people with and without a mask

Here is a sample of images after data preprocessing. The images can be seen in Fig. 3.

Fig. 3. An example of image people with and without mask after data preprocessing

B. Feature Extraction Method (Fisherface)

1) PCA Algorithm

- Take training data images and their labels, then each image is transformed into a row vector through data preprocessing, then each row vector is made a matrix so that a matrix will be obtained, each row represents a different image ($\boldsymbol{X}$) with the associated image label ($\boldsymbol{Y}$).
- Calculate the average matrix ($\boldsymbol{\Psi}$), which is the average value of all training data images, as in (1).

$$\psi_k = \frac{1}{M}\sum_{m=1}^{M} x_{mk}; \quad m = 1,2,3,\dots,M \tag{1}$$

Here is an example of average image people with and without mask. The images can be seen in Fig. 4.

Fig. 4. An example of average image people with and without mask

- Calculate the difference matrix ($\boldsymbol{\Phi}$), which is the difference between all training data images and their average value, as in (2).

$$\phi_{mk} = x_{mk} - \psi_k \tag{2}$$

- Calculating the total scatter matrix ($\boldsymbol{S_T}$), by multiplying the difference matrix by the transpose, as in (3).

$$s_{t_{kk}} = \sum_{m=1}^{M} \phi_{km}\phi_{mk} \tag{3}$$

- Calculate the eigenvalue (λ_{PCA}) and eigenvector ($\boldsymbol{v_{PCA}}$) of the total scatter matrix, as in (4).

$$\boldsymbol{S_T v_{PCA}} = \lambda_{PCA}\boldsymbol{v_{PCA}} \tag{4}$$

- Sort the eigenvectors based on the eigenvalues of each eigenvector.
- Take the required $N_{\boldsymbol{v_{PCA}}} - C$ eigenvectors where $N_{\boldsymbol{v_{PCA}}}$ is the number of eigenvectors and C is the number of classes.
- Take the selected set of eigenvectors then each eigenvector is transformed into a row vector, then each row vector is made a matrix so that a matrix that each row represents each eigenvector ($\boldsymbol{W_{PCA}}$) will be obtained.
- Calculate the eigenface weight matrix ($\boldsymbol{\Omega_{PCA}}$), by multiplying the difference matrix by the $\boldsymbol{W_{PCA}}$ transpose matrix, as in (5).

$$\omega_{PCA_{ml}} = \sum_{k=1}^{K} \phi_{mk} w_{PCA_{kl}}; \quad l = 1,2,3,\dots,L \tag{5}$$

2) LDA Algorithm

- Take the eigenface weight matrix ($\boldsymbol{\Omega_{PCA}}$) and the label matrix ($\boldsymbol{Y}$) to be used as input.

- Calculates the overall mean matrix ($\boldsymbol{\Psi}^{\boldsymbol{\Omega}_{PCA}}$) and the average on the same label ($\boldsymbol{\Psi}_{\boldsymbol{c}}^{\boldsymbol{\Omega}_{PCA}}$) from the eigenface weight matrix $\boldsymbol{\Omega}_{\boldsymbol{PCA}}$, as in (6) and (7).

$$\psi_l^{\Omega_{PCA}} = \frac{1}{M}\sum_{m=1}^{M} \omega_{PCA_{ml}}; \quad l = 1,2,3,\dots,L \tag{6}$$

$$\psi_{cl}^{\Omega_{PCA}} = \frac{1}{M^{(c)}}\sum_{m^{(c)}=1}^{M^{(c)}} \omega_{PCA_{m^{(c)}l}}; \quad m^{(c)} = 1,2,3,\dots,M^{(c)} \tag{7}$$

- Calculating the distribution between classes scatter matrix ($\boldsymbol{S_B}$) from the eigenface weight matrix $\boldsymbol{\Omega_{PCA}}$, as in (8).

$$\boldsymbol{S_b} = \sum_{c=1}^{M^{(c)}} M^{(c)} \left(\boldsymbol{\Psi_c^{\Omega_{PCA}}} - \boldsymbol{\Psi^{\Omega_{PCA}}}\right)\left(\boldsymbol{\Psi_c^{\Omega_{PCA}}} - \boldsymbol{\Psi^{\Omega_{PCA}}}\right)^T \tag{8}$$

- Calculating the distribution within classes scatter matrix ($\boldsymbol{S_W}$) from the eigenface weight matrix $\boldsymbol{\Omega_{PCA}}$, as in (9).

$$\boldsymbol{S_w} = \sum_{c=1}^{C}\sum_{m=1}^{M^{(c)}} \left(\boldsymbol{\Omega}_{PCA_{m^{(c)}}} - \boldsymbol{\Psi_c^{\Omega_{PCA}}}\right)\left(\boldsymbol{\Omega}_{PCA_{m^{(c)}}} - \boldsymbol{\Psi_c^{\Omega_{PCA}}}\right)^T \tag{9}$$

- Calculate the eigenvalue (λ_{LDA}) and eigenvector ($\boldsymbol{v_{LDA}}$) of the $\boldsymbol{S_B}$ and $\boldsymbol{S_W}$ matrix, as in (10).

$$(\boldsymbol{S_w})^{-1}\boldsymbol{S_b}\boldsymbol{v_{FLD}} = \lambda_{FLD}\boldsymbol{v_{FLD}} \tag{10}$$

- Sort the eigenvectors based on the eigenvalues of each eigenvector.
- Take the required $C-1$ eigenvectors where C is the number of classes.
- Take the selected set of eigenvectors then each eigenvector is transformed into a row vector, then each row vector is made a matrix so that a matrix that each row represents each eigenvector ($\boldsymbol{W_{LDA}}$) will be obtained.
- Calculating the Fisherface matrix ($\boldsymbol{W_{OPT}}$), by multiplying the matrix $\boldsymbol{W_{LDA}}$ by $\boldsymbol{W_{PCA}}$. matrix, as in (11).

$$w_{OPT_{l^*k}} = \sum_{l=1}^{L} w_{FLD_{l^*l}} w_{PCA_{lk}}; \quad k = 1,2,3,\dots,K \tag{11}$$

Here is an example of Fisherface of image people with and without a mask. The images can be seen in Fig. 5.

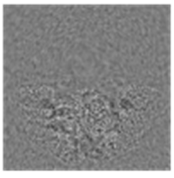


Fig. 5. An example Fisherface of image people with and without a mask

- Calculate the Fisherface weight matrix ($\boldsymbol{\Omega_{LDA}}$), by multiplying the difference matrix by the $\boldsymbol{W_{OPT}}$ transpose matrix, as in (12).

$$\omega_{FLD_{ml^*}} = \sum_{k=1}^{K} \phi_{mk} w_{OPT_{kl^*}}; \quad l^* = 1,2,3,\dots,L^* \tag{12}$$

C. Classification Algorithm

- Test data is taken from the results of data preprocessing and then stored into an $\boldsymbol{X^{TEST}}$ matrix and image labels are stored in a $\boldsymbol{Y^{TEST}}$ matrix where y_u is a label for $\boldsymbol{X_t^{TEST}}$ images.
- Calculate the difference matrix ($\boldsymbol{\Phi^{TEST}}$), which is the difference between all test data images and their average value, as in (2).
- Calculate the Fisherface weight matrix ($\boldsymbol{\Omega_{LDA}^{TEST}}$), by multiplying the difference matrix by the $\boldsymbol{W_{OPT}}$ transpose matrix, as in (12).
- *Classify each test data Fisherface weight ($\boldsymbol{\Omega_{LDA}^{TEST}}$) using the SVM model based on the training data Fisherface weight ($\boldsymbol{\Omega_{LDA}}$).*

D. Analysis SVM Model

Data training and testing are divided into several main scenarios. Each scenario has a different purpose in measuring the performance of different scenario. Testing the classification model using the 10-fold cross validation technique, means that 5000 data is divided into 10 folds (parts). A combination of 9 different folds is combined and used as training data, the remaining 1-fold is used as testing data. In this section, the test results of each scenario are presented to see the performance support vector machine method in face recognition of people with and without wearing a mask. The performance of each scenario is measured based on calculation results accuracy, precision, and recall.

1) Kernel Function Performance Comparison

This stage aims to obtain the most optimal kernel function to be applied in the final model prediction in this study. Kernel functions to be used in this test scenario are Linear Kernel, Polynomial Kernel, Radial Base Function (RBF) Kernel and Sigmoid Kernel. Comparison of test results is seen based on accuracy, precision, and recall on every kernel function.

TABLE I. KERNEL FUNCTION COMPARISON ACCURACY

Fold	Kernel Function Accuracy (%)			
	Linear	Polynomial	RBF	Sigmoid
1	99.80%	99.80%	99.80%	99.80%
2	99.60%	99.80%	99.60%	99.60%
3	99.80%	99.80%	99.80%	99.80%
4	99.80%	100%	99.80%	99.80%
5	100%	100%	100%	100%
6	99.80%	99.80%	99.80%	99.80%
7	99.80%	99.80%	99.80%	99.80%
8	99.40%	99.40%	99.40%	99.40%
9	99.60%	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%	99.60%

Table I shows that the comparison of performance prediction accuracy each scenario using SVM with various kernel functions has a sufficient accuracy value good. From the calculation of 10-fold cross-validation for this accuracy, the highest average percentage accuracy is obtained from the SVM prediction with the Polynomial Kernel function by 99.76%, followed by the use of Linear Kernel, RBF Kernel, and Sigmoid Kernel with 99.72% accuracy percentage.

Table II shows that the comparison of performance prediction precision for each scenario using SVM with various kernel functions has a sufficient precision value good. From the calculation of 10-fold cross-validation for this precision, the highest average percentage precision is obtained from the SVM prediction with the Polynomial Kernel function by 99.79%, followed by the use of Linear Kernel, RBF Kernel, and Sigmoid Kernel with 99.76% precision percentage.

TABLE II. KERNEL FUNCTION COMPARISON PRECISION

Fold	Kernel Function Precision (%)			
	Linear	Polynomial	RBF	Sigmoid
1	100%	100%	100%	100%
2	99.20%	99.60%	99.20%	99.20%
3	99.60%	99.60%	99.60%	99.60%
4	100%	100%	100%	100%
5	100%	100%	100%	100%
6	100%	100%	100%	100%
7	100%	100%	100%	100%
8	99.60%	99.60%	99.60%	99.60%
9	99.60%	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%	99.60%

Table III shows that the comparison of performance prediction recalls each scenario using SVM with various kernel functions has a sufficient recall value good. From the calculation of 10-fold cross-validation for this recall, the highest average percentage recall is obtained from the SVM prediction with the Polynomial Kernel function by 99.72%, followed by the use of Linear Kernel, RBF Kernel, and Sigmoid Kernel with 99.68% recall percentage.

2) C Parameter Performance Comparison

This stage aims to obtain the most optimal regularization parameter (C) to be applied in the final model prediction in this study. Regularization parameters (C) to be used in this test scenario are $C = 1, C = 5, C = 10$. Comparison of test results is seen based on accuracy, precision, and recall on every regularization parameter (C).

TABLE III. KERNEL FUNCTION COMPARISON RECALL

Fold	Kernel Function Recall (%)			
	Linear	Polynomial	RBF	Sigmoid
1	99.60%	99.60%	99.60%	99.60%
2	100%	100%	100%	100%
3	100%	100%	100%	100%
4	99.60%	100%	99.60%	99.60%
5	100%	100%	100%	100%
6	99.60%	99.60%	99.60%	99.60%
7	99.60%	99.60%	99.60%	99.60%
8	99.20%	99.20%	99.20%	99.20%
9	99.60%	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%	99.60%

TABLE IV. REGULARIZATION PARAMETER (C) COMPARISON ACCURACY

Fold	Regularization Parameter (C) Accuracy (%)		
	$C = 1$	$C = 5$	$C = 10$
1	99.80%	99.80%	99.80%
2	99.80%	99.80%	99.80%
3	99.80%	99.80%	99.80%
4	100%	100%	100%
5	100%	100%	100%
6	99.80%	99.80%	99.80%
7	99.80%	99.80%	99.80%
8	99.40%	99.40%	99.40%
9	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%

TABLE V. REGULARIZATION PARAMETER (C) COMPARISON PRECISION

Fold	Regularization Parameter (C) Precision (%)		
	$C = 1$	$C = 5$	$C = 10$
1	100%	100%	100%
2	99.60%	99.60%	99.60%
3	99.60%	99.60%	99.60%
4	100%	100%	100%
5	100%	100%	100%
6	100%	100%	100%
7	100%	100%	100%
8	99.60%	99.60%	99.60%
9	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%

Table IV shows that the comparison of performance prediction accuracy for each scenario using SVM with various regularization parameter (C) has a sufficient accuracy value good. From the calculation of 10-fold cross validation for this accuracy, the average percentage accuracy is obtained from the SVM prediction with the regularization parameter $C = 1, C = 5, C = 10$ by 99.76% accuracy percentage.

Table V shows that the comparison of performance prediction precision for each scenario using SVM with various regularization parameters (C) has a sufficient precision value good. From the calculation of 10-fold cross-validation for this precision, the highest average percentage precision is obtained from the SVM prediction with the regularization parameter $C = 1, C = 5, C = 10$ by 99.80% accuracy percentage.

TABLE VI. REGULARIZATION PARAMETER (C) COMPARISON RECALL

Fold	Regularization Parameter (C) Recall (%)		
	$C = 1$	$C = 5$	$C = 10$
1	99.60%	99.60%	99.60%
2	100%	100%	100%
3	100%	100%	100%
4	100%	100%	100%
5	100%	100%	100%
6	99.60%	99.60%	99.60%
7	99.60%	99.60%	99.60%
8	99.20%	99.20%	99.20%
9	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%

Table VI shows that the comparison of performance prediction recall each scenario using SVM with various regularization parameter (C) has a sufficient recall value good. From the calculation of 10-fold cross validation for this recall, the highest average percentage recall is obtained from the SVM prediction with the regularization parameter $C = 1, C = 5, C = 10$ by 99.72% accuracy percentage.

E. System Result

To determine whether the system is running well and properly it is necessary to test the model with the following process.

1) Training Process

The first process is the training process. This process aims to generate the weight of each image of training images ($\boldsymbol{W_{OPT}}$).

2) Image Recognition Process

This process is to recognize the test image with SVM method. From the analysis SVM model, we choose Polynomial Kernel, and $C = 1$. The goal is how accurate the system is to recognize the test image.

3) Image Recognition Results

The following is the results of face which can be seen in Table VII.

TABLE VII. FACE RECOGNITION FINAL RESULTS

Fold	Accuracy (%)	Precision (%)	Recall (%)
1	99.80%	100%	99.60%
2	99.80%	99.60%	100%
3	99.80%	99.60%	100%
4	100%	100%	100%
5	100%	100%	100%
6	99.80%	100%	99.60%
7	99.80%	100%	99.60%
8	99.40%	99.60%	99.20%
9	99.60%	99.60%	99.60%
10	99.60%	99.60%	99.60%

IV. CONCLUSION

Face recognition using Fisherface and SVM methods using face mask dataset produces a high accuracy. This is indicated by the value of the average prediction accuracy of the test data obtained by 99.76%, average prediction precision by 99.79%, and average prediction recall by 99.72%. Face recognition using Fisherface and SVM methods is not only capable of performing an introduction to the test face images with different color components of the training image and a sketch of the original image. This method also work to noise induced images and the blurring effect on the images.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] Y. H. Lin, C. H. Liu, and Y. C. Chiu, "Google searches for the keywords of 'wash hands' predict the speed of national spread of COVID-19 outbreak among 21 countries," Brain. Behav. Immun., vol. 87, pp. 30–32, 2020.

[2] O. O. Fadare and E. D. Okoffo, "Covid-19 face masks: A potential source of microplastic fibers in the environment," Sci. Total Environ., vol. 737, p. 140279, 2020.

[3] S. Setiati and M. K. Azwar, "COVID-19 and Indonesia," Acta Med. Indones., vol. 52, no. 1, pp. 84–89, 2020.

[4] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," Meas. J. Int. Meas. Confed., vol. 167, p. 108288, 2021.

[5] F. Mahmud, M. T. Khatun, S. T. Zuhori, S. Afroge, M. Aktar, and B. Pal, "Face recognition using Principle Component Analysis and Linear Discriminant Analysis," in 2nd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2015, 2015, pp. 1–4.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE Transactions on pattern analysis and machine intelligence, vol. 19, no. 7, pp. 711–720, 1996.

[7] M. Anggo and La Arapu, “face recognition using Fisherface method,” in Journal of Physics: Conference Series, vol. 1028, no. 1, p. 12119, 2018

[8] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–587.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[10] P. Kottarathil, Face Mask Lite Dataset. Kaggle Data. Kaggle, 2020.

Stroke Risk Prediction Model Using Machine Learning

Nugroho Sinung Adi[ab1], Richas Farhany[ab2], Rafidah Ghina[ab3], Herlina Napitupulu[bc4]
[a]*Department of Economic and Business, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Mathematics, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]sinung19002@mail.unpad.ac.id, [2]farhany19001@mail.unpad.ac.id, [3]ghina19008@mail.unpad.ac.id, [4]herlina@unpad.ac.id

Abstract*— Stroke is a disease that will occur when the blood flow to the brain is disrupted or diminished, causing the cell to die and it will lead to disability or even death. According to the WHO, stroke is the second deadliest disease in the world. Currently, the development of the Industrial Revolution 4.0 is collaborating with the health sector by using Machine Learning to make an accurate prediction of stroke that is highly valuable for early intervention and treatment. This research uses machine learning to predict a patient who is at high risk to get a stroke. The research will use three machine learning algorithm models, including Naive Bayes, Decision Tree, and Random Forest. The prediction use patient health history as the attribute in each mode. After testing for each of these methods. The first method that has the best accuracy is the Random Forest method with 94,781% accuracy, followed by the Decision Tree method with 91,906% accuracy and the least accurate is the Naïve Bayes Method with 89,976% accuracy. Based on these results, it can be concluded that the Random Forest method has the highest accuracy than other methods.

Keywords—Classification, Machine Learning, Decision Tree, Naive Bayes, Random Forest, Stroke

I. INTRODUCTION

Stroke or known as the term Brain Circulatory Disorder is a disease that will occur when the blood flow to the brain is disrupted or diminished. And when that part of the brain is not getting the proper blood flow, the cell will not get the oxygen and the nutrients, causing it to die[1]. Stroke is currently the second-highest death worldwide. WHO says that every year, about fifteen million people around the world suffer from a stroke, and 87% of people who have died from stroke often occur in low- and middle-income countries [2]. According to the diagnosis of health workers, the number of stroke sufferers in Indonesia in 2013 was estimated at 1,236,825 people from all recorded stroke patients, and as many as 80% were ischemic stroke types. Stroke is a serious medical emergency that needs to be solved because the death toll from having a stroke is rising every year.

From the Framingham study, several lists can be causing stroke, including systolic blood pressure, diabetes Mellitus, cigarette smoking, the use of anti-hypertensive therapy, prior cardiovascular disease, alcohol, atrial fibrillation, and age. Patients in the early stage of the stroke will have hearing loss, visual decline or Amnesia, Abnormal behavior or Dementia, and other minor symptoms [3]. Currently, the prognosis of stroke is mostly recovered with disabilities such as motor neuron disorders, memory disorders, and speech disorders. This perceived disability causes stroke sufferers to have a low quality of life.

The problem that often occurs in stroke, is an attack that comes suddenly. These sudden attacks are the reason why the prediction is very important to be able to know the simpler way because all of the calculations are performed using a computer system [4]. These characteristics make using machine learning techniques suitable for the medical field because they can predict a patient from having a stroke more efficient.

Machine Learning is proven to improve the provision in optimization and classification for creating intelligent systems [5]. In stroke prediction, Machine learning can work automatically with patient data and be used directly to predict stroke risk in patients. If the stroke can be predicted then the patient can have an early and less harmful medication. Detecting early stroke can help with the cost of the treatment, it is less expensive.

Classification is a technique to form a data model that has not been classified, therefore the model can be used to classify new data [6]. In data mining, many algorithms can be used in the classification process, namely Decision tree, Naive Bayes K-Nearest neighbor, Support Vector Machine, Random Forest, and many other algorithms.

In conducting this research, there are 3 data mining algorithms used, namely, *Naive Bayes, Decision Tree,* and *Random Forest.* Out of these three data mining algorithms, the result will be compared to which one is the most accurate algorithm for predicting stroke.

II. RESEARCH METHODS

A. Naïve Bayes Algorithm

Naive Bayes Algorithm is a classification based on the statistical probability that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset. This algorithm uses the Bayes theorem and assumes all attributes are independent or not interdependent given by the value of the class variable [7].

$$P(C\ |x) = \frac{P(x|C)P(C)}{P(x)} \quad (1)$$

To explain the Naive Bayes method, please note that the classification process requires several instructions to determine what class is suitable for the sample being analyzed. Therefore, the above Naive Bayes method is adjusted as follows:

$$P(C|F1 \ldots\ldots Fn) = \frac{P(C)P(F1 \ldots. Fn\ |C)}{P(F1 \ldots. Fn)} \quad (2)$$

The advantage of Naive Bayes is that this method requires a bit of training data to determine the parameter range to be used in the classification because an independent variable takes only the variation of a variable in a class

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

needed to define the classifier without the entire covariance matrix [11].

The flow of the *Naïve Bayes Algorithm* can be seen in Fig. 1.

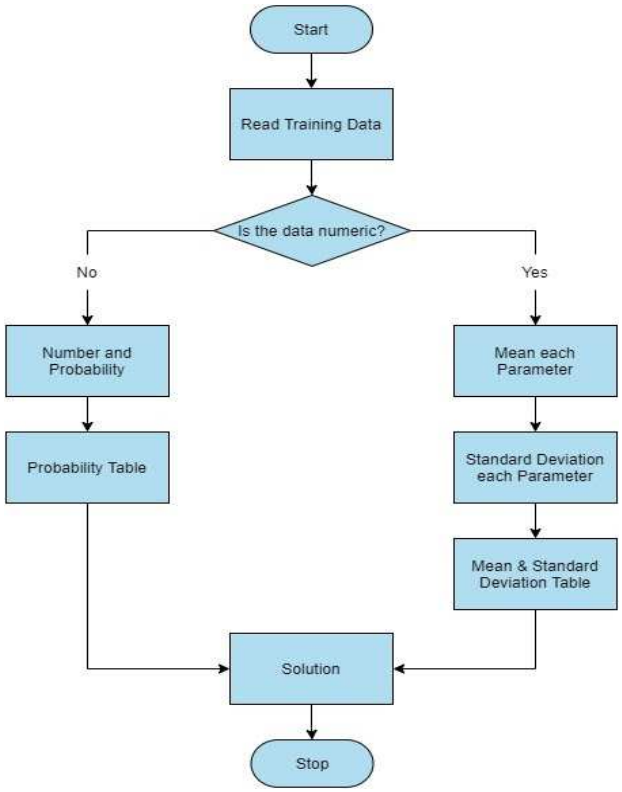


Fig. 1. Naïve Bayes Method Flow

B. Decision Tree Algorithm

A Decision Tree is a method in classification techniques that exist in data mining. A Decision Tree is a tree where each node represents a test of an attribute and the leaf node provides classification [8]. This method can also be useful for data exploration, finding a hidden relationship between a group of candidate input variables and a target variable.

In the method decision tree, data is usually expressed in the form of a table with records and attributes. An attribute states that a parameter is called the criteria in the formation of the tree. For example, if you decide to play tennis, the criteria that will be considered are temperature, weather, and wind. If one of the attributes is an attribute that represents solution data per power item it is referred to as a result attribute. There are several algorithms used in the formation of a decision tree, one of which is the C4.5 algorithm.

The C4.5 algorithm is a development of the ID3 Algorithm. Both of these algorithms were created by a researcher in the field of artificial intelligence named J. Rose Quinan in the late 1970s.

In general, the steps of the C4.5 algorithm in making a decision tree are:

1. Choose an attribute as the root.
2. A harrow for each value
3. Divide cases into branches
4. Repeat the process on each branch until all cases in the branch have the same class.

To select an attribute as the root, the highest gain value is based on the existing attributes. Gain can be calculated using the formula as shown in equation 1 below.

$$Entropy(s) = \sum_{i=1}^{n} P_i * log_2 Pi \quad (3)$$

Meanwhile, to calculate the entropy value can be used with the following equation

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|s|} * Entropy(S_i) \quad (4)$$

C. Random Forest Algorithm

Random forest is an extension of a decision tree and was first introduced by Leo Breiman. A random forest consists of a combination of decision trees in which each tree rests on a random vector value sampled independently and with the same distribution for all trees in the forest. Classification in random forest consists of a collection of classifications $\{h(x, \Theta_k), k = 1, 2, \ldots\}$ where $\{\Theta_k\}$ is an independently distributed random vector and each tree assigns a voting unit to the most popular class in input x [9]. The following explanations are given for the sake of identify the accuracy of the Random Forest.

1. Random Forests Converge

Centralizing random forests by determining the margin function can determine more accurate results. If the classification ensemble is $h_1(x), h_2(x), \ldots, h_k(x)$ with a random training set from the random vector distribution Y, X. The margin can be determined by the following equation:

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max av_k I(h_k(X) = j). \quad (5)$$

The indicator function is $I(\cdot)$. The margin function is used to measure how far the average number of votes in X, Y for a class exceeds the average vote for other classes. The larger the margin obtained, the more accurate the classification results.

2. Strength and Correlation

The upper bound on the random forest can be derived for generalization error by

$$PE^* \leq \bar{\rho}(1 - s^2)/s^2 \quad (6)$$

Two things that influence the generalization error are the strength of the individual classifiers in the *forest* and the correlation between them on the margin function. The ratio c/s2 is the correlation divided by the power to the power of two. The smaller this ratio, the better the accuracy of the random forest.

D. Classification Metrics

When performing classification predictions, there are four types of outcomes that could occur.

- True positives (TP): Cases where the positive class is predicted as positive by the model
- False positives (FP): Cases where the class was negative but the model predicted as positive
- True negatives (TN): Cases where the negative class is predicted as negative by the model.

- False negatives (FN): Cases where the class was positive but the model predicted as negative

These four outcomes are plotted on a confusion matrix, which can be seen in Fig. 2 below.

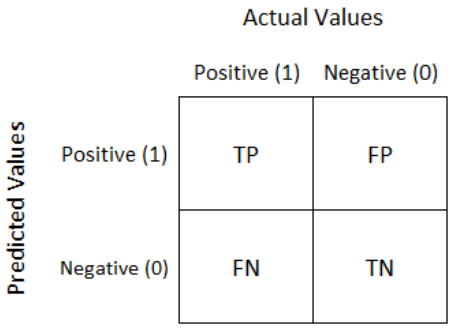


Fig. 2. Confusion Matrix

The four metrics used to evaluate a classification model are accuracy, precision, recall, and F-Measure [12].

1) Accuracy is defined as the percentage of correct predictions for the testing data. It is calculated by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \quad (7)$$

2) Precision is a percentage of positive instances out of the total predicted positive instances.

$$Precision = \frac{TP}{(TP+FP)} \quad (8)$$

3) Recall is a percentage of positive instances out of total actual positive instances.

$$Recall = \frac{TP}{(TP+FN)} \quad (9)$$

4) F-Measure is a measure of test's accuracy that considers both the precision and the recall of the test to compute the score.

$$F - Measure = \frac{2(Recall \times Precision)}{(Recall \times Precision)} \quad (10)$$

By using these four indicators as benchmarks to define the evaluation parameters, it is easy to determine whether a model can be classified as a good or a bad model.

III. METHODOLOGY

A. Research Procedure

The procedure performed in testing the algorithms consists of six steps, as shown in Fig. 3. Including (1) collecting data, (2) data labeling, (3) data preparation, (4) data partitioning, (5) data classification, and (6) evaluation.

In this research, the tool used to carry out the above procedure (see Fig. 3) and test the algorithms is the Knime Analytics Platform. The workflow performed in Knime can be seen in Fig. 4 below.

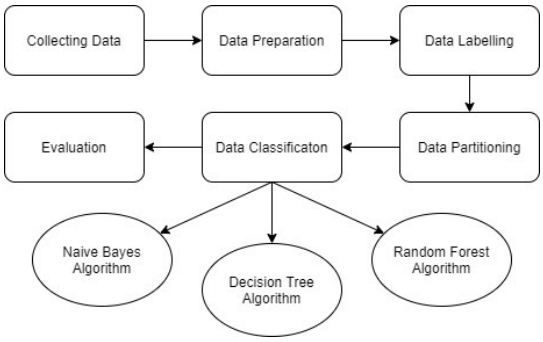


Fig. 3. Research Procedure Flowchart

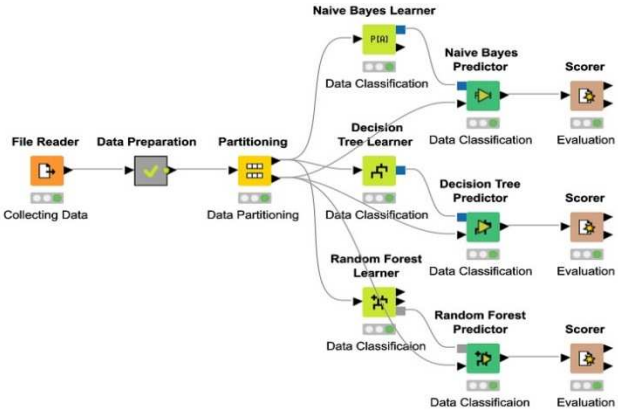


Fig. 4. KNIME Workflow

B. Collecting Data

The stroke dataset used to classify stroke predictions was taken from Kaggle in September 2021 [10]. The dataset is shown in Table I, it consists of 12 attributes, including:

TABLE I. LIST OF ATTRIBUTES IN DATASET

No	Attributes	Description
1	Id	unique identifier
2	Gender	"Male","Female", or "Other"
3	Age	age of the patients
4	Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5	heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6	ever_merried	"No" or "Yes"
7	work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8	residence_type	"Rural" or "Urban"
9	average_glucose_level	average glucose level in blood
10	Bmi	body mass index
11	Smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
12	Stroke	1 if the patient had a stroke or 0 if not

C. Data Labelling

The dataset is labeled in the 'stroke' column. There are 2 labels in the dataset, label ‘0’ indicates that the patient is not diagnosed with a stroke, and label ‘1’ indicates that the patient is diagnosed with stroke. According to Fig. 5, there are 4861 patients are not diagnosed with stroke and 249 patients are diagnosed with stroke.

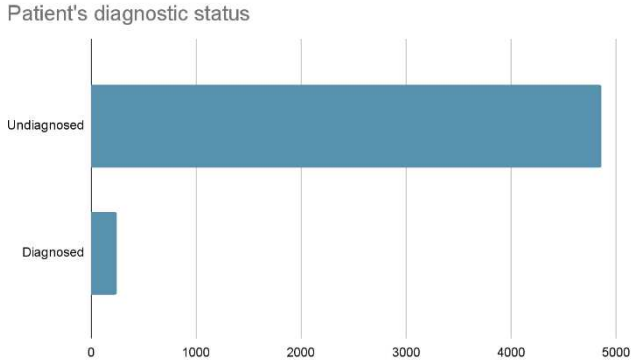


Fig. 5. Patient’s Diagnostic Status

D. Data Preparation

In this step, the dataset will be prepared before testing with the algorithms in the next step, which can be seen in Fig. 6.

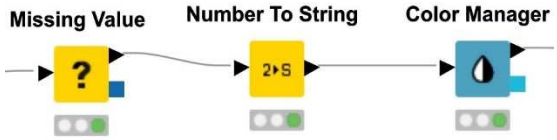


Fig. 6. Data Preparation KNIME Workflow

The preparations include:

- overcome missing values in the dataset. There are 201 missing values in the 'bmi' attribute. So, we fill the missing value data with the average of 'bmi' value,
- change the data type of 'hypertension', 'heart_disease', and 'stroke' attributes to string because the data from the attribute shows 'yes' and 'no' statements, but it's represented by using numbers '0' and '1',
- assign a color to the final result that will differentiate between diagnosed and undiagnosed stroke patients

E. Data Partitioning

Partitioning is done to divide the data randomly into two datasets, the first is 70% to be used as a training data and 30% to be used as a testing data. Training data is used to create and train a model, then testing data is used to test the model which has been made.

F. Data Classification

After the dataset has been prepared and partitioned, the next step is to test the algorithms. In the Naïve Bayes algorithm, a Naïve Bayes learner is used to create a model from training data, and use a Naïve Bayes predictor to test the model with testing data. Then in a decision tree algorithm, a decision tree learner is used to create a model from training data, then use a decision tree predictor to test the model with testing data. And in the random forest algorithm, a random forest learner is used to create a model from training data, then use a random forest predictor to test the model with testing data.

IV. Result and Analysis

The following are the results of the evaluation of the output obtained from using the three methods of machine learning classification algorithms.

A. Test Results with Naive Bayes

The following is the result of evaluating the output using the Naive Bayes classification algorithm method that is shown in Table II.

TABLE II. Naive Bayes Confusion Matrix

Stroke/ Prediction (Stroke)	1	0
1	27	50
0	119	1337
Correct Classified	1364	
Wrong Classified	169	
Accuracy	88.976%	
Error	11.024%	
Cohen’s Kappa (K)	0.189	

B. Test Results with Decision Tree

The following is the result of evaluating the output using the Decision Tree classification algorithm method that is shown in Table III.

TABLE III. Decision Tree Confusion Matrix

Stroke/ Prediction (Stroke)	1	0
1	13	64
0	60	1395
Correct Classified	1408	
Wrong Classified	124	
Accuracy	91.906%	
Error	8.094%	
Cohen’s Kappa (K)	0.131	

C. Test Results with Random Forest

The following is the result of evaluating the output using the Random Forest classification algorithm method that is shown in Table IV.

TABLE IV. Random Forest Confusion Matrix

Stroke/ Prediction (Stroke)	1	0
1	0	77
0	3	1453
Correct Classified	1453	
Wrong Classified	80	
Accuracy	94.781%	
Error	5.219%	
Cohen’s Kappa (K)	- 0.004	

D. Comparison Results of the Three Methods

After a trial of three classification algorithm methods, namely Naive Bayes, decision tree, and random forest are done, the comparisons of the evaluation output results are shown in Table V.

TABLE V. TEST RESULTS OF THREE TYPES OF METHODS

No	Parameter	Method Type		
		Naive Bayes	Decision Tree	Random Forest
1	Accuracy	89,976%	91,906%	94,781%
2	Precision	0.185	0.178	0
3	Recall	0.351	0.169	0
4	F-Measure	0.242	0.173	NaN
5	Kopen Kappa	0.189	0,131	-0.004

From the test results above, the three models have a fairly good performance. By doing a 70/30 data split. A random forest has the highest accuracy of 94.26%, then a decision tree with an accuracy value of 92.917% and a Naive Bayes accuracy value of 89.4%.

V. CONCLUSION & RECOMMENDATION

This research is conducted to predict stroke risk, As the second-highest death worldwide, detecting stroke in an early stage will help the patient to have better and less harmful medication, reduce the cost of medication, help aid people's health, predict accurate outcomes, help upgrade the healthcare value, and, the most important thing, it can help to save people lives.

The highest accuracy value was obtained by Random Forest with a result of 94.26%, in the test to predict stroke disease using 12 variables with a total data of 5,109 data. Random Forest has the advantage in classifying data because it works for data that has incomplete attributes, and is good for handling large sample data.

For further research, an application or smart system suggested to be made in predicting the diagnosis of stroke, by adding other algorithms in machine learning to create better and more accurate models. It is recommended also to add attributes to the dataset such as recent strenuous activity, and occupations to strengthen the prediction.

ACKNOWLEDGMENT

The Author thanks the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

[2] M. S. Phipps and C. A. Cronin, "Management of acute ischemic stroke," The BMJ. 2020, doi: 10.1136/bmj.l6983

[3] C. Chin et al., "An automated early ischemic stroke detection system using CNN deep learning algorithm," 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017, pp. 368-372, doi: 10.1109/ICAwST.2017.8256481.

[4] J. N. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning-based model for prediction of outcomes in acute stroke," Stroke, vol. 50, no. 5, pp. 1263–1265, 2019, doi: 10.1161/STROKEAHA.118.024293.

[5] L. von Rueden et al., ''Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," in IEEE Transactions on Knowledge and Data Engineering, 2021, doi: 10.1109/TKDE.2021.3079836.

[6] P. Shimpi, S. Shah, M. Shroff and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 603-607, doi: 10.1109/ICCMC.2017.8282537.

[7] Geoffrey I Webb, Naïve Bayes, Monash University, Springer, 2017

[8] S. S. Gavankar and S. D. Sawarkar, "Eager decision tree," 2017 2nd International Conference for Convergence in Technology (I2CT), 2017, pp. 837-840, doi: 10.1109/I2CT.2017.8226246.

[9] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019

[10] Fedesoriano, Stroke Prediction Dataset, Melbourne, Australia: Kaggle, 2015. Accessed on: September 25, 2021. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata

[11] A. Yudhana, D. Sulistyo, and I. Mufandi, "GIS-based and Naïve Bayes for nitrogen soil mapping in Lendah, Indonesia," *Sens. Bio-Sensing Res.*, p. 100435, 2021.

[12] H. Wu and F. J. Meng, "Review on evaluation criteria of machine learning based on big data," in Journal of Physics: Conference Series, 2020, vol. 1486, no. 5, p. 52026.

The Effect of Facial Attributes in Identifying Gender Using Facial Recognition

Athallah Muhammad Ariq[ab1], Yunardi Denni Tri[ab2], Zharief Dhiya Uzh[ab3], Herlina Napitupulu[ac4]
[a]*Department of Digital Business, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Mathematics, Universitas Padjadjaran*
Sumedang, Indonesiae-mail: [1]ariq19004@mail.unpad.ac.id, [2]denni19001@mail.unpad.ac.id, [3]dhiya19003@mail.unpad.ac.id, [4]herlina@unpad.ac.id

***Abstract*—Facial Recognition technology has become part of people's daily lives. This technology can be used in various ways such as face filter applications, mobile lock screens, to security cameras. This study aims to experiment in making models that use facial recognition attributes to identify gender by using five types of classification analysis methods. Based on the research that we conducted using several methods and measuring it based on the evaluation parameters, it can be seen that the model that uses Random Forest method has the highest level of accuracy among other methods, with an accuracy of 97.502%. Therefore, the Random Forest is the best model in determining the gender of the sample that we studied.**

Keywords— facial recognition, facial attributes, gender, classification analysis.

I. INTRODUCTION

Face recognition is a task that humans can easily perform in their daily lives. Even in adverse conditions, such as poor lighting or aging-related facial changes, humans can easily recognize faces [1]. Facial Recognition in recent years was not only an expensive technology but also less fast of speed in analysing compared to the human brain. However, after further development, it has become part of people's daily lives. This technology is used in various ways, such as face filter applications, mobile lock screens, to security cameras.

In the Facial Recognition measurement system, there are about 80 to 90 unique points on each of everyone's faces. Starting from the length of the jawline, the distance between the eyes, the shape of the cheekbones, to the depth of the eyes. These points will be measured and used as a code, named "faceprint", which will later identify the shape of the face and stored in a computer database [2].

There are at least two situations that benefit from Facial Recognition technology. First, its ability to identify the face of criminals. Facial Recognition can easily recognize the faces of criminals in public if the data has been previously recorded. By using the right regulations, this technology can be combined with public CCTV so that it can quickly detect if there are criminals. But this technology does not decide whether the criminal should be arrested or not. It will be the duty of the police to decide. However, it provides greater transparency in the decision-making process on the opportunity to stop and intervene in searches. Second, the advantage that can be obtained at the airport. Facial Recognition technology can be used for convenience in checking boarding passes, which is faster and more accurate than manual checking. The use of this technology can also increase security at the airport [3].

The help of machine learning is not only for analysing facial shapes but also for identifying human facial expressions. This method is usually called facial expression recognition (FER). By analysing multiple images of human faces, such as happy, angry, or scared, the FER method can help machines understand the types of human facial expressions. The potential application of FER is also quite influential, such as detection of abnormal human behaviour, and health management, and other similar tasks [4].

This study aims to create the optimal model that uses facial recognition attributes to identify gender. We use five types of classification analysis methods, namely Random Forest, Decision Tree, SVM, Probabilistic Neural Network (PNN), and RProp MLP. Random Forest, Decision Tree, and SVM are three types of classification analysis methods that are often used. We also experimented with using two methods that were rarely used, namely Probabilistic Neural Network (PNN) and RProp MLP.

II. METHODS

A. Nodes

1) Random Forest

Random decision forests also known as random forest is a method that operates by constructing multiple decision trees during the training phase. The decision of the majority of the trees is chosen by the random forest as the final decision, it has earned a wide interest in the research community that spawned a significant number of papers [5]. Random forest is a good way to prevent overfitting without sacrificing bias. It's simply a collection of decision trees whose results are aggregated into one final result. It is a strong ensemble modelling technique and much more robust than a single decision tree.

2) Decision Tree

Decision Tree Organizes a series of roots in a tree structure, it is one of the most practical methods for non-parametric supervised learning. Decision trees are commonly used for regression and classification problems due to their nice interpretability [6]. There are some important terminologies such as Root Node, Splitting, to determine the optimal splitting point, a splitting criterion is required [7], Decision Node, Leaf/Terminal Node, and Pruning.

3) Support Vector Machine (SVM)

Support Vector Machine is one of the best nonlinear supervised machine learning models. SVM is a widely-used machine learning method and has been successfully used in many fields such as face detection, handwritten digit recognition and text automatic classification, etc [8]. SVM

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

will help us identify an ideal hyperplane that categorizes fresh samples in one-dimensional space given a set of labelled training data. The hyperplane is a surface that divides a space into two parts where each class lies on either side.

4) Probabilistic Neural Network (PNN)

PNN is a modified version of the radial basis function that seeks to estimate the probability density function. PNN is a kind of feed-forward artificial neural network, which can approach a Bayes-optimal solution, and has been found suitable for pattern recognition and classification [9]. The advantage of a PNN is PNN generally trains faster and often produces more accurate models.

5) RProp MLP

The MLP method is used to create multilayer feedforward networks. MLP adjusts the weight updates on a local level based on the error function's behaviour. Multi-layer perceptron (MLP) is a neural network using simple models to investigate and solve difficult tasks such as prediction models. MLP develops robust algorithms to solve difficult problems [10].

B. Evaluation Parameters

In evaluating the model, we have made, we use four indicators as given below,

1. Accuracy is the ratio of the prediction of True Positive or True Negative with the overall data.
$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$
2. F-measure is a weighted comparison of the average precision and recall.
$$F - measure = \frac{2(Recall \times Precision)}{(Recall + Precision)}$$
3. Recall is the ratio of true positive predictions compared to the overall data that are true positive.
$$Recall = \frac{(TP)}{(TP + FN)}$$
4. Precision is the ratio of positive correct predictions compared to the overall positive predicted results.
$$Precision = \frac{(TP)}{(TP + FP)}$$

By using these four indicators as our reference in determining the evaluation parameters, we can more easily determine whether a model can be categorized as a good or bad model [11].

III. METHODOLOGY

A. Analysis

We are using an ASUS A409J laptop with an intel i3 10th gen processor. For data processing, we use the Knime application to analyse our research. Knime is considered an easier application to use because there is a visualization of the algorithm to be used and has a large enough community base to discuss everything in it. *Fig. 1* is a form of data processing flow that we have created. First, we import the dataset that we want to use and then check the quality of the data. Data cleaning is the process of detecting, diagnosing, and correcting incorrect data [12]. We were lucky to get a dataset that was clean from all kinds of abnormalities so that we could immediately carry out data processing with classification analysis. Before performing the nodes shown in *Fig. 1*, we use *partitioning* node to divide the dataset into 80% of the training dataset while the rest will be the testing dataset.

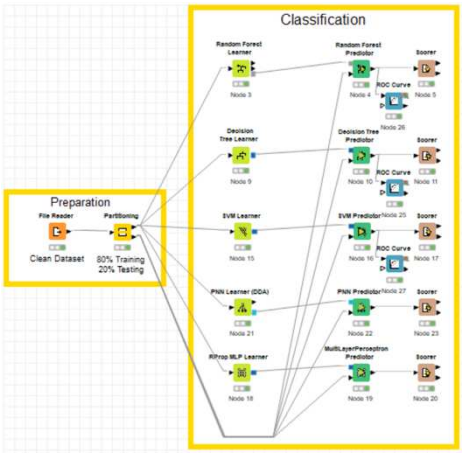


Fig. 1. Knime workflow

B. Dataset

We use a dataset on facial features for gender determination published by Jify Issadeen on Kaggle [13]. This dataset consists of eight label columns with data entry of 5001 entries. The eight columns include:

1. *long hair:* This column contains 0's and 1's where 1 is "long hair" and 0 is "not long hair".
2. *forehead width cm:* This column is in CM's. This is the width of the forehead.
3. *forehead_height_cm:* This is the height of the forehead and it's in centimeter
4. *nose wide:* This column contains 0's and 1's where 1 is "wide nose" and 0 is "not wide nose".
5. *nose long:* This column contains 0's and 1's where 1 is "Long nose" and 0 is "not long nose".
6. *lip_thin:* This column contains 0's and 1's where 1 represents the "thin lips" while 0 is "Not thin lips".
7. *distance to lip long:* This column contains 0's and 1's where 1 represents the "long distance between nose and lips" while 0 is "short distance between nose and lips"
8. *gender:* This is either "Male" or "Female".

The dataset contains several facial attributes that can be used as parameters in identifying the sample. With these attributes we will classify each sample as male or female.

IV. RESULTS

A. Results

We conducted several experiments on each method used. Initially, we tried to create a model according to the default parameters of each method. We use the results of these parameters as a basis for experimenting to find a model with a higher accuracy value than the basic results. We changed each parameter to see its impact on the accuracy results of the model under development.

1) Random Forest

We target the gender column as the target column that contains values for the model to learn. All attributes that exist in the dataset are used to be a model consideration in determining the score of the model to be made. We conducted

several split criterion experiments and the number of models used to obtain optimal results using the information gain ratio with a total of 800 models.

TABLE I. RANDOM FOREST CONFUSION MATRIX

Gender	Male	Female
Male	493	14
Female	11	483
Corrected Classified		976
Wrong Classified		25
Accuracy		97.502%
Error		2.298%

From Table I Knime workflow above, it can be seen that this Random Forest model has a high accuracy rate of 97.502%. Below is Fig. 2 which visualizes the accuracy of this model in classifying samples. For the female class, this model has a P of 0.996.

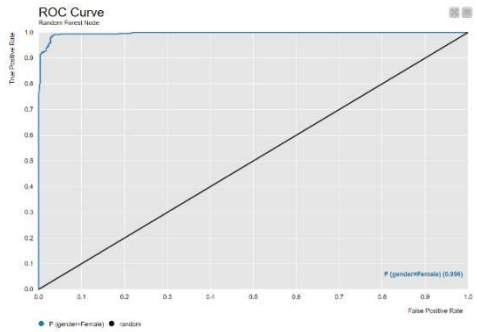


Fig. 2. ROC curve for random forest node

2) Decision Tree

The gender column remains the target of the use of this node. We use the gain ratio calculation and do pruning with the MDL method. By default, this algorithm will limit the size of the decision tree by pruning, splitting split points, and skipping columns that do not have a clear domain.

However, we decided not to split the split points to get better model accuracy. As a result, we get a model with an accuracy rate of 97.303%. We modeled the ROC curve as shown in Fig. 3. It can be seen that the curve in Fig 3, *P* has a curve that is close to 1 so that this model can be said to be a good model.

TABLE II. CONFUSION MATRIX OF DECISION TREE

Gender	Male	Female
Male	493	14
Female	13	481
Corrected Classified		974
Wrong Classified		27
Accuracy		97.303%
Error		2.697%

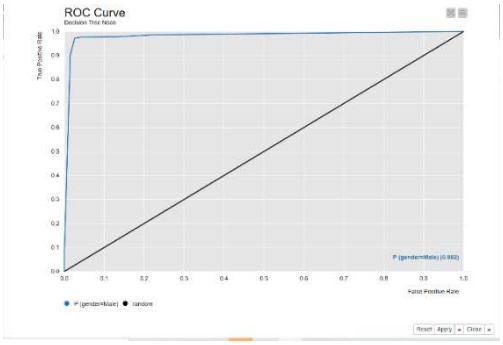


Fig. 3. ROC curve decision tree node

3) Support Vector Machine (SVM)

The class column from the use of this method is gender with an overlapping penalty of 1 because it is considered a good configuration in general. The kernel method used in this SVM method is a polynomial with the power, bias, and gamma parameters of each being 1.

TABLE III. CONFUSION MATRIX OF SVM

Gender	Male	Female
Male	490	17
Female	18	476
Corrected Classified		966
Wrong Classified		35
Accuracy		96.503%
Error		3.497%

From the confusion matrix above, it can be seen that this SVM model has a fairly high level of accuracy, which is 96.503%. To describe the level of model accuracy using the SVM method, you can use the ROC curve as shown in Fig. 4 below.

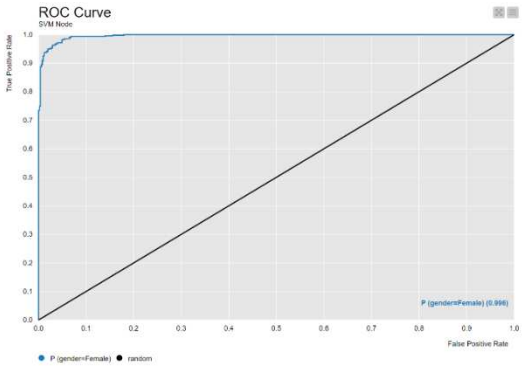


Fig. 4. ROC curve for SVM node

4) Probabilistic Neural Network (PNN)

In using PNN, the first thing that becomes the configuration of this node is how the algorithm overcomes missing values. We used the best guess method and checked the two advanced options but did not use the maximum number epoch method. The gender column is set as the calculation target column for this model. For the upper boundary of activation for conflicting rules (theta plus) we use the default configuration at 0.2 while for the lower boundary of activation for non-conflicting rules (theta minus) we set it at 0.3.

TABLE IV. CONFUSION MATRIX OF PNN

Gender	Male	Female
Male	493	14
Female	11	483
Corrected Classified		959
Wrong Classified		42
Accuracy		95.804%
Error		4.196%

5) RProp MLP

We made some parameter changes in this node configuration. Firstly, we increase the maximum number of iterations to 150 times, 2 hidden layers, and 15 hidden neurons per layer. The target column with this method is the gender column by ignoring missing values if any.

TABLE V. CONFUSION MATRIX OF MLP

Gender	Male	Female
Male	493	14
Female	13	481
Corrected Classified		968
Wrong Classified		33
Accuracy		96.703%
Error		3.297%

B. Comparison

In this section, we evaluate the results of the five models that have been created. We compare each model according to the evaluation parameters that we have set. The results of the comparison can be seen in the following table.

V. CONCLUSION

Facial Recognition technology in recent years was not only an expensive technology, but also its speed in analyzing was not as fast as the human brain. However, after further development, it has become part of people's daily lives. This technology can be used in various ways, such as face filter applications, mobile lock screens, to security cameras. This study aims to experiment in making models that use facial recognition attributes to identify gender by using five types of classification analysis methods, namely Random Forest, Decision Tree, SVM, Probabilistic Neural Network (PNN), and RProp MLP.

Based on the research that we conducted using several methods and measuring it based on the evaluation parameters, it can be seen that the results of the five methods we used have an accuracy rate close to the numbers ranging from 95-97%. But it also can be seen that the model using the Random Forest method has the highest level of accuracy among other methods with an accuracy rate of 97,502%. For other parameters such as precision, recall, and f-measure, Random Forest also remains at the top if the values of both classes are on average. The level of accuracy can be the main indicator when the levels of TP and TN have a close difference. Therefore, the Random Forest model is the best model in determining the gender of the sample that we studied.

TABLE VI. COMPARISON TABLE

No	Methods	Gender	Precision	Recall	F-measure	Accuracy
1	Random Forest	Male	0.978	0.972	0.975	0.975
		Female	0.972	0.978	0.975	
2	Decision Tree	Male	0.974	0.972	0.973	0.973
		Female	0.972	0.974	0.973	
3	SVM	Male	0.965	0.966	0.966	0.965
		Female	0.966	0.964	0.965	
4	PPN	Male	0.957	0.961	0.959	0.958
		Female	0.959	0.955	0.957	
5	MLP	Male	0.972	0.963	0.967	0.967
		Female	0.962	0.972	0.967	

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] K. Sato, S. Shah, and J. K. Aggarwal, “Partial face recognition using radial basis function networks,” in Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998, 1998, pp. 288–293.

[2] D. S. AbdELminaam, A. M. Almansori, M. Taha, and E. Badr, “A deep facial recognition system using computational intelligent algorithms,” *PLoS One*, vol. 15, no. 12 December, pp. 1–27, 2020.

[3] R. V. Virgil Petrescu, “Face Recognition as a biometric application,” J. Mechatronics Robot., vol. 3, no. 1, pp. 237–257, 2019.

[4] B. Yang, J. Cao, R. Ni, and Y. Zhang, “Facial expression recognition using Weighted Mixture Deep Neural Network based on Double-Channel Facial Images,” IEEE Access, vol. 6, pp. 4630–4640, 2017.

[5] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019

[6] C. H. Hsu, “Optimal decision tree for cycle time prediction and allowance determination,” IEEE Access, vol. 9, pp. 41334–41343, 2021.

[7] P. Fernandez-Gonzalez, C. Bielza, and P. Larranaga, “Random Forests for regression as a weighted sum of k-potential nearest neighbors,” IEEE Access, vol. 7, pp. 25660–25672, 2019.

[8] Z. Sun, K. Hu, T. Hu, J. Liu, and K. Zhu, “Fast multi-label low-rank linearized SVM classification algorithm based on approximate extreme points,” IEEE Access, vol. 6, pp. 42319–42326, 2018.

[9] M. Hao *et al.*, “Stokes space modulation format identification for optical signals using probabilistic neural network,” IEEE Photonics J., vol. 10, no. 3, pp. 1–13, 2018.

[10] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[11] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on

Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[12] J. Van Den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, “Data cleaning: Detecting, diagnosing, and editing data abnormalities,” PLoS Med., vol. 2, no. 10, pp. 0966–0970, 2005.

[13] J. Issadeen, “Gender classification dataset”, 2020, Version 1. Retrieved September 27, 2021 from https://www.kaggle.com/elakiricoder/gender-classification-dataset.

Analysis of Prediction Data for the Third Wave of COVID-19 in Bogor Regency

Alfian Fadhil Labib[ab1], Difa Bagasputra Maulana[ab2], Sina Mustopa[ab3], Intan Nurma Yulita[ab4]
Mulya Nurmansyah Ardisasmita[c5], Dwi Agustian[c6]
[a]*Department of Computer Science, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Public Health, Faculty of Medicine, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]alfian18003@mail.unpad.ac.id, [2]difa18002@mail.unpad.ac.id, [3]sina18001@mail.unpad.ac.id, [4]intan.nurma@unpad.ac.id, [5]mulya@unpad.ac.id, [6]dwi.agustian@unpad.ac.id

***Abstract* — The COVID-19 pandemic is far from over. The government has carried out several policies to suppress the development of COVID-19 is no exception in Bogor Regency. However, the public still has to be vigilant especially now we will face a year-end holiday that can certainly be a trigger for the third wave of COVID-19. Therefore, researchers aim to make predictions of the increase in positive cases, especially in the Bogor Regency area to help the government in making policies related to COVID-19. The algorithms used are Gaussian Process, Linear Regression, and Random Forest. Each Algorithm is used to predict the total number of COVID-19 cases for the next 21 days. Researchers approached the Time Series Forecasting model using datasets taken from the COVID-19 Information Center &Coordination Center website. The results obtained in this study, the method that has the highest probability of accurate and appropriate data contained in the Gaussian Process method. Prediction data on the Linear Regression method has accurate results with actual data that occur with Root Mean Square Error 1202.6262.**

Keyword— COVID-19, Gaussian Process, Linear Regression, Random Forest, Bogor Regency

I. INTRODUCTION

Corona virus or COVID-19 is a new type of coronavirus that is transmitted to humans. This virus can attack anyone, babies, children, adults, the elderly, pregnant women, and breastfeeding mothers. The COVID-19 infection was first discovered in the city of Wuhan, China, at the end of December 2019[1]. This virus spread quickly and spread to other areas in China and most countries in the world, including Indonesia. Several ways have been taken by the government, such as lockdown, PPKM, to mass vaccination, but of course this is not a guarantee that the corona virus will be finished in the near future.

Covid-19 cases in Indonesia have shown a significant decline. active cases that previously reached 500,000 are now below 150,000 [2]. The government also continues to extend the implementation of Community Activity Restrictions (PPKM) and implements a mass vaccination program, because it is considered successful in reducing COVID-19 cases.

However, PPKM has now been given a lot of leeway, including in Bogor Regency, such as shopping centers starting to operate again with looser restrictions. Bogor Regency administratively consists of 434 villages / villages (417 villages and 17 villages), with a total of 3,941 RW and 15,874 RT covered in 40 sub-districts. The largest area in Cigudeg District covers an area of 17,726.24 ha, consisting of 15 villages and the smallest area is Ciomas District which has an area of 1,810.36 ha, consisting of 10 villages and 1 village[3]. The easing of rules also applies to restaurants. Previously, the rule for eating in restaurants was only 30 minutes, now it has been extended to 60 minutes and offices in non-essential fields are allowed to WFO with a maximum capacity of 25%. Bogor Regency itself had previously been a black zone in the COVID-19 distribution area. Of course, if people remain disobedient in preventing COVID-19, it is not impossible that the third wave of COVID-19 in Indonesia will appear in the near future.

Currently, we are approaching year-end holidays in December. Where if we look at the cause of the surge in COVID-19 cases in the past occurring after a long holiday, the government must of course start to be vigilant and at least have to pay attention to the current development of COVID-19 until December. If the addition of corona cases increases again, this will cause problems for the handling of COVID-19 in Indonesia. Corona virus has many impacts, both in the fields of economy, education, and health. Not only mass layoffs, from the part of the business owners themselves, they also experience losses. In addition to layoffs, demand, supply, production, faltered. Then the businesses went bankrupt, yes that's for sure as we often see. Of course, with the reduction in cases of the spread of the corona virus, it can slowly help overcome these problems.

Therefore, researchers conducted COVID-19 research to predict positive cases, the results of which are expected to help generate data that will be analyzed later to prevent a spike in the third wave later. The study uses a dataset from the COVID-19 Information and Coordination Center, where the data will be processed using the Gaussian Process, Linear Regression, and Random Forest algorithms. In order to produce predictions for the number of spikes in cases, the results are expected to assist the government in making policies. to prevent the spread of COVID-19, especially in Bogor Regency.

II. LITERATURE STUDY

In conducting this research process, several methods were used to analyze the data contained in PIKOBAR (West Java COVID-19 Information and Coordination Center), namely:

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

A. Time Series Forecasting

Time series forecasting is an area of machine learning that focuses on the attribute of time. This field focuses on analyzing a series of data that is sequential over the time, then predicting future data based on previous data [4]. The results of these predictions are not always accurate predictions, and the possibility of forecasting predictions can vary greatly, especially when dealing with variables that generally fluctuate in time series data and factors beyond our control. However, an insightful estimate of which outcome is more likely—or less likely—to occur than any other potential outcome. Frequently, the more complete the data we have, the more accurate the estimate.

B. Gaussian Process

The Gaussian Process is an alternative approach to the artificial neural network model which was first proposed by O'Hagan (1978). The Gaussian process is a distribution of functions where the mean and variance are a certain function. Gaussian process regression can be explained from the point of view of nonparametric Bayesian regression, namely by placing directly the Gaussian prior distribution for the regression functions f(x)[5].

C. Linear Regression

Linear Regression is a statistical procedure to calculate the value of the dependent variable from other independent variables [6]. Linear Regression measures the relationship between two variables. This is a modeling technique in which the dependent variable is predicted based on one or more independent variables. Linear Regression here performs statistical tests that are applied to the data set to define and measure the relationship between the variables under consideration.

D. Random Forest

Random Forest is a Supervised Machine Learning algorithm used in Classification and Regression problems. This Random Forest method is an ensemble of learning methods using a decision tree as the base of the classifier to be built and combined [7]. Random Forest Algorithm, also known as random decision forests, is a particular type of ensemble algorithm that utilizes a combination of decision trees based on subsets of datasets. A Random Forest Algorithm does not create decision trees from smaller decision trees, but uses decision trees in parallel for prediction. The Random Forest Algorithm is also usually more accurate than a single decision tree although it is less intuitive [8].

There are three important aspects in using the Random Forest method, namely: First, doing bootstrap sampling to build a predictive tree. Second, each decision tree predicts a random predictor. Third, then forest random predicts by combining the results of each decision tree by means of majority vote for classification or average for regression [9].

III. METHODOLOGY

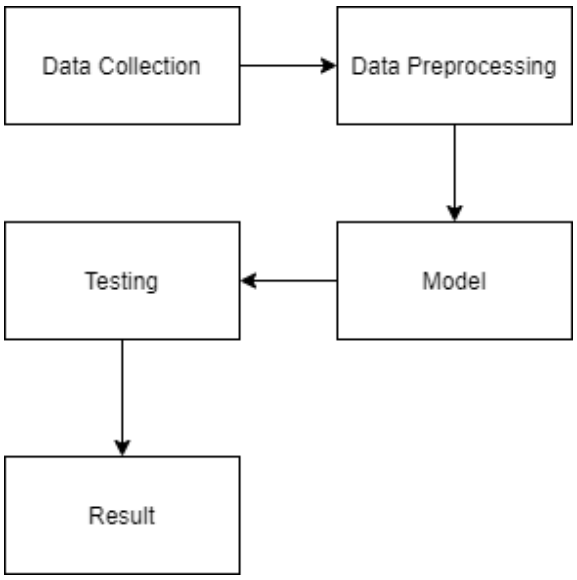


Fig. 1. Flowchart methodology flow

A. Description of Dataset

The dataset taken from the Pikobar website, contains attributes consisting of dates starting from August 1, 2020 to September 27, 2021, the code of the province of West Java, the name of the province, namely West Java, the code of the district or city in West Java, the name of the district in West Java, close contact or close contact with people affected by the covid-19 virus, close contact quarantined, close contact discarded or recovered, suspected, suspected isolated, suspected discarded, probable or possible covid-19 believed to be a suspect with severe ARI or respiratory failure due to fluid-filled lung alveoli (ARDS), probable isolated, probable discarded, probable death, confirmed, confirmed completed, confirmed death, close contact growth, discarded close contact growth, quarantined close contact growth, suspected growth discarded, growth suspect discarded, growth suspect isolated, probable growth, probable discarded growth, probability growth e isolated, probable growth died, confirmed growth, confirmed growth completed, confirmed growth died. The data contained is related to the growth of Covid-19 data in the West Java province, especially the data that will be used is data included from Bogor Regency data[10].

B. Attribute Variables

Attribute variables that will be used by researchers in this study are:

1) District Code, which contains the code of the district that will be tested by researchers.
2) Date , which is the date on which the addition of COVID-19 cases occurred.
3) Total confirmation, namely the total active cases in Bogor Regency.

In this test, the researcher only used the date attribute, district code, and total confirmation. It aims to collect data on positive active cases in Bogor Regency only.

C. Data Collection

The data collection process was taken from the website of the Information and Coordination Center (Pikobar: url (https://pikobar.jabarprov.go.id/table-case)) COVID-19 West Java Province and downloaded on 27 September 2021.

D. Data Preprocessing

The data collected from Pikobar was reduced back to the Bogor Regency dataset. There are several stages carried out in Data Preprocessing, including :

1) Data Cleaning

Perform data cleaning for missing values and data cleaning for noise data. It is necessary to systematically correct the data through algorithms. So it can save more costs and time.

2) Data Transformation

Convert the data into a form suitable for the analysis method. Here the researcher performs attribute selection to sort out which attributes will be analyzed.

3) Data Reduction

The data used is 423 rows, this is needed as data from Bogor Regency and to limit the data collection so as to increase storage efficiency while reducing the cost of money and time.

E. Model

The model formation process from the existing dataset is formed from the training data process using the Gaussian Process, Linear Regression, and Random Forest algorithms. Utilization of several algorithms is intended as a calibration of model rules to get optimal results in predicting data. The model was corrected using the Root Mean Squared Error (RMSE).

The RMSE calculation process is by subtracting the actual value from the forecast value then squared and adding up the overall results and then dividing by the amount of data. The results of these calculations are then calculated again to find the value of the root of the square root. The formula can be seen as follows[11]:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

F. Testing

The testing process is carried out using 3 algorithms, namely using the Gaussian Process, Linear Regression, and Random Forest algorithms. The specified data testing period will also be 21 days.

G. Result

Of all the processes carried out, through a data modeling of all the datasets available on the Pikobar website, through a data mining process, we get a prediction of the results and the formation of patterns that can be identified based on the planned size to obtain knowledge information as a result of data processing. mining.

H. Device and Application Specifications

1) Software used :

- Microsoft Excel

2) Computer application development hardware with minimum requirements :

- 2.6 GHz or faster processor
- RAM 4GB

IV. Results and Analysis

The data testing process is carried out with adjusted settings for each algorithm used. The customized settings are in the same Evaluation stage for each algorithm. For Gaussian Process, Linear Regression, and Random Forest use 0.01 for evaluation. The error calculation will be displayed later using RMSE.

TABLE I. RMSE Result (Root Mean Square Error)

Method	RMSE
Gaussian Process	5048.3356
Linear Regression	1202.6262
Random Forest	1891.9064

RMSE results against the applied method can be seen in table 1, can be seen the results of Linear Regression smaller than Random Forest this means the accuracy of linear regression algorithms is more accurate than Random Forest and Gaussian Process if you look at the results of RMSE (Root Mean Square Error).

The results of data processing and case testing with the model tested for the prediction process for the total confirmed Covid-19 data in Bogor Regency. The prediction results are obtained from September 28, 2021 to October 18, 2021 within a period of 21 days as follows:

Fig. 2. Linear Regression Data Validation Graph

The graphic depiction in Figure 2 shows a comparison of the data visualized with the red line as the real number of confirmed positive cases of Covid-19 and the blue line showing the prediction results of positive confirmed cases using the Linear Regression method. From the existing results, the actual data has insignificant differences which means that the prediction results from the method are quite accurate.

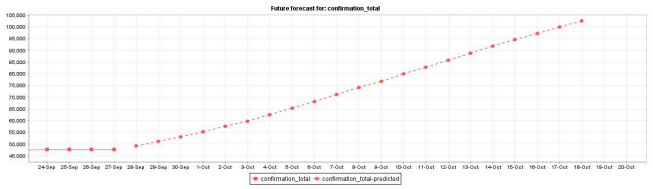


Fig. 3. Predictive data graph with Linear Regression

In Figure 3, after the training data process and data testing were carried out, the data obtained from the Covid-19 prediction in Bogor Regency using the Linear Regression method within a period of 21 days, starting from September 28, 2021 to October 18, 2021, the graph shows an increase in Covid-19 confirmed numbers. The data is entered into the following table :

TABLE II. PREDICTION DATA USING LINEAR REGRESSION IN 21 DAY

Date	Total Confirmation
28/09/2021	49263.8252
29/09/2021	51041.87
30/09/2021	53046.2784
01/10/2021	55320.0321
02/10/2021	57515.3846
03/10/2021	59694.6809
04/10/2021	62448.8786
05/10/2021	65235.5555
06/10/2021	68099.2656
07/10/2021	71128.0064
08/10/2021	74119.2222
09/10/2021	76817.6669
10/10/2021	79857.6643
11/10/2021	82836.3403
12/10/2021	85769.8411
13/10/2021	88759.7129
14/10/2021	91731.8106
15/10/2021	94482.0957
16/10/2021	97138.8918
17/10/2021	99879.0564
18/10/2021	102514.2663

Fig. 4. Gaussian Process Data Validation Graph

The graphic depiction in Figure 4 shows a comparison of the data visualized with the red line as the real number of confirmed positive cases of Covid-19 and the blue line showing the prediction results of positive confirmed cases using the Gaussian Process method. Just like Linear Regression, the actual data has insignificant differences which means that the prediction results from the method are quite accurate.

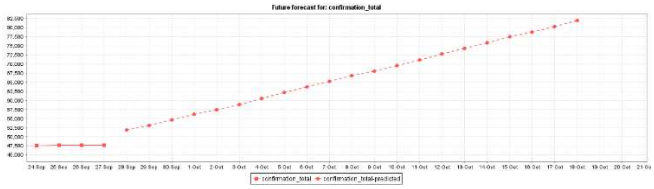


Fig. 5. Predictive data graph with Gaussian Process

In Figure 5, the results of the Covid-19 predictions in Bogor Regency have increased significantly from day to day with details in the following table :

TABLE III. PREDICTION DATA USING GAUSSIAN PROCESS IN 21 DAY

Date	Total Confirmation
28/09/2021	51802.0669
29/09/2021	53160.3779
30/09/2021	54616.9083
01/10/2021	56124.5457
02/10/2021	57377.4224
03/10/2021	58878.8462
04/10/2021	60538.6666
05/10/2021	62102.294
06/10/2021	63633.9467
07/10/2021	65202.0309
08/10/2021	66767.7102
09/10/2021	68032.7586
10/10/2021	69515.1241
11/10/2021	71144.2301
12/10/2021	72698.4081
13/10/2021	74254.5176
14/10/2021	75849.9052
15/10/2021	77448.2222
16/10/2021	78753.0539
17/10/2021	80282.7917
18/10/2021	81967.7051

Fig. 6. Random Forest Data Validation Graph

The graphic depiction in Figure 6 shows a comparison of the data visualized with the red line as the real number of confirmed positive cases of Covid-19 and the blue line showing the prediction results of positive confirmed cases using the Random Forest method. In the prediction results using this method, the visualization of predictive data has the same results and is in accordance with the actual data.

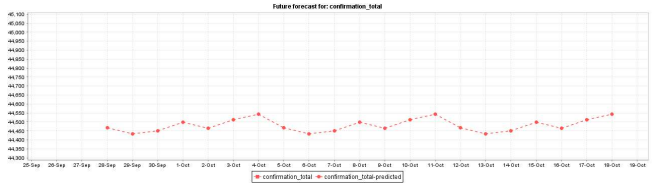


Fig. 7. Predictive data graph with Random Forest

In Figure 7, the data obtained from the Covid-19 predictions in Bogor Regency experienced a significant increase with details in the following table :

TABLE IV. PREDICTION DATA USING RANDOM FOREST IN 21 DAY

Date	Total Confirmation
28/09/2021	44468.6779
29/09/2021	44432.6159
30/09/2021	44449.9991
01/10/2021	44499.6871
02/10/2021	44465.8704
03/10/2021	44511.9344
04/10/2021	44544.1959
05/10/2021	44468.6779
06/10/2021	44432.6159
07/10/2021	44449.9991
08/10/2021	44499.6871
09/10/2021	44465.8704
10/10/2021	44511.9344
11/10/2021	44544.1959
12/10/2021	44468.6779
13/10/2021	44432.6159
14/10/2021	44449.9991
15/10/2021	44499.6871
16/10/2021	44465.8704
17/10/2021	44511.9344
18/10/2021	44544.1959

V. CONCLUSION

Based on the results of the tests that we have done using the Linear Regression Algorithm, Gaussian Process, and Random Forest, we get the results that the highest data accuracy according to RMSE Results is in the Linear Regression method with 1202.6262. But, if you look at the prediction results, the method that has the highest probability of accurate and appropriate data is found. on the Gaussian Process method. We can see this through the graph for each method, where on the data validation graph it can be seen that the visualization of predictive data in the Random Forest method has accurate results with actual data that occurs. After that, the prediction data graph shows data that is quite consistent on the prediction result graph in the Gaussian Process method.

With this conclusion, the government can find out the prediction of the increase in corona cases for the next 21 days. With these results, it can also be taken into consideration for the government, especially the Bogor Regency government, so that they can review the regulations that have been made and be aware of the coming Third Wave of Covid-19. Because according to the results of the prediction data from the study, data confirmed positive for Covid-19 showed a significant increase in the last 21 days.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] C. QIN, et al, " Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in Wuhan, China", Clinical infectious diseases, 71.15: 762-768, 2020. R. Tosepu, J. Gunawan, D. S. Effendy, H. Lestari, H.Bahar, and P. Asfian," Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia", Science of the total environment, 725, 138436, 2020.

[2] Portal Resmi Kabupaten Bogor. (n.d.). Retrieved October 26, 2021, from https://bogorkab.go.id/pages/pemerintahan.

[3] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu,"Financial time series forecasting with deep learning: A systematic literature review: 2005–2019", Applied Soft Computing, 90, 106181, 2020

[4] D. Maulud, and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", Journal of Applied Science and Technology Trends 1.4 : 140-147, 2020.

[5] K. Khushbu and Y. Suniti, “Linear Regression Analysis Study”. New Delhi, India: University of Delhi, 2018.

[6] V. Y. Kulkarni and P. K. Sinha, “Effective Learning and Classification using Random Forest Algorithm,” InternationalJournalof Engineering and InnovativeTechnology, vol. 3, no. 11, pp. 267–273, 2014.

[7] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019.

[8] R. Robi, N. Dwi, “Forecasting New Student Candidates Using the Random Forest Method,” Semarang, Indonesia: Universitas PGRI Semarang. DOI : 10.24843/LKJITI.2020.v11.i01.p05.

[9] I. N. Yulita et. al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", In Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[10] M. Ćalasan, S. H. A. Aleem, and A. F Zobaa, "On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function", Energy conversion and management, 210, 112716, 2020.

Classification of Water Potability Using Machine Learning Algorithms

Muhammad I'tikafi Khoirul Haq[ab1], Fauzian Dwi Ramadhan[ab2], Fatimah Az-Zahra[ab3], Linda Kurniawati[bc4], Afrida Helen[b5]
[a]*Department of Geophysics, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Business Administration, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]muhammad18360@mail.unpad.ac.id, [2]fauzian18002@mail.unpad.ac.id, [3]fatimah18001@mail.unpad.ac.id, [4]linda.kurniawati@unpad.ac.id, [5]helen@unpad.ac.id

***Abstract*— Clean water is one of the basic needs of everyday life. Recently, an ongoing process has been shown to improve water quality, making water less suitable for use. To solve this problem, research is done using a machine learning model. The Decision Tree Algorithm is used by Naïve Bayes algorithm in this type of machine learning to support drinking water quality. The two types of performance are compared in this work. K-fold cross credentials are used to evaluate our machine learning model. Results obtained in the decision tree algorithm have the best results in the configuration with an accuracy value of 97.23%.**

Keywords—classification, decision tree, k-fold cross-validation, machine learning, naïve bayes, water potability

I. INTRODUCTION

Water is the most important source of life to sustain many living things [1] . All living things on earth depend on water to grow and sustain life. In order to use this water in everyday life, it is significant to have access to clean water without contamination. Persistent pollution of water will affect the life of living organisms using water. Recent studies have shown that water quality continues to decline, making it less suitable for use [2]. The main reason for the decline in water levels is human and industrial activity. In addition, the use of chemical fertilizers for agriculture and changes in land use are other factors that also affect water quality [3] [4]. The World Health Organization (WHO) says that the challenges of providing safe drinking water for human needs are at an alarming rate. This is also due to the increasing pollution in urban areas [5]. Poor water management can hurt human life as an external cause of various diseases [6]. The United Nations (UN) reports that about 1.5 million people die each year due to waterborne diseases.

This drinking water problem really requires to be addressed so that it does not become a global problem. Therefore, the effort is required to maintain the quality of clean water. One way to maintain water quality is to look at the quality of the water and the environment around the equipment. Water quality testing is done to determine if the water is suitable for drinking or for humans. Tests are usually done by looking at the characteristics of water as a function of physical, biological, and chemical conditions. Water quality can be measured in several ways, one of which is based on the Water Quality Index (WQI). The water sample will be tested according to the parameter to be considered, but of course, the test takes a long time. Of course, other methods are needed to determine the water level quickly to solve this issue. The technique that can be used is the grading system using machine learning.

Various classification methods can be used to describe water power, including Decision Meter (DT), Naïve Bayes, Vector Machine Support (SVM), K-Nearest Neighbor (KNN), K-Means, and more. Numerous previous studies have been conducted on various aspects of machine learning design. Sulaiman et al. dividing water levels in Malaysia using an artificial neural network (ANN) [7]. They et al. compared SVM, KNN, and Naïve Bayes algorithms to extract water quality in India [1]. Radhakrishnan et al. conducted the water level analysis using SVM, DT, and Naïve Bayes methods [3]. Numerous previous studies have shown that mechanical processing systems can be used to obtain the right amount of water for consumption or use.

Therefore, in this project, we have put together two machine learning techniques that can be used to improve the quality of the water. Different classification methods encourage us to compare several classification methods that can be used. In this study, a classification system in the form of Decision Tree (DT) and Naïve Bayes (NB) was used to describe the quality of water energy. We have selected a decision tree algorithm that takes into account the benefits of this algorithm. This algorithm is non-parametric and does not require any consideration for the distribution of data entry, and requires minimal knowledge of data preparation to work effectively with a large data set. We also selected Naïve Bayes as a comparison algorithm because this algorithm is one of the most efficient algorithms. Naïve Bayes' main idea is that each character is independent, and their needs are equal. K-fold cross-validation validation is implemented in this function to test the model's performance. We consider some advantages of using cross-validation since it helps determine the model's strength to some extent. Cross-validation is also the key to determining the degree of model overfit [8].

II. MATERIALS AND METHOD OF ANALYSIS

A classification is a form of data analysis in extracting a model used to predict class labels [9]. There are two steps in data classification; the first is training data, which is the step in forming a classification model and using the model to predict the class on the data. In this study, two algorithms will be used to classify the data; there are Decision Tree algorithm and the Naïve Bayes algorithms.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

A. Decision Tree

The decision tree algorithm is the most commonly used decision-making and ranking system. This algorithm will find a solution to the problem by integrating the parameters of the nodes to form a plant system [10]. Decision trees are a predictive model of decision-making using a process-based or systematic approach. Each tree has branches that represent a characteristic that must be filled to move on to the next branch until it ends in a leaf where there are no other branches. The concept of data in the decision tree is that of the data displayed in the form of a table composed of attributes and records. These attributes such as the parameters used are used as criteria when designing a tree.

B. Naïve Bayes

Naïve Bayes is a statistical method by calculating the possibility of similarities between old cases and new cases in the case database [11]. Naïve Bayes has a high degree of accuracy and good speed when compared to large databases. Naïve Bayes included a supervised learning environment where early data were sought during learning in the form of training data to conduct classification. During classification, the probability value of each class in the given entry will be calculated. A class that is as probable as the class from the input data will be used. Naïve Bayes is the simplest Bayes theorem collection, as it is able to reduce the complexity of the collection as easily as possible. Naïve Bayes' design is as follows [11]:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (1)$$

Where the X is the dataset, C_i is the i-th pattern solution class, where i is the number of class label data (given class) and the $P(C_i \mid X)$ is the probability of input criteria X with class label C_i .

C. K-fold Cross-Validation

In this study, we used K-fold cross-validation to evaluate model performance. K-fold cross-validation or more commonly known as cross-validation is one of the most popular methods used in the analysis of ML models. In the backup, the K folds are divided into data sets. In each case, each ring will be used once as test data and the remaining bends will be used as training data. By passing the confirmation, we get more accurate analysis results because the model is analyzed with all the data. Determining the best K value in itself depends on a number of factors, including the amount of data, the number of variables that depend on it, and the number of self-changes. Overall, for large data (over 5000) the value K = 5 is sufficient to provide better results. The number of datasets less than 5000 is usually used for K = 10, but the choice of K should also take into account the variables in the used dataset [8].

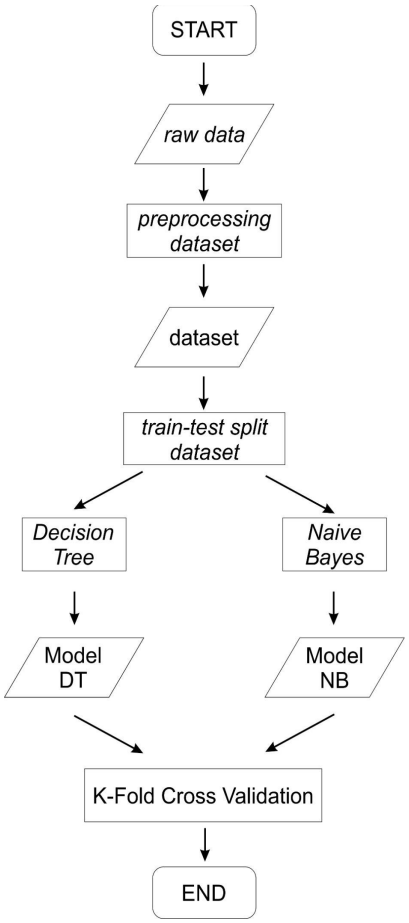


Fig. 1. Flowchart of this work

D. Dataset

A public dataset from kaggle [12] has been considered for training and testing machine learning models. This dataset contained nine features and two labels as indicators of the suitability of water to be used. There are pH values, Hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, potability. The pH value is data on the degree of acidity, which can be an indicator of acid or alkaline conditions of water. The turbidity data depends on the amount of solids present in the suspended state, which is used to indicate the quality of waste disposal with respect to colloidal matter. Solid TDS data show the ability of water to dissolve various organic and organic minerals. Water with a high TDS value indicates the water is highly mineralized. The hardness data is defined as the capacity of water to precipitate soap caused by Calcium and Magnesium. The sulfate content data show how much of the substance is in the water. The chloramine content data is to determine how much chloramine content is in water, which is usually used as the main disinfectant in public water systems. The conductivity data show how the nature of water in conducting electricity, where pure water is a good insulator with increasing ion concentration, will increase the conductivity of the water. The organic carbon content data show a measure of the total amount of carbon in organic compounds in pure water. The trihalomethanes (THM) content data can indicate the amount of chlorine and the temperature that need to be used to treat the water. The data has been equipped with a label consisting of two labels, there are labels 0 for 0 for non-potable water and 1 for potable water.

III. METHODOLOGY

The objective of this work is to develop a machine learning model to classify water potential with the flow, as shown in Fig. 1. Raw data obtained from public datasets will be preprocessed data by eliminating missing values in the dataset with total 2011 datasets. The dataset then separated for the training and testing process. At this stage, we try several variations for the proportion of tests to get the best accuracy in each classification algorithm. We use a percentage test size of 25% for Decision Tree and 20% for Naïve Bayes. For the Naïve bayes model, we use four types of Naïve bayes to see which Naïve bayes model has the best accuracy. The Decision Tree and Naïve Bayes models were evaluated using k-fold cross-validation with K = 5 and K = 10. Determining the value of k considers the number of datasets, the number of dependent variables, and the independent variable [8]. Machine learning model created in the Jupyter Notebook environment using the python programming language version 3.8.5 and the Scikit-learn library. The computer specifications used to create and run the machine learning model created in this study use an AMD A9-9425 processor with a frequency of 3.1 GHz with 2C + 3G compute core and 8 GB RAM.

IV. RESULT AND DICUSSION

In this work, we use two machine learning algorithms consisting of Decision Tree and Naïve Bayes to classify water potability data. The machine learning model validation in this study uses K-Fold Cross validation with K=5 and K=10 for each generated model.

A. Decision Tree

Classification was performed using a resolution tree with a test rate of 25%. The model is then validated using K-Fold cross-validation with two different values of K. support using K-fold cross-validation with K = 5 and K = 10 given as shown in Table I. All confirmation results using this decision tree show that the performance of the model is considered very good because it produces an average of 96.82% for K = 5 and 97.22% for K = 10. This high can be caused by a large. The number of data sets used to model as well as the small number of variables based on it is only 2 depending on the variance. We make a confusion matrix to see the performance of the decision tree model. Fig. 2 shows the results of a confusion matrix of decision trees, while the model is almost entirely predictable with label.

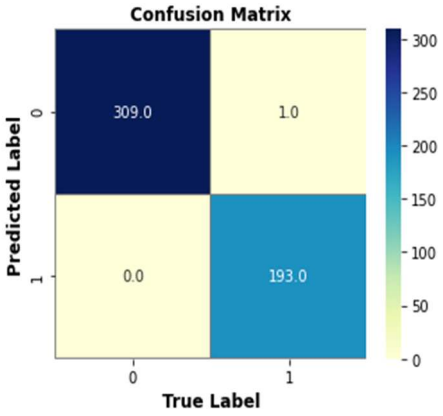


Fig. 2. Confusion Matrix results for the Decision Tree model

TABLE I. MODEL ACCURACY USING K-FOLD CROSS VALIDATION WITH K=5 AND K=10 FOR THE DECISION TREE MODEL

No	K-Fold Cross Validation Accuracy	
	k=5	k=10
1	0.96039604	0.98039216
2	0.96039604	0.96078431
3	0.99009901	0.98039216
4	0.99	0.92
5	0.94	1
6		1
7		0.96
8		1
9		0.96
10		0.96
Average	0.968178218	0.972156863

B. Naïve Bayes

The classification was performed using the Naïve Bayes algorithm by generating four different types of NB with the same test rate, which was 20%. Four types of NB models were performed in this study, namely: Gaussian, Bernoulli, Multinomial and Complement. All NB type results were obtained using the K-Fold cross confirmation, as well as K = 5 and K = 10. The accuracy of each type is shown from Table II-V where it can be seen that each NB type has different performance. The highest average was obtained by the Bernoulli Naïve Bayes model using K = 5 with an average of 62.53%. This accuracy value indicates the performance of the NB model which is much lower than the performance of the Model Plant Resolution, Also for Multinomial and supplementary models, the precision is even lower. We also compute statistical metrics to evaluate the performance of the machine learning model of two Naïve Bayes models with the highest precision, Gaussian Naïve Bayes and Bernoulli Naïve Bayes. Table VI and VII show that the Gaussian Naïve Bayes and Bernoulli Naïve Bayes models perform poorly in label 1 classification, which can be seen from the Memory, Precision and F1-Score values. We conclude that these factors are reduced compared to the performance of these two models.

C. Performance comparison of Decision Tree and Naïve Bayes models

Based on the results of the analysis of water potability data, the performance of the model using the tree decision algorithm was found to be better than the model using the Naïve Bayes algorithm. The highest average for the Tree Decision type was 97.22%, while the highest average for the Naïve Bayes type was 62.53%. These results can be analyzed that Decision Tree has the best performance due to, in this case, the dataset only consists of two labels. This trend is also similar to previous research conducted by Radhakrishnan et al [2] which showed that in this case the Decision Tree algorithm

gave the best performance compared to other classification algorithms.

TABLE II. MODEL ACCURACY USING K-FOLD CROSS VALIDATION WITH K=5 AND K=10 FOR EACH TYPE OF GAUSSIAN NAÏVE BAYES MODEL

No	Gaussian Naïve Bayes	
	K-Fold Cross Validation Accuracy	
	k=5	k=10
1	0.65432099	0.65853659
2	0.56790123	0.58536585
3	0.7037037	0.51219512
4	0.525	0.6
5	0.6	0.75
6		0.75
7		0.6
8		0.55
9		0.55
10		0.65
Average	0.610185	0.620609756

TABLE III. MODEL ACCURACY USING K-FOLD CROSS VALIDATION WITH K=5 AND K=10 FOR EACH TYPE OF BERNOULLI NAÏVE BAYES MODEL

No	Bernoulli Naïve Bayes	
	K-Fold Cross Validation Accuracy	
	k=5	k=10
1	0.61728395	0.6097561
2	0.62962963	0.63414634
3	0.62962963	0.63414634
4	0.625	0.625
5	0.625	0.625
6		0.625
7		0.625
8		0.625
9		0.625
10		0.625
Average	0.62530864	0.610185

TABLE IV. MODEL ACCURACY USING K-FOLD CROSS VALIDATION WITH K=5 AND K=10 FOR EACH TYPE OF MULTINOMIAL NAÏVE BAYES MODEL

No	Multinomial Naïve Bayes	
	K-Fold Cross Validation Accuracy	
	k=5	k=10
1	0.58024691	0.65853659
2	0.54320988	0.48780488
3	0.44444444	0.53658537
4	0.55	0.575
5	0.5375	0.4
6		0.475
7		0.5
8		0.575
9		0.525
10		0.55
Average	0.53108	0.53108

TABLE V. MODEL ACCURACY USING K-FOLD CROSS VALIDATION WITH K=5 AND K=10 FOR EACH TYPE OF COMPLEMENT NAÏVE BAYES MODEL

No	Complement Naïve Bayes	
	K-Fold Cross Validation Accuracy	
	k=5	k=10
1	0.58024691	0.65853659
2	0.54320988	0.48780488
3	0.44444444	0.53658537
4	0.55	0.575
5	0.5375	0.4
6		0.475
7		0.5
8		0.575
9		0.525
10		0.55
Average	0.53108	0.53108

V. CONCLUSION

In this study, a model for classifying water potability has been made using two machine learning algorithms, there are Decision trees and Gaussian Naïve Bayes with the validation method used is k-fold cross-validation. The two models produce different performance. The Decision Tree model has very good results with an accuracy value of 97.22%. The Gaussian Naïve Bayes model is a better model than other Naïve Bayes models but has poor results with an accuracy value of 62.53% where most of them fail to classify water potability.

TABLE VI. STATISTICAL METRICS OF GAUSSIAN NAÏVE BAYES

Gaussian Naïve Bayes				
Labels	**Precision**	**Recall**	**F1-Scores**	**Support**
0	0.67	0.90	0.77	252
1	0.62	0.27	0.38	151
Accuracy			0.67	403
Macro Avg	0.65	0.59	0.57	403
Weight Avg	0.65	0.67	0.62	430

TABLE VII. STATISTICAL METRICS OF BERNOULLI NAÏVE BAYES

Bernoulli Naïve Bayes				
Labels	**Precision**	**Recall**	**F1-Scores**	**Support**
0	0.63	1.00	0.77	252
1	0.00	0.00	0.00	151
Accuracy			0.63	403
Macro Avg	0.31	0.50	0.38	403
Weight Avg	0.39	0.63	0.48	403

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] T. Aldhyani, M. Al-Yaari, H. Alkahtani and M. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," Applied Bionics and Biomechanics, vol. 2020, 2020.

[2] N. Radhakrishnan and A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," in 5th International Conference on Communication and Electronics Systems (ICCES), 2020.

[3] Q. Zhang, X. Shi, B. Huang, D. Yu, I. Öborn, K. Blombäck, H. Wang, T. F. Pagella and F. L. Sinclair, "Surface water quality of factory-based and vegetable-based peri-urban areas in the Yangtze River Delta region, China," Catena, vol. 69, no. 1, pp. 57-64, 2007.

[4] M. Hussain, S. M. Ahmed and W. Abderrahman, "Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia," Journal of environmental management, vol. 86(1), pp. 297-307, 2008.

[5] M. Forde, R. Izurieta, B. Ôrmeci, M. Arellano and Mitchell, Water Quality in the Americas, The Inter-American Network of Academies of Sciences IANAS, 2019.

[6] T. Fredrick, M. Ponnaiah, M. V. Murhekar, J. K. David, S. Vadivoo and V. Joshua, "Cholera outbreak linked with lack of safe water supply following a tropical cyclone in Pondicherry, India, 2012.," jounal of healt, population, and nutrition., vol. 33(1), p. 31, 2012.

[7] Sulaiman, Khadijah, L. H. Ismail, M. A. M. Razi, M. S. Adnan and R. Ghazali, "Water Quality Classification Using an Artificial Neural Network (ANN)," in IOP Conference Series: Materials Science and Engineering, 2019.

[8] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," Computational Statistics, vol. 36, 2021.

[9] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[10] Babič, Š. Hleb, P. Kokol, V. Podgorelec, M. Zorman, M. Šprogar and M. M. Štiglic, "The art of building decision trees," Journal of medical systems, vol. 24, pp. 43-52, 2000.

[11] D. Berrar, "Bayes' theorem and naive Bayes classifier," Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, pp. 403-4012, 2018.

[12] A. Kadiwal, "kaggle," 25 April 2021. [Online]. Available: https://www.kaggle.com/adityakadiwal/water-potability. [Accessed 25 September 2021].

Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle Coupon Recommendation Data

Dicky Rahma Hermawan[ab1], Mohamad Fahrio Ghanial Fatihah[ab2], Linda Kurniawati[bc3], Afrida Helen [ab4]
[a]*Department of Computer Science, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Business Administration, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]dicky19002@mail.unpad.ac.id, [2]mohamad19001@mail.unpad.ac.id, [3]linda.kurniawati@unpad.ac.id, [4]helen@unpad.ac.id

***Abstract*— Coupons are one of the media used to increase sales and invite customers to repurchase products. A study to investigate the effectiveness of the distribution of coupons, especially coupons for restaurants and bar, can be carried out by collecting data through an in-vehicle survey. The data can then be analyzed using classification techniques in data mining. This paper presents a classification on the problem of in-vehicle coupon recommendation to determine the decision of coupon acceptance through the J48, Random Tree, and Random Forest decision tree classification algorithm. The dataset used consists of 23 attributes including the class Y attribute which indicates the receipt of coupons by customers. The performance of the three algorithms is evaluated to determine the best classification algorithm by looking at accuracy, time to build the model, and other variables that appear in the class classification experiment. The results reveal that the Random Tree classification algorithm takes the least amount of time (0.28 seconds) and has the lowest accuracy (67.38%). The J48 algorithm is more accurate than the Random Tree algorithm (72.79%) but takes significantly longer time (0.36 seconds). The Random Forest technique has the best accuracy (77.0%), but the time it takes for model creation is substantially longer than the Random Tree and J48 algorithms (10.89 seconds).**

Keywords—classification, data mining, J48, random tree, random forest

I. INTRODUCTION

Coupons are one of the media used to increase sales. Coupons are distributed with the intention of inviting potential customers to make more purchases. The coupon distribution target can be in the form of regular customers or non-customers. When a customer makes a trial purchase on a product being sold, it is also expected that the product will continue to be purchased by the customer. Coupons can also be distributed to specific target customers, for example the best customers with pre-set preferences. This is considered to have built a lot of customer loyalty [1]. Several studies have attempted to investigate how the attitude of potential customers when a coupon is given to them as an antecedent variable and whether the coupon will be used. To collect customer data, especially in different scenarios, the questionnaire method is widely used. Through the data collected, an analysis was carried out to investigate the correlation of the response variations from customers to the effectiveness of coupon distribution. In case of analyzing data and concluding it, data mining is an effective tool.

Data mining is a method that aims to find and examine a structured pattern in the data to gain an understanding. The basic principles in data mining include analyzing data through various directions, categorizing, and finally making conclusions from visible patterns. The pattern can be broken down into information to predictions. In the case of machine learning and data mining, the knowledge structures and structural descriptions obtained in a data are as important as their ability to sample new data. In this case, people use data mining not only to make predictions, but to gain new knowledge that emerges from a dataset [2].

Data mining can also be referred to as *knowledge discovery in database* (KDD) which follows the following steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evolution, knowledge evolution, and data reduction. There are also many techniques in data mining, including classification, clustering, data pre-processing, pattern recognition, association, and visualization. Classification is a process in data mining to describe and distinguish data classes and concepts through searching for classification models [3]. The purpose of the classification is to use a model that has been successfully searched to determine the class of unknown data labels. Currently, there are several classifier algorithms available. The classifier that is commonly used and has good performance is a decision tree [4]. Decision trees are expressive enough to model many data partitions that cannot be achieved in classifiers such as logistic regression and Support Vector Machines (SVM) which only rely on one decision scope. Decision trees are flexible in processing data with a combination of real types and categorical features and attributes that have missing data.

Classifiers such as J48, Logistic Model Tree (LMT), Random Tree, Simple Cart, Random Forest, and Reduced Error Pruning (REP) Tree are part of a decision tree and are used for the purpose of classifying datasets. In the decision tree, the results of the analysis are in the form of sequential rules that lead to a certain class or value and form a tree structure. A positive classification is obtained if there is a rule path to a positive leaf [5]. This can be used to derive conclusions from datasets that have classes with dichotomous labels such as yes and no.

In this study, we use the in-vehicle coupon recommendation dataset as a data source in conducting comparative analysis on decision tree-based classification

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

algorithms, namely J48, Random Tree, and Random Forest. We are comparing these three data mining classifiers to attain the most accurate decision tree classifiers for this type of case.

II. Research Methods

A. J48 Algorithm

J48 is a form of classifier that uses the C4.5 algorithm and is part of the classification method in data mining. The C4.5 method is a well-known and commonly used technique for categorizing data with numerical and categorical properties. The results of the classification procedure within the sort of rules are frequently utilized to estimate the value of the new record's discrete type property. The C4.5 algorithm is a progression of the ID3 algorithm. The development is carried out in terms of the ability to overcome missing data, the ability to handle continuous data, and the ability to prune. The C45 algorithm has an advantage over the ID3 algorithm. The advantage is in the way it gathers data. Because C4.5 employs the gain ratio as a metric of attribute selection, C4.5 is superior [6].

The J48 method works on the principle of dividing data into ranges based on attribute values for items in the training data set. Missing values, which are values for elements that can be predicted based on what is known about attribute values in other rows, are ignored by the J48 method [7].

J48 uses a greedy technique, in which decision trees are built by recursively separating attributes from top to bottom, with the topmost attribute being the most influential attribute of the attributes beneath it. J48 employs the pessimistic pruning approach, which determines whether to prune a portion of the tree depending on the expected error rate. The J48 algorithm starts with a root node, which is then subdivided into another section of the node as a result of evaluating the attribute variable to see if it meets the test value. If the test result is a node that can be tested again, it is referred to as a branch; if it cannot be tested again or is the final result, it is referred to as a leaf, also known as a label or class [10].

The stages of the J48 algorithm in the construction of a decision tree are as follows:

1. Make attribute the root attribute.

2. For each value, make a branch.

3. Separate the cases into branches.

4. Repeat the process for each branch until all cases on that branch have the same class.

The largest gain value of the existing attributes is used to choose an attribute as the root. The following equation is used to compute the gain [8]:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si) \quad \dots (1)$$

Where:

S = set of instances

A = attribute

n = number of partitions of A

|Si| = number of cases on partition i

|S| = number of cases in S

The basic formula for the entropy is as follows:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2 pi \quad \dots\dots\dots\dots\dots (2)$$

Where:

S = case set

A = feature

n = number of partition S

pi = proportion of Si to S

B. Random Forest Algorithm

Random Forest is a classifier made up of a series of tree-structured classifiers {h (x, k) k=1, 2, ….}, where {Θk} is a uniformly distributed random vector and each tree awards one vote unit to the most popular class at input x. Random Forest is made up of a huge number of individual decision trees that work together as a set, as the name implies. In Random Forest, each tree generates a class prediction, with the most votes being the model prediction [8].

The key of Random Forest is that each model has a low correlation. An uncorrelated model can generate ensemble forecasts that are more accurate than any individual prediction, just as low-correlated assets (such as stocks and bonds) combine to create a portfolio that is larger than the sum of its parts. The trees will shield one other from each other's errors if they don't continually make the same mistakes. Therefore, an uncorrelated model can produce ensemble predictions that are more accurate than any single prediction. While some trees will be incorrect, many others will be correct. As a result, the trees might move in the same direction as a group. Random Forest constructs each tree using two methods: bagging and feature randomness, to generate a forest of uncorrelated trees [9].

Because decision trees are highly dependent on the data they are trained on, even little modifications to the training set might result in drastically different tree architectures. Random Forest takes use of this by allowing each tree in the data set to be randomly sampled via replacement, resulting in a new tree, this process known as bagging (bootstrap aggregation). When splitting a node in a normal decision tree, it considers all candidate features and chooses the one that creates the maximum separation between the observations at the left and right nodes. Each tree in the Random Forest, on the other hand, can only choose from a random subset of features. This causes greater variety among the model's trees, which leads to a reduced correlation between them and increased diversification, this process known as feature randomness.

Trees are not only trained on diverse data sets (bagging) in the Random Forest algorithm, but they also use distinct features to make judgments (feature randomness). As a result, an uncorrelated tree emerges, which both supports and protects each other from errors.

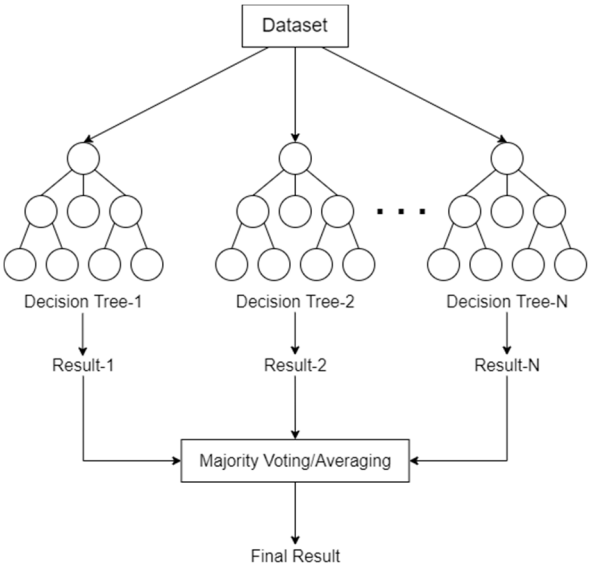


Fig. 1. Diagram of Random Forest algorithm

C. *Random Tree Algorithm*

Random Tree is a supervised classifier that produces many unique learners. The Random Tree algorithm is a type of ensemble learning algorithm. It builds a decision tree using the principle of bagging to generate a random set of data. Each node in a traditional tree is split using the best split across all variables.

The Random Tree algorithm can deal with both classification and regression problems. Leo Breiman and Adele Cutler invented this algorithm. A forest is an ensemble of prediction trees known as a random tree. The following is how the classification works: Random Tree takes an input feature vector, classifies it with each tree in the forest, and then assigns the class label with the most "votes" to the input feature vector. The Random Tree answer in regression is the average of all the responses from all the trees in the forest.

Random Tree is a machine learning algorithm that combines two algorithms: a single model tree and the Random Forest concept. A model tree is a decision tree that has a linear model for each leaf that is optimized for the local subspace specified by that leaf. Random Forests have been demonstrated to considerably increase the performance of single decision trees, using two randomization methods used to generate tree diversity. As in bagging, the training data is first sampled using a substitute for every single tree. Second, instead of computing the best split for each node all the time, when building a tree, only a random subset of all attributes is considered for every node, and thus the best alternative for that subset is calculated. Random Tree takes a different strategy, splitting the median of several attributes to roughly offset the trees. Recently, the approximate approach for calculating medians was described. As long as the data is close to the median, this process only requires two linear scans [10][11].

III. METHODOLOGY

The flowchart to find the model in this study using data mining for each classifier is as follows.

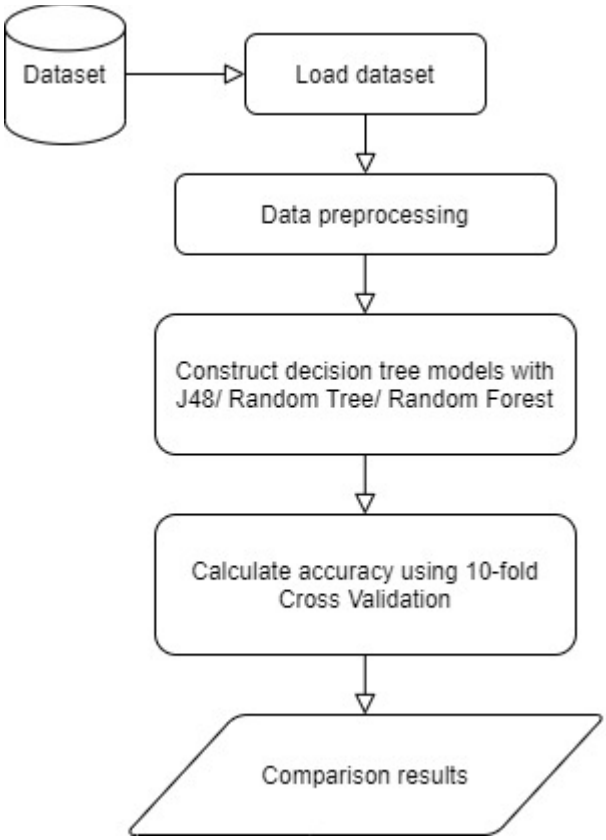


Fig. 2. Data mining process flowchart

The method proposed for classification of in-vehicle coupon recommendation dataset is implemented using a computer with AMD Radeon R2 with 2GB of memory, 4 GB DDR3-SDRAM memory, and AMD E2 1.5GHz processor.

A. *Dataset Preparation*

The dataset used in this study was sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation) which was collected through a survey on Amazon Mechanical Turk, a crowdsourcing website that owned by Amazon. There are 12,684 rows of data and 26 attributes in the dataset. The problem of this classification experiment is to see whether potential customers will accept the coupons offered under certain circumstances, such as the place of validity of the coupons, demographic attributes, and certain contextual attributes. For a positive answer which means that the potential customer will 'redeem the coupon immediately' or 'exchange it later but before the expiration date' is labeled with 'Y = 1'. For negative answers which mean the potential customer is 'not interested in receiving or exchanging coupons' is labeled with 'Y = 0'. The list of each attribute and their meaning are as follows.

- destination: Driving destination
- passenger: Passengers in the vehicle
- weather: Weather conditions when driving
- temperature: Temperature (°F) when driving
- time: Time when driving
- coupon: Coupon exchange type

• expiration: Does the coupon expire in one day or in two hours?

• gender: Gender of the driver

• age: Age of the driver

• maritalStatus: Marital status of the driver

• hasChildren: Does the rider have children?

• education: Education status of the driver

• occupation: Job role of the driver

• income: Yearly income of the driver

• car: Type of the vehicle used when driving

• Bar: How many times does the driver go to the bar every month?

• CoffeeHouse: How many times does the driver go to the coffeehouse each month?

• CarryAway: How many times does the driver order take away food each month?

• RestaurantLessThan20: How many times have the driver go to a restaurant that averaged less than $20 per person per month?

• Restaurant20To50: How many times have the driver go to a restaurant that averaged less than $20 - $50 per person per month?

• toCouponGEQ5min: Does it take more than 5 minutes to get to the restaurant/bar to redeem the coupon?

• toCouponGEQ15min: Does it take more than 15 minutes to get to the restaurant/bar to redeem the coupon?

• toCouponGEQ25min: Does it take more than 25 minutes to get to the restaurant/bar to redeem the coupon?

• directionsame: Is the restaurant/ bar in the same direction as the current destination?

• direction_opp: Is the restaurant/ bar in a different direction from its current destination?

• Y: Will the coupon be accepted and exchanged by the driver?

B. *Data Preprocessing*

First, data preprocessing on the dataset is carried out to improve the quality of the data. Things that are done at this stage include:

• Remove redundant attributes such as an attribute containing only one unique value (*toCouponGEQ5min*) and an attribute whose value is complementary to another attribute (*directionsame* with *direction_opp*).

• Removing attribute that have a very large amount of empty data (> 90%) (*car*).

• There are several types of attributes that are recognized as strings with several different data values < 10. These attributes are then converted into nominal type.

• There are several attributes that have an empty data row (missing). Therefore, the attributes with empty data are filled with the value that has the highest number of occurrences.

The remaining attributes are 23 and have nominal type. This data is ready to be used for the classification stage.

C. *Training the Model and Testing Using 5-fold, 7-fold, and 10-fold Cross Validation*

The preprocessed data are then loaded to train data mining algorithms, namely J48, Random Tree, and Random Forest for classification purposes. Parameters for each algorithm are set to default values.

After training, the classification model is then evaluated using stratified k-fold cross validation. This technique works by partitioning the dataset into k parts. Each part is used as test data and the rest is used as training data until k repetitions are completed. This time, we chose values of k=5, k=7, and k=10. In the field of applied machine learning, k=10 is a popular choice. 10-fold cross validation is carried out because it is suitable to avoid biased results and can produce good classification results. Hyperparameters tuning is possible only with original training set. This allows keeping test set as an unseen dataset for selecting the final model. Through this technique, all the correct or incorrect classification results from each iteration will be recorded and then the classification accuracy will be generated. The best classifier is known by comparing the accuracy values of each algorithm.

D. *Search for the Most Important Attributes (Feature Selection)*

After evaluating the model, feature selection will be carried out to find the attributes that are most important or have the most impact on the classification model. Feature selection was done by calculating the information gain. The information gain (also known as entropy) for each attribute of the output variable was calculated. The values for the entries range from 0 (no information) to 1 (a lot of information) (maximum information). Information gain value is higher for attributes that contribute more information, while it is lower for attributes that do not contribute much information.

IV. Results and Discussion

Test results with 5-fold validation can be seen in Table I below.

TABLE I. Table for classification performance, accuracy, and training time for creating respective models using 5-fold cross validation

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,174	3,510	72.33%	0.35
Random Forest	9,665	3,019	76.20%	3.18
Random Tree	8,475	4,209	66.82%	0.05

TABLE II. Confusion Matrix Table for The Entire Decision Tree using 5-Fold Cross Validation

Classifier	a	b	Parametric Variable
J48	3,302	2,172	0
	1,338	5,872	1

Classifier	a	b	Parametric Variable
Random Forest	3,647	1,827	0
	1,192	6,018	1
Random Tree	3,438	2,036	0
	2,173	5,037	1

TABLE III. TABLE FOR CLASSIFICATION PERFORMANCE, ACCURACY, AND TRAINING TIME FOR CREATING RESPECTIVE MODELS USING 7-FOLD CROSS VALIDATION

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,211	3,473	72.62%	0.1
Random Forest	9,718	2,966	76.62%	2.69
Random Tree	8,369	4,315	65.98%	0.04

TABLE IV. CONFUSION MATRIX TABLE FOR THE ENTIRE DECISION TREE USING 7-FOLD CROSS VALIDATION

Classifier	a	b	Parametric Variable
J48	3,301	2,173	0
	1,300	5,910	1
Random Forest	3,655	1,819	0
	1,147	6,063	1
Random Tree	3,409	2,065	0
	2,250	4,960	1

Test results with 7-fold validation can be seen in Table III below. Test results with 10-fold validation can be seen in Table V. Table VI shows the performance of decision tree in this study.

TABLE V. TABLE FOR CLASSIFICATION PERFORMANCE, ACCURACY, AND TRAINING TIME FOR CREATING RESPECTIVE MODELS USING 10-FOLD CROSS VALIDATION

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,233	3,451	72.79%	0.29
Random Forest	9,778	2,906	77.09%	3.1
Random Tree	8,547	4,137	67.38%	0.14

TABLE VI. CONFUSION MATRIX TABLE FOR THE ENTIRE DECISION TREE USING 10-FOLD CROSS VALIDATION

Classifier	a	b	Parametric Variable
J48	3,311	2,163	0
	1,288	5,922	1
Random Forest	3,704	1,770	0
	1,136	6,074	1
Random Tree	3,488	1,986	0
	2,151	5,059	1

TABLE VII. TABLE OF ATTRIBUTES RANKING

Rank	Value	Attributes
1	0.050799	coupon
2	0.015898	CoffeeHouse
3	0.012838	passanger
4	0.01237	destination
5	0.01218	expiration
6	0.01002	time
7	0.007814	weather
8	0.007665	toCoupon_GEQ25min
9	0.006874	occupation
10	0.004818	toCoupon_GEQ15min
11	0.004511	Bar
12	0.003854	Restaurant20To50
13	0.003591	age
14	0.003186	temperature
15	0.002821	income
16	0.002779	maritalStatus
17	0.002477	CarryAway
18	0.00239	education
19	0.001496	has_children
20	0.001395	gender
21	0.001327	RestaurantLessThan20
22	0.000153	direction_same

From Table 7, it can be seen that the test using 10-fold cross-validation has better accuracy than 5-fold and 7-fold cross-validation. Table 5 shows all the results after testing using stratified 10-fold cross validation. It is known that Random Forest has the highest accuracy (77.09%), and Random Tree has the lowest accuracy (67.38%). However, Random Tree has the fastest time for model training (0.14 seconds) while Random Forest has the slowest (3.1 seconds). Based on all the results obtained, it is found that Random Forest has the highest accuracy results, but the training time is far longer. From the attribute ranking table, it shows that coupon is the most influential attribute, with an information gain value of 0.050799.

V. CONCLUSION

This study shows how to classify coupon recommendation data in vehicles using three data mining decision tree classification algorithms: J48, Random Forest, and Random Tree. The classification was done using stratified 10-fold cross-validation and based on the given data set. The results reveal that the Random Tree classification algorithm takes the least amount of time (0.14 seconds) and has the lowest accuracy of the three algorithms (67.38%). The J48 algorithm is more accurate than the Random Tree algorithm (72.79%), but the time it takes is significantly longer (0.29 seconds). The Random Forest technique has the best accuracy (77.09%), but the time it takes to develop the classification model is substantially longer than the Random Tree and J48 algorithms (3.1 seconds). Thus, if less time is required, the Random Tree method can be used, and if great accuracy is required, the Random Forest approach can be employed. From the results, it can also be concluded that coupon is the most influential attribute with an information gain value of 0.050799 so that the type of coupon is the most considered by customers in choosing whether they will accept the coupon or not.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] S. Barat and L. Ye, "Effects of Coupons on Consumer Purchase Behavior: A Meta-Analysis," Journal of Marketing Development and Competitiveness, vol. 6, no. 5, pp. 131-145, 2012.

[2] S. R. Kalmegh, "Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree, and LADTree for Classification of Indigenous News Data," International Journal of Emerging Technology and Advanced Engineering, vol. 5, no. 1, 2015.

[3] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[4] S. Kiranmai and A. J. Laxmi, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy," Protection and Control of Modern Power Systems, vol. 3, no. 29, pp. 1-12, 2018.

[5] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, K. E and P. MacNeille, "A bayesian framework for learning rule sets for interpretable classification," Journal of Machine Learning Research, vol. 18, pp. 1-37, 2017.

[6] U. Bashir and M. Chachoo, "Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System," International Journal of Network Security & Its Applications (IJNSA), vol. 9, no. 4, 2017.

[7] N. Saravanan and V. Gayathri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)," International Journal of Computational Intelligence and Informatics, vol. 7, no. 4, 2018.

[8] M. Schounlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal: Promoting communications on statistics and Stata, vol. 20, no. 1, 2020.

[9] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019

[10] W. Gata, G. Y. E. Patras, R. Hidayat, R. Fatmasari, S. Tohari, B. and N. K. Wardhani, "Prediction of Teachers' Lateness Factors Coming to School Using C4.5, Random Tree, Random Forest Algorithm," in 2nd International Conference on Research of Educational Administration and Management (ICREAM 2018), Bandung, 2019.

[11] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest data mining algoritms for microarray data analysis," International Journal on Advanced Electrical and Computer Engineering (IJAECE), vol. 3, no. 4, 2016.

Comparison of Adolescent Vaccination Data Accuracy by Urban Village in DKI Jakarta Province in July 2021 Using Several Data Mining Methods

Nadine Annisa Heartman
Research Center for Artificial Intelligence and Big Data
Computer Science Department
Universitas Padjadjaran
Sumedang, Indonesia
nadine19002@mail.unpad.ac.id

Yuela Thahira
Computer Science Department
Universitas Padjadjaran
Sumedang, Indonesia
yuela19001@mail.unpad.ac.id

Ruth Rebecca Ovelin
Computer Science Department
Universitas Padjadjaran
Sumedang, Indonesia
ruth19004@mail.unpad.ac.id

Afrida Helen
Research Center for Artificial Intelligence and Big Data
Computer Science Department
Universitas Padjadjaran
helen@unpad.ac.id

***Abstract* — Since 2020, the outbreak of the Coronavirus disease has begun to enter the territory of Indonesia. For a year and a half, various efforts have been made to reduce the number of deaths caused by this pandemic. One of the efforts made by the government is the provision of vaccinations for the community, especially for adolescents. This is one way to attract people's interest to vaccinate and also make it easier for the government and the system to process vaccination data, especially for youth vaccination. The purpose of this study is to determine the accuracy of the data on adolescents who have been vaccinated in the DKI Jakarta province in July 2021 by using several methods of data mining. Of the three data mining methods used in this study, the JRip method produces the highest percentage of accuracy, which is 100%.**

Keywords* — *Comparison, vaccination, data mining, ZeroR, Naive Bayes Classifier, JRip.

I. INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Most people infected with this virus will experience mild to moderate respiratory problems and recover without requiring special treatment. However, some will become seriously ill and require medical attention [1]. Older people and people with underlying medical conditions such as cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop these diseases into more serious illnesses. Anyone can fall ill from contracting the disease COVID-19 and become seriously ill or die at any age. The virus can spread through fluid particles in an infected person's mouth or nose in tiny fluid particles when they cough, sneeze, talk, or even breathe. These particles range from larger respiratory droplets to smaller aerosols. So please practice breathing etiquette, such as coughing with bent elbows, and stay home and self-isolate until you recover if you feel unwell.

According to news reported by Detik.com, in Indonesia itself, news emerged of people with COVID-19 disease on Monday, March 2, 2020 [8]. At that time, President Joko Widodo announced that there were 2 Indonesians infected with the Coronavirus, namely a 31-year-old woman and a 31-year-old woman and a 64-year-old mother. It is suspected that the case began with a meeting of a 31-year-old woman with a Japanese citizen in Indonesian territory in February 2020.

Since it was first announced that the Coronavirus disease was in Indonesia, its spread has become more widespread. In January 2021, new positive cases of Corona in a day could reach around 12,000 cases every day. This figure briefly decreased in April but then rose significantly again in July 2021, with new positive cases of Corona per day reaching 55,000 cases. In addition to positive cases, there is also data on deaths that increase to 1,600 deaths per day [9]. Because of this detrimental impact on society, preventive action is needed from both the government and the community.

The best way to prevent and slow transmission is to be well informed about the disease and how the virus spreads. Protect yourself and others from infection by maintaining a distance of at least 1 meter from others, wearing a well-fitting mask, and frequently washing hands or using an alcohol-based rub. In addition, people aged 12-60 years can also get vaccinated according to the shifts that have been given by the local government [3]. By holding vaccinations and implementing stricter health protocols through the PPKM (*Pemberlakuan*

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

Pembatasan Kegiatan Masyarakat or Enforcement of Community Activity Restrictions), the number of positive cases and deaths in Indonesia caused by Corona, can be suppressed quickly. Now, new positive cases of Corona persist at around 1,900 new cases per day and 117 deaths per day.

With the aim of reducing the number of new positive cases of Corona and the death rate by seeing a decrease in the impact of Coronavirus after the administration of a vaccine by the government, researchers conducted this study to prove that the decline in cases was influenced by the number of vaccinations that had been given, especially at the age of teenagers in the Jakarta area. In addition, through this study, the researcher also compared the accuracy of the data from the three methods used so that it was found which method was the most suitable for processing the dataset in this case. This study uses a dataset from the Jakarta Open Data DKI Jakarta Province. The data will be processed using three methods, namely the ZeroR method, the Naive Bayes Classifier method, and the JRip Classifier method.

II. Literature Study

In the process of this research, three methods of data mining were used to make comparisons in this study, namely the comparison of adolescent vaccinations by urban villages in the province of DKI Jakarta in July 2021. The classifier methods used are ZeroR, Naive Bayes Classifier, and JRip. The reason for using these methods is because they are the simplest classification methods that rely on targets and all predictors. The methods are useful for determining baseline performance as a benchmark for other classifications. The classifications are based on the most frequent class predictions. In addition, these methods are also fast and efficient in dealing with missing values with approximate probability estimates. Further explanation of the method, as follows :

A. ZeroR Method

ZeroR is the simplest classification method that relies on the target and ignores all predictors. The classification of the ZeroR method only predicts the majority category (class). The ZeroR method is useful for determining baseline performance as a benchmark for other classification methods, although this method lacks predictability power. ZeroR classifier method always classifies to the largest class - the largest [4].

B. Naive Bayes Classifier Method

Naive Bayes Classifier is a classification method rooted in Bayes' theorem. This method uses statistical methods to calculate the probability of a class having a group of attributes and determine which class is the most optimal. The main characteristic of the Naive Bayes Classifier is a very strong (nave) assumption of the independence of each condition or event [2]. The basis of the Naive Bayes Classifier theorem is the Bayes formula, if there are two separate events (e.g. A and B), then the Bayes theorem is formulated as follows in Eq. 1.

$$P\ (A|B) = \frac{P(A)\ .P\ (B|A)}{P(A)} \quad (1)$$

C. JRip Classifier Method

JRip Classifier is a learning implementation of the propositional rule, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP. The algorithm consists of two stages, namely, the Building stage with sub-points Grow phase, Prune phase, and the Optimization stage) [6].

III. Methodology

In this research process, several comparisons of methods were carried out to classify adolescent vaccination data by urban village in DKI Jakarta province in July 2021 using data mining. The methodological stages used in this study can be described with a flowchart as shown in Figure 1.

As can be seen in Figure 1, the first thing that must be done to conduct this research is to collect a dataset on adolescent vaccination by urban villages in the DKI Jakarta province in July 2021. After the dataset is collected, then do preprocessing with two types, namely preprocessing using data which still have missing values and also preprocessing using data whose missing values have been removed. After preprocessing, the thing that must be done is to classify the data using methods in data mining. Do a trial using the selected method, here we use 3 methods, namely the ZeroR method, the Naive Bayes Classifier method, and the JRip method [5]. After the results of the trial come out, then do a comparison of the methods used.

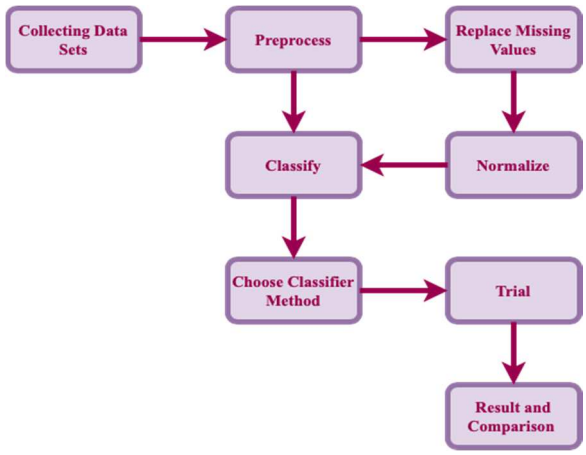


Fig. 1. Research Methodology Flow

A. Collecting Dataset

The first thing to do before preprocessing and classifying is to collect the dataset that will be used in this study. The dataset that will be used in this research is taken from the Open Data Jakarta website which presents data from all Units and Work Units in the DKI Jakarta Provincial Government. The dataset was taken on September 30, 2021. The initial attributes in the dataset are 8 attributes, namely *tanggal, kode_kelurahan,*

wilayah_kota, kecamatan, kelurahan, penerima_vaksinasi, dosis, and *jumlah_vaksin* (date, village_code, city_region, subdistrict, *kelurahan*, recipient_vaccination, dose, and number of vaccines). In addition, this dataset has 14,418 data.

B. Preprocess

In preprocessing, the youth vaccination dataset by kelurahan in the DKI Jakarta province in July 2021 has missing values. Prior to preprocessing, data that has missing values are classified to see a comparison of the percentage of data accuracy before and after the removal of missing values. In preprocessing the data, there are two stages of the process, namely the removal of missing values and normalizing the data. With missing values in the data, it will interfere with the classification process and some data mining methods cannot process [10]. Therefore, it is necessary to eliminate missing values in the Adolescent Vaccination Amount dataset by Urban Village in DKI Jakarta Province in July 2021. By carrying out this process, the existing data can be referred to as clean data. After that, the second process is carried out, namely Normalize, by normalizing all numeric values in the given data set to the default range [0.0, 1.0].

C. Classifier Method Selection on Classify

In analyzing the performance of several classification techniques, a comparison was made on three methods in data mining, namely the ZeroR method, the Naive Bayes Classifier method, and the JRip Classifier method to select the best method for classifying datasets from the Adolescent Vaccination Amount Data by Urban Village in DKI Jakarta Province Month July 2021. Using the existing classification methods in data mining in the form of ZeroR, Naive Bayes Classifier, and JRip Classifier.

TABLE I. DATA WITH MISSING VALUES

No.	Method	Correctly Classified Instances	Incorrectly Classified Instances
1	ZeroR 10%	2616	2628
2	ZeroR 50%	1451	1472
3	ZeroR 90%	299	311
4	Naive Bayes Classifier 10%	5212	32
5	Naive Bayes Classifier 50%	2871	52
6	Naive Bayes Classifier 90%	609	1
7	JRip Classifier 10 %	5231	13
8	JRip Classifier 50 %	2916	7
9	JRip Classifier 90 %	608	2

TABLE II. DATA WITHOUT MISSING VALUES

No	Method	Correctly Classified Instances	Incorrectly Classified Instances
1	ZeroR 10%	10360	2616
2	ZeroR 50%	5758	1451
3	ZeroR 90%	1143	299
4	Naive Bayes Classifier 10%	12939	37
5	Naive Bayes Classifier 50%	7201	8
6	Naive Bayes Classifier 90%	1441	1
7	JRip Classifier 10 %	12958	18
8	JRip Classifier 50 %	7208	1
9	JRip Classifier 90 %	1442	0

D. Trial

After the data set is extracted, to perform the classification algorithm, a testing phase is carried out with different percentage splits. The data to be tested is dose (nom) totaling 14,418 data. From the data set, it was found that the amount of correctly classified instances and incorrectly classified instances differs from each method. Table 1-4 are the result for each method. From the data, the value of correctly classified instances and incorrectly classified instances from each method can also be defined in percentage form as shown below:

TABLE III. DATA WITH MISSING VALUES

No.	Method	Correctly Classified Instances %	Incorrectly Classified Instances %
1	ZeroR 10%	49.885 %	50.114 %
2	ZeroR 50%	49.640 %	50.359 %
3	ZeroR 90%	49.016 %	50.983 %
4	Naive Bayes Classifier 10%	99.389 %	0.610 %
5	Naive Bayes Classifier 50%	98.221 %	1.779 %
6	Naive Bayes Classifier 90%	99.836 %	0.163 %
7	JRip Classifier 10 %	99.752 %	0.247%
8	JRip Classifier 50 %	99. 760 %	0.239 %
9	JRip Classifier 90 %	99.672 %	0.327%

TABLE IV. DATA WITHOUT MISSING VALUES

No	Method	Correctly Classified Instances %	Incorrectly Classified Instances %
1	ZeroR 10%	79.840 %	20.160 %
2	ZeroR 50%	79.872 %	20.127 %
3	ZeroR 90%	79.264 %	20.735 %
4	Naive Bayes Classifier 10%	99.714 %	0.2851 %
5	Naive Bayes Classifier 50%	99.889 %	0.111 %
6	Naive Bayes Classifier 90%	99.930 %	0.069 %
7	JRip Classifier 10 %	99.861 %	0.138%
8	JRip Classifier 50 %	99.986 %	0.013%
9	JRip Classifier 90 %	100,000%	0%

IV. RESULTS AND ANALYSIS

This section represents the results obtained after testing the three classification methods in data mining with different percentage splits. The test results of the classification methods are shown by using the scoring method. The scoring method used to calculate the accuracy of this classification is by looking at the percentage of correctly classified instances [7].

TABLE V. COMPARISON OF DATA ACCURACY BEFORE REMOVING MISSING VALUES

Method	10%	50%	90%	% with Highest Accuracy
ZeroR	49.88558352 %	49.64078002 %	49.01639344 %	10
Naive Bayes Classifier	99.38977879 %	98.22100582 %	99.83606557 %	90
JRip	99.75209764 %	99.76052001 %	99.67213115 %	50

It can be seen in Table 5, before removing the missing values on the data the percentage value of the accuracy of the ZeroR method was very small overall. In the ZeroR method, the highest percentage of accuracy is 49.885 % with a percentage split of 10%. For the Naive Bayes Classifier method, the percentage of scattered accuracy is 99.836 % with a percentage split of 90% and for the JRip method, the percentage of scattered accuracy is 99.760 % with a percentage split of 50%. It can be seen that the Naive Bayes Classifier method and the JRip method have very high accuracy values compared to the ZeroR method. Thus, it was found that the highest accuracy value before removing missing values using the three methods was 99.836 % in the Naive Bayes Classifier method with a percentage split of 90%.

TABLE VI. COMPARISON OF DATA ACCURACY AFTER REMOVING MISSING VALUES

Method	10%	50%	90%	% with Highest Accuracy
ZeroR	79.839 %	79.872 %	79.264 %	50
Naive Bayes Classifier	99.714 %	99.889 %	99.930 %	90
JRip	99.861 %	99.986 %	100, 000%	90

It can be seen in Table 6, after removing the missing values in the data, the percentage value of the accuracy of the ZeroR method remains the smallest percentage. In the ZeroR method, the highest percentage of accuracy is 79.872 % with a 50% split percentage. For the Naive Bayes Classifier method, the percentage of scattered accuracy is 99.930% with a percentage split of 90% and the JRip method has a percentage spread of 100% accuracy with a percentage split of 100%. It can be seen that the JRip method has a very high accuracy value compared to the ZeroR and Naive Bayes Classifier methods. This is different from before, in data that has not been removed for missing values, the Naive Bayes Classifier method is the most accurate method. Then it was found that the highest accuracy value after eliminating missing values using the three methods was 100% in the JRip method with a percentage split of 90%.

The classification results are compared by looking at the numbers on Correctly Accuracy (CA), Precision, Recall and F-Measures. The scoring method is used to calculate the classification accuracy of the data using Correctly Classified Instances & Incorrectly Classified Instances. CA, Precision, Recall and F-Measure are the scoring methods used for this test. CA is used to calculate subset accuracy, precision is used for intuitive classification accuracy, recall is used to measure the classification ratio, F-Measure is used to find the harmonic mean (average value) of Precision and Recall, with the best value being 1 and the worst being 0. Table 7-9 shows the result of scoring from each data with missing values, and Table10-12 without missing values.

TABLE VII. PERCENTAGE SPLIT OF 10%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.499	0.499	0.499	0.666
Naive Bayes Classifier	0.994	0.994	0.994	0.994
JRip	0.998	0.998	0.998	0.998

TABLE VIII. PERCENTAGE SPLIT OF 50%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.496	0.496	0.496	0.663
Naive Bayes Classifier	0.982	0.983	0.982	0.982
JRip	0.998	0.998	0.998	0.998

TABLE IX. PERCENTAGE SPLIT OF 90%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.490	0.490	0.490	0.658
Naive Bayes Classifier	0.998	0.998	0.998	0.998
JRip	0,997	0.997	0.997	0.997

TABLE X. PERCENTAGE SPLIT OF 10%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.798	0.798	1.000	0.888
Naive Bayes Classifier	0.997	0.997	0.997	0.997
JRip	0.999	0.999	0.999	0.999

TABLE XI. PERCENTAGE SPLIT OF 50%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.799	0.799	1.000	0.888
Naive Bayes Classifier	0.999	0.999	0.999	0.999
JRip	1.000	1.000	1.000	1.000

TABLE XII. PERCENTAGE SPLIT OF 90%

Method	Accuracy	Precision	*Recall*	*F-Measure*
ZeroR	0.793	0.793	1.000	0.884
Naive Bayes Classifier	0.999	0.999	0.999	0.999
JRip	1.000	1.000	1.000	1.000

The results of the scoring on the data before removing missing values show the best accuracy values produced by the JRip method and the Naive Bayes method but the best accuracy is dominated by the JRIP method with Accuracy value of 0.998, Precision of 0.998, Recall of 0.998 and F-Measure of 0.998. The same accuracy value is also obtained by the Naive Bayes method at a percentage split of 90% with a Precision, Recall and F-Measure value of 0,998. Meanwhile, the scoring results on the data after removing missing values show the best accuracy value produced by the JRip method with Accuracy value of 1,000, a precision of 1,000, a recall of 1,000 and an F-Measure of 1,000 for all percentage splits.

V. CONCLUSION

This paper has analyzed a dataset using three types of test methods in data mining that show different results and percentages. Based on the research above, it was found that the classification method with data that still has missing values is using Naive Bayes with an accuracy rate of about 99.84% for the most accurate results and ZeroR with an accuracy rate of about 49.89% for the least accurate results. While the classification method with data that has no missing value is JRip with an accuracy rate of 100% for the most accurate results and ZeroR with an accuracy rate of 79.87% for the least accurate results. This study is useful for the government and medical professionals, especially for the urban area of DKI Jakarta, to collect and analyze data by estimating the predicted rate and percentage of adolescents who have been vaccinated in the area. In addition, this study also aims to develop strategies and take appropriate preventive measures to reduce the spread of Covid-19 cases in the region.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[2] S .Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models", Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 787–795, 2013. https://doi.org/10.2478/amcs-2013-0059

[3] S. Balamurugan, M. Ashwini, and P. Waranyoo, "Emergence of novel coronavirus 2019-ncov: Need for Rapid Vaccine and Biologics Development",2020;9(2):148. https://doi.org/10.3390/pathogens9020148

[4] K. Ahmed and T. Jesmin, "Comparative analysis of data mining classification algorithms in type-2 diabetes prediction data using weka approach", Internat. J. Sci. Eng., vol. 7(2), pp.155-160, October 2014. https://doi.org/10.12777/ijse.7.2.155-160

[5] L .J. Muhammad, M. M. Islam, S.S. Usman et al, " predictive data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery", SN COMPUT. SCI. 1, 206 (2020). https://doi.org/10.1007/s42979-020-00216-w

[6] I. M. Tarun, D. G. Bobby, and B. T. Tanguilig III, "Generating Licensure Examination Performance Models Using PART and JRip Classifiers: A Data Mining Application in Education", Int. J. Computer and Communication Engineering, vol. 3, no. 3, May 2014.

[7] I. E. Putri, D. Rahmawati, and Y. Azhar, " Comparison of data mining classification methods to detect heart disease ", Jurnal PILAR Nusa Mandiri, vol. 16, no. 2, September 2020. https://doi.org/10.33480/pilar.v16i2.1388

[8] T. P. Latchoumi, L. Jayakumar, T. P Ezhilarasi, L. Parthiban and K. Mahalakshmi "Comparison of classification techniques on data mining", International Journal of Pure and Applied Mathematics, vol. 118, no.1, pp. 357-264, 2018. https://doi.org/10.12732/ijpam.v118i11.43

[9] T. P. Velavan, and C. G. Meyer, The COVID-19 epidemic, Tropical medicine & international health, 25(3), 278, 2020

[10] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019 https://doi.org/0.25126/jtiik20185958

Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method

Paquita Putri Ramadhani
Research Center for AI dan Big Data
Department of Computer Science,
Universitas Padjadjaran
Sumedang, Indonesia
paquita16001@mail.unpad.ac.id

Setiawan Hadi
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
setiawanhadi@unpad.ac.id

Rudi Rosadi
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
r.rosadi@unpad.ac.id

***Abstract*— Competition in the business world is getting tighter, making companies will continue to look for ways so that the company can continue to compete and also have profits that continue to rise every year. Various ways will be carried out by the company to continue to be able to maintain the best performance in the value of business processes for the company. One way for companies to be able to continuously monitor the company's performance is to integrate data owned by the company. The use of Data Warehouse technology will simplify and speed up the data integration process in a company and the Business Intelligence Dashboard will also assist companies in monitoring company performance in the form of data visualization.**

Therefore, this research will design a Data Warehouse system and build a business intelligence dashboard using Superstore Europe data which aims to monitor sales performance and also the delivery of goods using the PostgreSQL database. In this study using the Kimball Lifecycle method.

Keywords—Data Warehouse, Kimball Lifecycle Method, PostgreSQL Database, Business intelligence

I. INTRODUCTION

The usage and growth of information technology can no longer be ignored, and it is happening at a faster rate than before, resulting in an ever-growing data. The data created is not just a collection of unused facts, merely if we process the data properly, the data will have meaning, value, and several benefits that can be used as a source of information and a part of an analysis requirement. This is possible by using Data Warehouse technology.

The usage of Data Warehouse technology in a company or superstore will simplify or speed up the integration process, particularly with the data-level integration. A company or superstore department can have its own business procedures, which are then integrated into one or more apps. Data Warehouse's ability to present reports with a much more quickness and flexibility is a reason why companies or superstores are starting to use Data Warehouses.

The use of business intelligence in an organization, particularly in a firm, can result in an increase in revenue for the company. Because the company can process data properly according to the business needs of the company or superstore, the company can find out which steps, strategies, and decisions that must be taken to increase the company's revenue.

This paper presents a design of Data Warehouse technology which will later produce a report in the form of a business intelligence dashboard containing information about business needs at a superstore in Europe. The implementation of this Data Warehouse will use the Kimball Lifecycle method and the PostgreSQL database.

II. LITERATURE REVIEW

A. Data Warehouse

Data Warehouse is a term that refers to a data repository that is handled independently from the operational database. The Data Warehouse system allows multiple application systems to be integrated. They support the information processing by providing a strong foundation of historical data consolidation for analysis purposes. [1].

1) *Characteristics of Data Warehouse*

a) *Subject Oriented*

Data warehouse is organized around key subjects such as customers, suppliers, products, and sales. A Data Warehouse, rather than focussing on an organization's day-to-day operations and transaction processing, concentrates on modeling and analyzing data for decision makers [2].

b) *Integrated*

The term "integrated" means "combined into a single whole.". In this example, the integrated characteristic specifies that the data warehouse was created by combining multiple sources, including relational databases, flat files, and online transaction record. Data cleaning and the technique of data integration are done to ensure the consistency of naming convention, code structure, attribute size, etc.

c) *Non volatile*

Non-volatile means that the data that enters the Data Warehouse, must not be affected by data manipulation in the form of edit, update and delete. The data within the Data Warehouse cannot be changed even slightly. While

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

the database, also known as OLTP, allows for data manipulation in the form of insert, edit, update, and delete operations.

d) *Time variant*

In general, a company's operating system may or may not have a time element in its data, but with Data Warehouse, the data is stored to provide information from a historical perspective (eg. the last 5-10 years) by adding a historical dimension to the data obtained from operational databases or OLTP.

2) *Data Warehouse Architecture*

Architecture of data warehouse is constructed by three perts. These elements are linked together, namely Load Manager, Warehousing Manager, Data Access Manager (Query Manager) [3].

3) *Extract, Transform, Load* (ETL)

ETL is a process that must be passed in the shaping of a Data Warehouse which aims to collect, filter, process and combine relevant data from various sources to be stored in the Data Warehouse. There are 3 ETL processes, namely extract, transform, and load.

4) *Dimensional Modelling*

Dimensional modeling is a technique used to present analytical data to deliver a data that business users can understand. Dimensional modeling has 2 main components, namely fact tables and dimension tables.

5) Data Warehouse Schema

Data Warehouse has several schemas in its design. The schemas in the Data Warehouse are as follows:

a) *Stars Schema*

Stars schema is the simplest Data Warehouse schema. Stars schema can be easily identified by the location of the fact table and dimension table. If the former is in the middle and the former is at the star nodes.

b) *Snowflake Schema*

The main difference between star schema and snowflake schema is that snowflake schema consists of normalized dimensions whereas star schema consists of non-normalized dimensions.

c) *Fact Constellation Schema (Galaxy Schema)*

The main difference between star/snowflake schema and galaxy schema is that star/snowflake schema consists of only one fact table whereas galaxy schema always consists of multiple fact tables. In a galaxy schema, dimension tables can also be used by other fact tables.

6) *Online Analytical Processing (OLAP)*

Online Analytical Processing or abbreviated OLAP is a method to extract useful information so as to provide answers to the analysis process requests [3]. OLAP also collects information from various systems and provides a display of information/summary to management, while data mining is used to find hidden patterns in data.

B. Business Intelligence

Business Intelligence (BI) describes a concept and method on how to improve the quality of business decision making based on data-based systems [4]. BI is often likened to briefing books, report and query tools, and executive information systems since by using business intelligence, a company can process and visualize the data more regularly with the help of tools, making it easier for companies to find data which may be of use to the company.

C. Kimball Lifecycle Method

Kimball Lifecycle provides an overall framework that cumulates various Data Warehouse or business intelligence implementation activities [5]. Kimball Lifecycle has several stages, these stages are:

1) Program/project planning.
2) Program/project management.
3) Business requirement definition.
4) Technology Track
5) Data track.
6) Business intelligence application track.
7) Deployment, Maintenance, and Growth

D. PostgreSQL Database

PostgreSQL is a free and open-source Database Management System (DBMS). In addition, PostgreSQL is an object-relational (ORDBMS-object-relational-DBMS). PostgreSQL emphasizes extensibility, creativity, and compatibility. In a contender with large relational database vendors such as Oracle, MySQL, SQL Server, and others [6].

III. RESEARCH METHOD

A. Data Warehouse Design and Development

In designing and building a Data Warehouse, the author uses the Kimball Lifecycle method. In the process of developing a Data Warehouse using the Kimball Lifecycle method, several processes are interrelated, namely program/project planning, program/project management, business requirements definition, technology track, data track, business intelligence application track, deployment and maintenance.

1) *Project planning.*

At this stage, identification and planning are carried out for the need for data warehouse development. There are 3 stages, namely the identification of the scope, objectives, and benefits as follows:

a) *Environmental Identification*

Information on what business process needs may be derived from this Superstore data during the scope identification phase. There are 2 areas of business processes that can be taken, namely in the sales and logistics department. The sales department can support reports on the value of profits and also the value of sales in the last 4 years. Meanwhile, in the logistics section, the objectives to be taken are as supporting reports to monitor the value of on time in full or on time in full (OTIF) as an indicator of logistics performance

b) *Propose*

This Data Warehouse development aims to implement the Data Warehouse system in the data superstore with the goal of supporting the report, namely monitoring the profit value as well as the revenue value for four years in the sales process, and monitoring the value of on-time, in full and the value on time in full (OTIF) as an indicator of logistics performance which measures the efficiency and accuracy of delivery in the logistics department. In addition, it can be an evaluation tool in supply chain management or called supply chain management,

especially in sales and logistics by providing useful information for the analysis and decision-making process. Advantage

The advantage of establishing a Data Warehouse and business intelligence system is that it may assist monitoring in the sales and logistics departments, which will be utilized as a decision-making reference.

2) *Project management.*

At this stage, monitoring and tracking of problems is carried out during the construction of the Data Warehouse and business intelligence dashboard. This is done to ensure that Kimball Lifecycle activities stay on track and in sync.

3) *Business requirement definition.*

At this stage, an identification process is carried out that is useful in the business intelligence development process. And the results of the identification process obtained are business processes in the sales department regarding the sales process (revenue and profit) and the logistics department regarding the delivery process (key performance indicators on time in full).

Key Performance Indicator on Time in Full (OTIF KPI) is a measurement of logistics or delivery performance in a supply chain.

In general, the OTIF KPI is calculated by taking into account the number of shipments that meet the delivered on time and delivered in full values. The following is the formula for calculating the OTIF KPI:

$$\text{OTIF}(\%) = \frac{OTIF\ Delivery\ Amount}{Total\ Delivery\ Amount} \times 100 \qquad (1)$$

Total revenue is the calculation of the amount of income earned by the company by selling its goods and or services over a certain period of time, for example, per month or per year. This total revenue is obtained based on the calculation of the production level multiplied by the price level. The following is the formula for calculating total revenue:

$$\text{Total Revenue} = \text{Item Price} \times \text{Quantity} \qquad (2)$$

Profit is the net profit obtained after subtracting the total revenue from the total cost. Total costs include the costs of making goods, products, and services offered by the company. The following is the profit calculation formula:

$$\text{Profit} = \text{Total } Revenue - \text{Total Cost Profit} \qquad (3)$$

The data is taken from data.world, namely Superstore Europe data. The data has 10,000 rows and 20 columns. The columns in this data are order_id, order_date, ship_date, ship_mode, customer_id, customer_name, segment, city, state, country, region, product_id, category, sub_category, product_name, product_type, sales, quantity, profit, arrived_quantity.

4) *Technology Track*

a) *Technical Architecture Design*

In this technical architecture design, we will create a technical framework for developing Data Warehouse. Looking at the various information needs and objectives of the research, the overall performance framework has been determined.

b) *Product Selection & Installation*

1. Software Specification

The software used in accordance with the needs of data warehouse development are:

a. Windows 10
b. Google Chrome.
c. Visual Studio Code
d. PostgreSQL *Database*

2. Hardware Specification

The computer hardware required for creating this web-based dashboard are listed below:

a. Processor: Intel Core i7-8565U 1.80GHz
b. Memory: 16 GB DDR4
c. Graphic: NVIDIA MX230 2GB
d. Operation System: Windows 10
e. Hard Drive: 1TB SSD.

5) *Data track*

At the data track stage, there are 3 activities carried out, first, dimensional modeling is the process of forming dimensions and facts as well as the Data Warehouse schema. Second, physical design is the process of forming metadata in the database, by detailing all the attributes of each dimension and facts that have been designed in the previous stage. And the third is ETL Design & Development, which is the process of extracting, transforming and loading from the source to the target environment.

a) *Dimensional Modelling*

At this stage, dimensional modeling is carried out by referring to the four-step Kimball method. Here are the steps in Kimball's four-step method:

1. *Choose the Business Process*
   Choosing the business process means determining the main subject. The business processes used in designing this Data Warehouse are business processes in sales and logistics in superstores.
2. *Declare the Grain*
   Grain or granularity is information that will represent data from a fact table. Choosing a grain means determining what a fact table will represent or represent.
3. *Identify the Dimensions*
   Identify and relate dimension tables to fact tables. 12 dimension tables are formed, namely kategori, subkategori, region, negara, provinsi, kota, produk, pesanan, segmen, mode_pengiriman, waktu, pelanggan.
4. *Identify the Facts*
   This process determines the fact table to be formed. 2 sales and logistic fact tables are formed.

Fig. 1 presents the relationships between tables of data warehouse superstore.

b) *Physical Design*

After carrying out the dimensional modeling stage, the physical design stage is carried out, namely the process of forming metadata in the Data Warehouse database, by detailing all the attributes of each dimension and facts that have been designed in the previous stage.

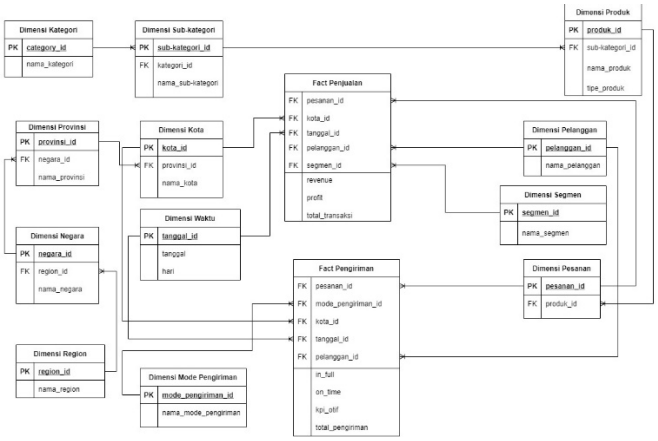


Fig. 1. Relationships Between Tables of Data Warehouse Superstore

c) *ETL Design and Development*

ETL is a process that is generally carried out to move data from one data source to another, including manipulation of the data. The series of ETL carried out are data extraction, data transformation, and data loading.

6) *Business intelligence application track.*

a) *Business Intelligence Application Design*

At this stage, the design of the appearance of the application that will be made is mainly the dashboard interface. At this design stage, the selection of applications or platforms that will be used to support business needs is carried out and also displays the business value of the solutions provided by business intelligence.

b) *Business Intelligence Application Development*

This stage contains the process of developing BI applications from the designs that have been carried out in the previous stage.

7) *Deployment, Maintenance, and Growth*

Deployment phase, what is done in this phase is testing the entire process, starting from the technology track, data track and business intelligent track. In launching the application in this study, it was carried out using heroku hosting. In the maintenance phase, after the business intelligence system is produced, an operational technique is needed to keep the system working optimally. In the growth phase, it occurs when the data increases, or the need for analysis increases. This has resulted in the Data Warehouse development project having to be restarted from scratch, be it a large-scale development or the addition of several new schemes.

IV. Results and Discussion

The implementation of this interface is the result of creating a web-based business intelligence application interface that aims to display the contents of the Data Warehouse. Making this application using the full-stack django framework with the python programming language.

a) *Extraction*

Extract is the process of selecting and retrieving data from one or more sources such as flat files or RDBMS. This stage aims to identify the right pieces of data, which will be entered into the flowchart of the ETL for further processing. In this process the data source comes from the PostgreSQL database.

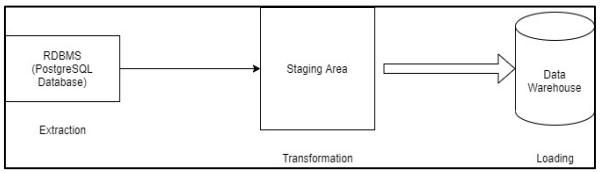
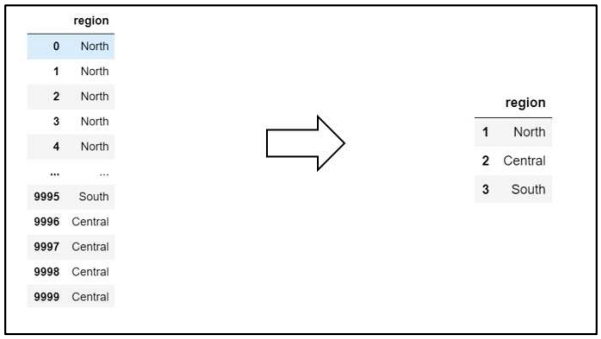


Fig. 2. ETL Schema

b) *Transformation*

In this process the data that has been taken in the extraction process will be cleaned and change the data from its original form into a form that suits the needs of the Data Warehouse.

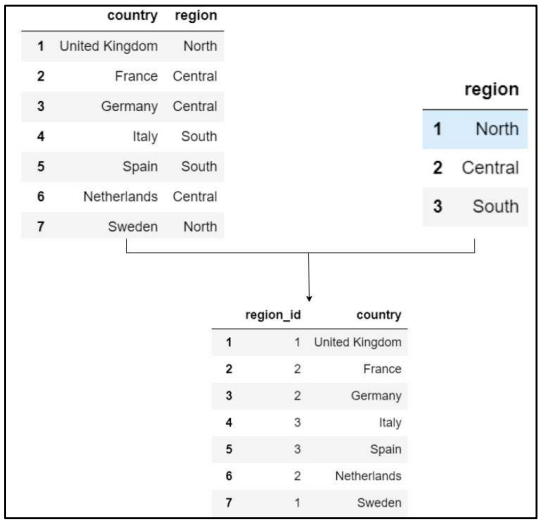
At this stage there are 2 processes carried out, namely removing duplicates and lookup. The process of deleting duplicate data is carried out on the dimension table except for the time dimension table. An example of the process of removing duplicates at the transformation stage can be seen in Fig. 16. The lookup process is to match the foreign key in each column in the table that has foreign key relations with other tables. An example of the lookup process at the transformation stage can be seen in Fig. 17.



	region
0	North
1	North
2	North
3	North
4	North
...	...
9995	South
9996	Central
9997	Central
9998	Central
9999	Central

	region
1	North
2	Central
3	South

Fig. 3. Example of Remove Duplicate



	country	region
1	United Kingdom	North
2	France	Central
3	Germany	Central
4	Italy	South
5	Spain	South
6	Netherlands	Central
7	Sweden	North

	region
1	North
2	Central
3	South

	region_id	country
1	1	United Kingdom
2	2	France
3	2	Germany
4	3	Italy
5	3	Spain
6	2	Netherlands
7	1	Sweden

Fig. 4. Example of *Lookup* Process

c) *Loading*

The last phase is loading, which is used to enter data into the final destination, which is the Data Warehouse. A SQL script can be used to enter data. In this process, the

data will be converted into a dimensional data store so that the data format is suitable to be applied to the analysis process and has been integrated with several data sources.

Before proceeding with the loading process (ETL process), a database schema is created in the Data Warehouse based on the tables that have been designed. The formation of the database schema in this Data Warehouse uses the fact constellation schema (galaxy schema).

V. CONCLUSION

The Kimball Lifecycle approach for building a Data Warehouse system involves multiple interconnected stages, namely program/project planning, program/project management, business requirements definition, technology track, data track, business intelligence application track, deployment maintenance and growth . The result of this design is in the form of a Data Warehouse schema, where in this study the schema used is a fact constellation schema or a galaxy schema by having 2 fact tables and 12 dimension tables.

In this business intelligence dashboard, it will display graphs to monitor sales performance in the Sales department and also the distribution of items in the Logistics department. There are 5 interface pages on the business intelligence dashboard, namely the login page, home page, sales dashboard page, logistics dashboard page, sales or delivery data table page.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques 3rd Edition. Elsevier, 2011.

[2] I. Pratama, Handbook Data Warehouse Teori dan Praktik Berbasiskan Open Source. Bandung: Informatika Bandung, 2018

[3] P. Bhatia, Data Mining and Data Warehousing Principles and Practical Techniques. Cambridge University Press., 2019.

[4] R. Sherman, Business Intelligence Guidebook From Data Integration to Analytics. Morgan Kaufmann, 2015.

[5] R. Kimball,et al., The Data Warehouse Lifecycle Toolkit Second Edition. Wiley, 2008.

[6] S. Juba, and A. Volkov, Learning PostgreSQL 10 - Second Edition. Packt, 2017.

[7] W. S. Vincent, Django for Beginners: Build websites with Python and Django. WelcomeToCode, 2018

[8] D. Rubio, Beginning Django: Web Application Development and Deployment with Python, Apress, 2017.

[9] N. George, Build a Website With Django 3: A complete introduction to Django 3, Build a Website with Django 3, 2019.

[10] B. Shaw, S. Badhwar, A. Bird, B. C. K.S. and C. Guest, Web Development with Django: Learn to build modern web applications with a Python-based framework, Packt Publishing, 2021.

Twitter's Hate Speech Multi-label Classification Using Bidirectional Long Short-term Memory (BiLSTM) Method

Refa Annisatul Ilma[ab1], Setiawan Hadi[b2], Afrida Helen[ab3]
[a]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[b]*Department of Computer Science, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]refaands@gmail.com, [2]setiawan.hadi@unpad.ac.id, [3]helen@unpad.ac.id

Abstract*— Since social media is one of the most likable products of techngetpeople get easier to express their opinions. Anyone be able to tell their opinion freely there. Unfortunately, its convenience has also become a boomerang for us, the easier every opinion conveyed the easier hate speech is expressed. This matter become the dark side of social media. Hate speech face us with a lot of dangers, such as violence, social conflict, even homicide. Therefore, preventing all of those dangers that might be occur because of hate speech is one of the prior things we need to do. This research was done as an attempt to take care of the dangers that could be done by hate speech. The attempt we tried to do is using multi-label text classification to predict hate speech with the Bidirectional Long Short-term Memory (BiLSTM) method. This multi-label text classification labelled every tweet in the dataset crawled from Twitter with 12 labels about hate speech. From this experiment, we obtained the best hyperparameter value that could achieve great performance with 82.31% accuracy, 83.41% precision, 87.28% recall, and 85.30% F1-score.

Keywords—Bidirectional Long Short-term Memory, GloVe, Hate speech, Multi-label Text Classification, Twitter.

I. INTRODUCTION

Nowadays, social media is one of the important aspects in people's life as social creature. People always crave for attention and interaction, makes social media being something people always use. Based on the data stated by HootSuite (https://www.hootsuite.com)—one of the websites that provide social media content management—stated that social media users in Indonesia reached 160 million persons counted per February 2020. That big number proves social media is important part in people's nowadays life, because social media give them many conveniences specially to communicate to the others wherever and whenever they want. Other than that, they also enjoy sharing their opinion about something through social media. Social media let everyone share their opinion comfortably [1].

However, that convenience not only has leverage in people's internet lives. They also have the other side that carried together with it, just like the two sides of a coin. Their freedom of speech for sharing their opinion in social media also lead to a bunch of negative impacts. One of those is the massively spreading of hate speech in social media. There are irresponsible persons that hiding in their anonymous social media and spreading hate speech towards certain group or person. Behind their social media account, they spread various types of hate speech, ranging from mild to severe and potentially bringing the conflict that could endanger the group or individual that targeted by the hate speech.

Because of its potential to lead to discrimination, violence, social conflict, even a homicide of a particular group or individuals, hate speech is considered as a serious crime [2]. The newest Bill of Act on Electronic Information and Transaction or Law no. 19 of 2016 even mentioned that penalty for the violation of spreading hate speech is sentencing. Unfortunately, the Bill of Act on Electronic Information and Transaction is not optimal enough to reduce the violation include hate speech as a result of too many ambiguous articles there. The ambiguous articles made there are no exact standards about which kind of expressions could be considered as hate speech. The standardization for considering hate speech is unclear. This is clearly the problem should be faced to reduce the adverse effects of hate speech [3].

Based on those, this research conducted to overcome it. We built classification algorithm that would be able to identify and classify hate speech. The general purpose of this research is classifying and labelling the hate speech from Indonesian tweet using BiLSTM algorithm. Multi-label classification also proposed so the model will be able to categorize and identify what kind of hate speech it is. The algorithm expected could help to identify, classify, and standardize the classification of hate speech.

II. LITERATURE STUDY

A. Hate Speech

According to Komnas HAM [2], hate speech is all actions, direct or indirect attempts based on hatred towards certain race, ethnic, religion, disability, gender, and sexual orientation so that group/individual with those background will experience discrimination, violence, social conflict, and even a homicide. Those actions could be done through any medium, including social media.

B. Multi-label Text Classification

Multi-label text classification is an approach that designed to completing task by giving every input text one or more category label that determined before. Multi-label classification seen as a bunch of binary classification tasks. Each label that assigned to the input will be classified with binary classification, so the output amount will be the same as the amount of label determined for classify the data.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

Nevertheless, the model has more task than ordinary binary classification. The model must understand the prediction rules and relations among all feature on each label, so that the model could label the data accurately [4], [5].

C. Data Preprocessing

Data preprocessing is the stage carried out to extract the data into a format that suits the processing needs. This process is needed considering that the data in real world is still raw and need processing to turn it into format suiting the needs. The raw data still contain noise, inconsistency, too many features, and have various format that made the model unable to process it.

Data preprocessing will change those disorganized data into structured data that could be accepted and processed by the model. This stage considered as crucial and complex because the result from this stage will be the representation from each data that used in learning process of the built machine learning or deep learning model [6].

1) Data Cleaning

Data cleaning is basic task that should be done to clean the raw data from unneeded elements that could interfere the data processing by the model. Those elements should be cleaned so that the data could be processed and the model would have better performance [7].

2) Case Folding

Case folding is a process converting all letters in the data into lowercase. This process intend to reduce feature redundancy caused by the case sensitivity [8].

3) Text Normalization

Text normalization is a process to transform informal vocabulary into the formal one. The purpose of this preprocessing stage is to reduce feature redundancy in the dataset that caused by different spelling. Indonesian language has so many informal spellings that actually has the same meaning, hence the text normalization should be done to synchronize it and lessen the redundancy. Feature redundancy could lead the deep learning's performance effectivity decreased [9].

4) Stopword Removal

Stopword removal is a process of removing stopword or most common words appear in the dataset from the data so that it could reduce the dimension of the data. Stopword used to not have significant meaning or sentiment that could influence the sentence, it is more likely a preposition, conjunction, or similar things that really often used in sentence. According to that reason, stopword would be better getting removed so it could lessen the feature that should be learned by the model [10].

5) Tokenization

Tokenization is a process to create token by cutting the data's sequence of strings. Tokenization will divide the data into several part, it can be a word, keyword, phrase, symbol, or even a character in string. Each elements that separated from the sequence of strings called token [11].

D. Global Vector (GloVe)

Global Vectors (GloVe) is unsupervised learning algorithm to represent vectors from words based on statistic of global word-word co-occurrences in a corpus combined with word context from the surrounding words. GloVe become a word embedding method that is used to translate a series of words in the form of strings into vectors so that they can be understood by machines [12].

E. Long Short-term Memory (LSTM)

Long Short-term Memory (LSTM) is an algorithm that modified from Recurrent Neural Network (RNN). LSTM solve RNN's problem of vanishing gradient [13]. LSTM replace the self-connected hidden unit with memory block so that block could be used to keep the information from the previous step [14], [15].

A LSTM unit consists of 4 components, that is forget gate, input gate, output gate, and memory cell. Figure 1 shows the architecture of a LSTM unit.

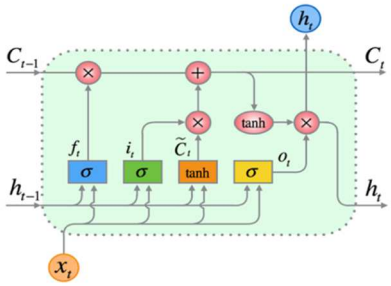


Fig. 1. Figure 1 LSTM Architecture

1) Forget Gate

Forget gate is part of the architecture that will determine the memory of the cell whether it should be forgotten or continued to the next process. This process will select the input for the next process. The following is the equation used to operate the forget gate:

$$f_t = \sigma(W_f \cdot x_t + Uf \cdot h_{t-1} + b_f \quad (1)$$

2) Input Gate

Input gate is the gate that receive the input and from the previous state. Here is the equation used to calculate the value in input gate:

$$i_t = \sigma(W_i \cdot x_t + Ui \cdot h_{t-1} + b_i) \quad (2)$$

3) Output Gate

Output gate is the part of the LSTM architecture that would modulate the memory output and output result for the next stage. The following is the equation used to calculate in output gate:

$$o_t = \sigma(W_o \cdot x_t + Uo \cdot h_{t-1} + b_o) \quad (3)$$

4) Memory cell

Memory cell is a part of the cell that would be activate vector. Memory cell consists of two components, i.e.: memory that have been forgotten (c_{t-1}) in forget gate and new memory that would be continued ($\tilde{c}.t$) to the next state. Here are the equations used to calculate the modulation gate:

$$\tilde{c}_t = tanh(W_c \cdot x_t + W_c \cdot h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (5)$$

F. Bidirectional Long Short-term Memory (BiLSTM)

Bidirectional LSTM, or also known as BiLSTM, are one of modification from unidirectional LSTM algorithm. BiLSTM combine the forward unit and the backward unit LSTM cell for each state, so that the input BiLSTM receive are from past and future of the data. The output from those two LSTM cells then will be combined so that they could be used not only to consider the data context from the forward direction, but also the data context from the backward direction [15]–[19].

G. Evaluation Metrics

Evaluation metrics is metrics that used to measure performance of a model so the model's performance could be analyzed and improved. In the classification model, the commonly used metrics are accuracy, precision, recall, F-measure, and ROC area. The accuracy of the labels classified by the model will be calculated by the scores of these metrics [20], [21].

1) Example-based Metrics

Example-based metrics work by evaluating the performance of each sample test separately and then the results of all calculations for each test sample are combined with all the data in the test set.

a) Accuracy

Accuracy is one of evaluation metrics that used to calculate model's accuracy on predicting the data tests. Unlike the binary or multi-class classification which only categorizes each data as correct and incorrect, the multi-label classification has one more category namely partially correct. Partially correct is used to mark data whose labels are not all identical to the actual label.

The following is the equation to calculate the accuracy rate, with Y as predicted label, Z as actual label and n as the amount of involved label:

$$A = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (6)$$

That equation is used to calculate the evaluation result per record data test, then average the whole data test to get the accuracy of the whole model.

2) Label-based Metrics

Label-based metrics work by evaluating the performance of the model through each label separately and then averaging the evaluation results of each label into the overall evaluation result. The results of this type of calculation not only provide the overall calculation of the dataset, but also provide the evaluation results of each label. This type of metric is calculated based on the results of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) from each label so that Recall, Precision, and F1-scores are obtained from each label.

a) Macro Averaged Measures

Macro averaged measures are formulas that are used to obtain the average results of label-based metrics measurements by calculating the average of the calculation results of each label. This method combines the results that have been calculated previously, so that these metrics will go through two calculation stages: calculating the metrics of each label and calculating the average of the label metrics as the overall calculation result. [22].

b) Micro Averaged Measures

Micro averaged measures are carried out by knowing each number of TP, TN, FN, and FP and then calculating them according to the required metrics formula. In contrast to macro averaged measures, which operate the results of each label operation first and then average them, this calculation only performs operations at the end after the label evaluation results are obtained.

III. Methodology

A. Dataset

The dataset used in this research is fetched from prior similar research topic about hate speech multi-label text classification. The dataset is open for public and could be accessed from the researcher's Github account [9] (https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection). From that source, we also use the dictionary of some informal spelling and the formal one. The dictionary is used for text normalization step in data preprocessing later.

Tweets in that dataset are collected by crawling in Twitter using Tweepy library. The data gathered from Twitter in seven months, starting from 20th March 2018 until 10th September 2018.

Total data in the dataset is 13,169 records and classified into 12 target labels. Those 12 labels are HS, Abusive, HS_Individual, HS_Group, HS_Religion, HS_Race, HS_Physical, HS_Gender, HS_Other, HS_Weak, KS_Moderate, and HS_Strong. Table 1 describe all the target labels. These labels obtained from the Focus Group Discussion (FGD) between the researcher of prior research and the staff of Direktorat Tindak Pidana Siber Badan Reserse Kriminal Kepolisian Republik Indonesia (Bareskrim Polri) as the responsible party for investigating cybercrimes in Indonesia.

From 13,169 data taken, Figure 2 shows the number of tweets labeled '1' for each label. These data, of course, do not always have only one label. One data record can have more than one label because the data it has is in the form of multi-label classification data. A record can have zeros, ones, or even multiple labels.

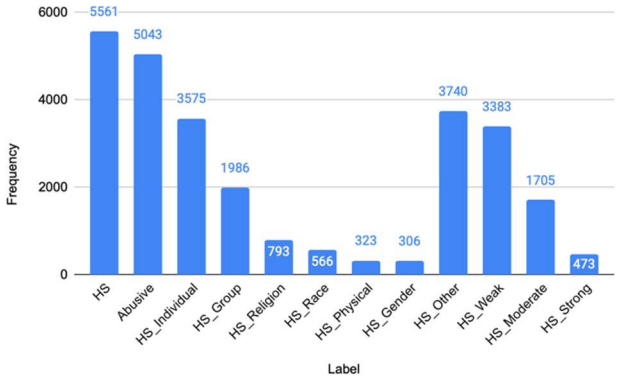


Fig. 2. Amount of tweet labeled as positive per each label

In this research, the data split performed with 9:1 ratio for data training and data test. The cross validation also applied to test the performance of the model.

B. Data Preprocessing

Data preprocessing stages proposed to clean the raw data that just acquired from its source. New obtained data from the real world need to pass the data preprocessing because the dataset still contains many noises and will complicate the classification. Data preprocessing prepare the dataset so that it's not only provides the maximum results of classification, but also effective and efficient performance.

The data preprocessing that proposed in this experiment are data cleaning, case folding, text normalization, stopword removal, and tokenization.

TABLE I. DESCRIPTION OF TARGET LABEL FROM THE DATASET[9]

Label	Explanation
HS	This label shows whether the tweet is classified as hate speech or not
Abusive	This label will be marked as '1' if the tweet contain any abusive words. Abusive words usually exist in a hate speech, but not every tweet contain hate speech included as hate speech.
HS_Individual	This label will shows the hate speech is addressed to certain individual
HS_Group	This label will shows the hate speech is addressed to certain group
HS_Religion	This label classify the tweet as a hate speech to certain religion (Islam, Protestant, Catholic, Buddha, Hindu, Confucianism) or beliefs.
HS_Race	This label classify the tweet as hate speech to certain race or ethnicity.
HS_Physical	This label classify the tweet as a hate speech towards someone or group with dissabilities or different physical appearances.
HS_Gender	This label classify the tweet as hate speech towards certain sexual orientation or certain gender.
HS_Other	This label classify the tweet as hate speech, insult, or slander that did not included as the other label before.
HS_Weak	This label rate the severity level of hate speech as weak hate speech. This kind of hate speech still considered on weak level because the hate speech addressed to certain individual without any incitement or provocation leads to open conflict, the hate speech only considered as personal problem.
HS_Moderate	This label rate the severity level of hate speech as moderate hate speech. The hate speech on this level of severity usually in the form of swearing, blasphemy, defamation, etc. that addressed towards certain groups or communities that could lead to open conflict.
HS_Strong	This label rate the severity level of the hate speech as strong hate speech. The hate speech with this severity level is in the form of swearing, blasphemy, defamation, and the other kind of hate speech contain incitement and provocation that lead them to open conflict.

C. Data Preprocessing

Data preprocessing stages proposed to clean the raw data that just acquired from its source. New obtained data from the real world need to pass the data preprocessing because the dataset still contains many noises and will complicate the classification. Data preprocessing prepare the dataset so that it's not only provides the maximum results of classification, but also effective and efficient performance.

The data preprocessing that proposed in this experiment are data cleaning, case folding, text normalization, stopword removal, and tokenization.

D. GloVe Word Embedding

After the preprocessing stages, we obtained the ready to use data in the token form. Those tokens of words should be transformed into a vector so that the classifier algorithm could understand it. In this stage, GloVe is used. GloVe will map words based on global words co-occurrence statistic and result of local text window to consider the context of each word.

E. Multi-label Text Classification Using BiLSTM Method

In this research, the model built with BiLSTM algoritm. The data in vector form then classified using the BiLSTM model.

The process inside BiLSTM algorithm is similar with LSTM algorithm because they actually formed by the same LSTM cell. The only difference between those two algorithms is the output. BiLSTM has two outputs in the end of the process from each LSTM. Those two outputs should be combined. There are some ways to combine the outputs, i.e.: concatenation, addition, subtraction and etc. According to that, the output of the cell would be concatenated before processed in the dense layer and then classified it by all the labels in the data.

F. Model Evaluation

In this experiment, multi-label classification model will combine the example-based and label-based metrics evaluation. For calculating the accuracy of the model, the metrics would be calculated using the example-based metrics while the precision, recall, and F1-score will be calculated using the label-based measures type micro averages measures.

IV. EXPERIMENT AND RESULT

A. Epoch

Epoch is the amount of iteration would be used to train the model. Epoch is one of the hyperparameters that affect model's performance. When the amount of epoch increased, model generalization towards training dataset will increase. However, if the epoch number is too big, the probability of overfitting happened is getting bigger and model's generalization will be lessening. So therefore, choosing the right amount of epoch before the model starts training is important. It will affect model's performance.

This experiment has been done by trying to adjust some epoch number for the model. The amount of epoch proposed for this experiment are 5, 10, 15, and 20. Table 2 shows the results of this experiment.

TABLE II. EXPERIMENT RESULTS FOR NUMBER OF EPOCHS ADJUSTMENT

Epoch	*Running Time* (s)	*Accuracy* (%)	*F_1 score* (%)
5	484	66.73	68.78
10	714	**69.90**	**73.21**
15	796	61.62	68.10
20	830	52.97	62.58

Based on the experiment results, the best performance obtained by model with 10 epochs reached 69.90% of accuracy and 73.21% of F1-score. From those results, the model tends to overfit when the number of epochs is getting bigger.

B. Data Dimension

In this research, we will experiment with dimension data or the size of data vector that would be used as input to the model. This data dimension value will affect the size of the word vector matrix. If the input vector matrix has a length less than the specified dimensions, then the remaining void will be filled with a zero vector. On the other hand, if the length of the data matrix exceeds the specified dimension value, the rest of the matrix will be removed.

Data dimension have a considerable influence on the performance of the model. Model should choose the right value of data dimension because it will determine the information that will be extracted by the model. Input with too big data dimension, will have too many zero-vector elements. Meanwhile, too small data dimension could omit some important contexts of the data.

In this experiment, we will test some value of data dimensions, i.e.: 50, 100, 150, and 200. Table 2 shows the results from the experiment of data dimensions adjustment. From the results, model with 100 data dimensions input has the best performance. It reaches 69.90% of accuracy and 73.21% of F1-score.

TABLE III. EXPERIMENT RESULTS OF DATA DIMENSION ADJUSTMENT

Data Dimension	*Running Time* (s)	*Accuracy* (%)	*F_1 score* (%)
50	437	71.05	72.01
100	714	**69.90**	**73.21**
150	1.250	67.28	69.60
200	1.504	69.36	69.65

C. Learning Rate

When training a model, learning rate is one of important factors that should be considered. Learning rate have a strong effect to the optimization of weight when training a model. Too big learning rate could pass the optimal point of the model, but too small learning rate will need longer time to reach the optimal point of the model and ineffective.

Table 4 shows the experiment results of adjusting the model's training rate. From the results, we obtained the best performance is when the learning rate in 0.005. The model reached 70.07% of accuracy and 75% of F1-score.

TABLE IV. EXPERIMENT RESULTS OF LEARNING RATE ADJUSTMENT

Learning Rate	*Running Time* (s)	*Accuracy* (%)	*F_1 score* (%)
0.005	850	**70.07**	**75.00**
0.01	714	69.90	73.21
0.015	788	59.42	68.60
0.02	689	35.75	55.03

D. LSTM Unit

LSTM Unit is parameter of hidden states value as one of the inputs for LSTM cell in the BiLSTM model. Hidden state is one of the hyperparameter that affect the model performance. Table 4 shows the experiment results by adjusting the number of LSTM unit. Based on the experiment of adjusting the number of LSTM unit, we obtained that this model the best performance when the LSTM unit is 100.

TABLE V. EXPERIMENT RESULTS OF LSTM UNIT AMOUNT ADJUSTMENT

LSTM Unit	*Running Time* (s)	*Accuracy* (%)	*F_1 score* (%)
10	474	66.47	66.98
25	577	71.07	70.34
50	714	75.06	75.21
100	1834	**77.79**	**78.03**

E. Classification Threshold

For classification model, threshold is the minimum limitation of prediction results to classify it as positive class. This threshold adjustment often applied to the model with imbalance dataset. Default of the threshold from binary classification with 0.5 is considered not good enough for the performance of model with imbalance dataset.

In this experiment, the dataset used is considered imbalanced because there are several labels that consist of a lot of data, but some other labels have too little data. To overcome the data imbalance, threshold adjustments are made to the prediction results so that they can provide maximum performance.

This threshold adjustment is carried out using the brute-force method, where the model will be tested with all possible thresholds applied to the model. This study will make this adjustment from 0 to 1.00 with a distance of 0.01 so that there will be 101 experiments conducted to determine the threshold value that gives the best performance.

From the brute-force test of the threshold value, we get the results as illustrated by the graph in Figure 3. The graph shows the F1-score of each threshold test. The F1-score of the model continues to increase until it reaches its peak, then after that it slowly decreases with increasing threshold. The best result is given by a threshold of 0.42 with 85.30% of F1-score compared to the default threshold of 0.5 which gives 84% of F1-score.

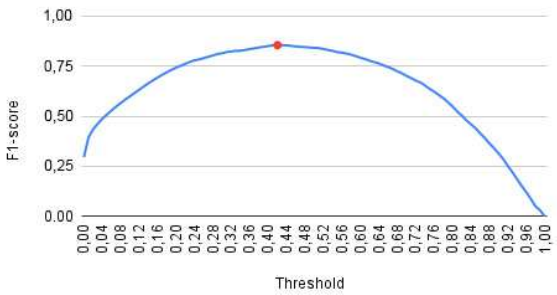


Fig. 3. Results of threshold brute-force method experiment

TABLE VI. TABLE 1 RESULT COMPARISON OF BEFORE AND AFTER THRESHOLD ADJUSTMENT

Before		After	
Metrics	**Score (%)**	**Metrics**	**Score (%)**
Accuracy	82.38	Accuracy	82.31
Precision	88.08	Precision	83.41
Recall	80.28	Recall	87.28
F1-score	84.00	F1-score	85.30

Table 6 shows the difference between before and after the threshold adjustment. The table explains more the detail about the changes that happened after the threshold adjusted.

V. CONCLUSIONS AND SUGGESTIONS

A. Conclusions

Based on the prior experiment we conducted, hate speech could be detected using Bidirectional Long Short-term Memory (BiLSTM) method. Not only detected it as hate speech, but the model also be able to predict the multi-label classification of the tweet. The best model's performance reached with the hyperparameter settings of 10 epochs, 100 vectors of data dimension, 0.005 value of learning rate, 100 LSTM units, and 0.42 value of threshold. From those hyperparameter setting, the best possible model's performance reached are 82.31% of accuracy, 83.41% of precision, 87.28% of recall, and 85.30% of F1-score.

B. Suggestions

There are some suggestions that could be done for further research for the same topic. Following are the suggestions mentioned before:

1. Based on the dataset, this dataset could perform better with hierarchical classification algorithm.
2. Some punctuations and emoticons in the dataset could be considered as a feature, because some of it could differ context and emotion of a sentence. It could improve the quality of prediction.
3. The dataset in this experiment considered as imbalance dataset, so that it would perform better if the dataset is balanced. The data amount of each label should not have too big difference. The data could be added or try some data balancing method.

ACKNOWLEDGEMENT

The Authors thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] F. I. Febriansyah and H. S. Purwinarto, "Pertanggungjawaban Pidana Bagi Pelaku Ujaran Kebencian di Media Sosial," J. Penelit. Huk. Jure, vol. 20, no. 2, p. 177, Jun. 2020, doi: 10.30641/dejure.2020.V20.177-188.

[2] N. Kholis, Buku Saku Penanganan Ujaran Kebencian (Hate Speech). Jakarta, Indonesia: Komisi Nasional Hak Asasi Manusia Republik Indonesia, 2015.

[3] I. A. Permatasari, "Implementasi Undang-Undang Informasi dan Transaksi Elektronik Dalam Penyelesaian Masalah Ujaran Kebencian Pada Media Sosial Implementation of Electronics Information and Transactiom in Completion of the Problem of Hate Speech on Social Media," Jurnal Penelit. dan Komun. Pembang., vol. 23, no. 1, pp. 27–41, 2019.

[4] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - Revisiting neural networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8725 LNAI, no. PART 2, pp. 437–452, 2014, doi: 10.1007/978-3-662-44851-9_28.

[5] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Magnet: Multi-label text classification using attention-based graph neural network," in ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence, 2020, vol. 2, pp. 494–505, doi: 10.5220/0008940304940505.

[6] K. Anandakumar and V. Padmavathy, "A Survey on Preprocessing in Text Mining," Int. J. Adv. Res. Comput. Sci., vol. 4, no. 9, pp. 79–91, 2013.

[7] G. Angiani *et al.*, "A comparison between preprocessing techniques for sentiment analysis in Twitter," 2016.

[8] R. F. N. Firmansyah, M. A. Fauzi, and T. Afirianto, "Sentiment Analysis pada Review Aplikasi Mobile Menggunakan Metode Naive Bayes dan Query Expansion," Doro Ptiik, vol. 8, 2016.

[9] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.

[10] M. A. Fauzi, R. F. N. Firmansyah, and T. Afirianto, "Improving sentiment analysis of short informal Indonesian product reviews using synonym based feature expansion," Telkomnika (Telecommunication Comput. Electron. Control., vol. 16, no. 3, pp. 1345–1350, 2018, doi: 10.12928/TELKOMNIKA.v16i3.7751.

[11] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010, vol. 5, no. 12, pp. 1320–1326, 2010, doi: 10.17148/ijarcce.2016.51274.

[12] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1532–1543, doi: 10.3115/v1/d14-1162.

[13] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling," COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap., vol. 2, no. 1, pp. 3485–3495, Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.06639.

[14] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[15] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction," pp. 1–11, Jan. 2018, [Online]. Available: http://arxiv.org/abs/1801.02143.

[16] H. Wang, R. Zhou, and G. Goos, Systems Engineering – WISE 2020. 2020.

[17] Y. Huang, Y. Jiang, T. Hasan, Q. Jiang, and C. Li, "Topic BiLSTM model for sentiment classification," ACM Int. Conf. Proceeding Ser., vol. Part F1376, pp. 143–147, 2018, doi: 10.1145/3194206.3194240.

[18] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," IEEE Access, vol. 7, no. c, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[19] Z. Hameed, B. Garcia-Zapirain, and I. O. Ruiz, "A computationally efficient BiLSTM based approach for the binary sentiment classification," in 2019 IEEE 19th International Symposium on Signal Processing and Information Technology, ISSPIT 2019, 2019, pp. 0–3, doi: 10.1109/ISSPIT47144.2019.9001781.

[20] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Oregon State Univ. Corvallis, pp. 1–25, 2010, [Online]. Available: http://people.oregonstate.edu/~sorowerm/pdf/Qual-Multilabel-Shahed-CompleteVersion.pdf.

[21] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, 2014, doi: 10.1109/TKDE.2013.39.

[22] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6913 LNAI, no. PART 3, pp. 145–158, doi: 10.1007/978-3-642-23808-6_10.

COVID-19 Social Safety Nets Sentiment Analysis On Twitter Using Gated Recurrent Unit (GRU) Method

Junia Adhani[ab1], Intan Nurma Yulita [ab2], Asep Sholahuddin [b3], Mulya Nurmansyah Ardisasmita [c4], Dwi Agustian [c5]
[a]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[b]*Department of Computer Science, Universitas Padjadjaran*
[c]*Department of Public Health, Universitas Padjadjaran*
Bandung, Indonesia
e-mail: [1]junia.adhani@gmail.com, [2]intan.nurma@unpad.ac.id, [3]asep.sholahuddin@unpad.ac.id, [4]mulya@unpad.ac.id, [5]dwi.agustian@unpad.ac.id

***Abstract*— One of the Indonesian government's programs in dealing with Covid19 problems is the Social Safety Net program which is given to the community, especially Covid19 assistance which is given every month to the community. Based on the assistance provided by the government, many people expressed their opinions through Twitter social media. This study aims to analyze the sentiment on Twitter tweets regarding the Social Safety Net Program from March to December 2020. The data collected is 4061 tweets data. The data is classified into two classes, namely positive and negative. The classification algorithm used is Gated Recurrent Unit (GRU). Hyperparameter testing is carried out in order to produce an optimal model. In the optimal GRU hyperparameter, when there are 10 GRU units, the activation function is sigmoid, the optimizer used is Adam, the batch size is 128, with 10 epochs of iteration and 0.2 dropout size. The GRU model produces an f1score of 92.09%, a precision of 90.34%, and a recall of 93.90%.**

Keywords*— *Sentiment Analysis, Gated Recurrent Unit, Social Safety Net, Twitter

I. INTRODUCTION

Early 2020 in Indonesia began with a global threat to health, namely the emergence of Corona Virus Disease 19 (Covid19). The Covid19 pandemic has greatly changed the order of people's lives, not only in Indonesia but also throughout the world. Activities that are usually carried out face-to-face such as activities at schools, campuses and offices can no longer be carried out normally but carried out virtual. Covid19 or in a scientific language called SARS CoV-2 has a very fast rate of spread, reportedly starting in the end of 2019 from Wuhan, China. In mid-2020, Covid19 has spread to 215 countries in the world with the number of infected as of October 23, 2020, as many as 41,994,442 people with the death toll reaching 1,142,744 and Indonesia is in 19th position in the world with a total of 377,541 cases [1].

Regarding the increasing number of Covid19 cases in Indonesia, the government has prepared a Social Safety Net Program. A social safety net is a form of government involvement in providing assistance to the poor and vulnerable to improve quality of community life [2]. The social safety net is expected to provide an alternative fiscal stimulus that is feasible to implement along with tax incentives and medical facilities, this is manifested in the issuance of Perpu No.1 of 2020, which states that additional funds to reduce Covid19 through social security are 110 billion Rupiah [3].

The Social Safety Net Program, which is a program from the government, is expected to help people affected by Covid19. Many people praised the government's program and some did not agree with the program. So that there are many people who give positive and negative views on the government program of the Social Safety Net. These positive and negative views are very helpful for the government in evaluating the Social Safety Net program, because if the community gives a positive response, it can be interpreted that this program is quite helpful for the community in overcoming the problems of their life needs and in accordance with the objectives of the program, and if the community response tends to be negative it will this should be an evaluation for the government and what kind of assistance is actually needed by the community on the ground. This community response can be seen directly using social media, namely Twitter. Twitter is one of the social media that is very widely used from various circles because with this application the Twitter user community can show its expression. The expression, which is presented in written form, is an individual's sentiment on the current events, especially the view of the Social Safety Net program.

Sentiment analysis is a method to analyze some data to find out human emotions. Sentiment analysis can see how the public's view of the "tweet" that is conveyed via Twitter to the government program. These "tweets" can be classified into the form of positive and negative views. It is hoped that with this sentiment analysis, it can be seen that public opinion is more likely to be positive or negative towards this Social Safety Net program. Sentiment analysis can be done with various methods or algorithms, one of which is RNN (Recurrent Neural Network). This method is a deep learning method but has weaknesses in processing longer sequence conditions and also has short-term memory problems. To overcome the problem of the RNN, the development of the RNN, namely the Gated Recurrent Unit (GRU) emerged.

Gated Recurrent Units were introduced to make each Recurrent Unit adaptively capture dependencies of different timescales. Gated Recurrent Units are a simpler alternative to Long Short-Term Memory and are also quite popular. Just like Long Short-Term Memory, Gated Recurrent Unit can also reduce the Vanishing Gradient problem of Recurrent Neural Network [4]. The Gated Recurrent Unit (GRU) is a special mechanism in training a neural network to have a

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

short-term memory through its internal mechanism known as a gate. Using the Gated Recurrent Unit (GRU) method in sentiment analysis because its layers help the neural network in memorizing contexts in interdependent sentence sequences [5]. Therefore, this sentiment analysis research uses the Gated Recurrent Unit method

II. Literature Review

A. Social Safety Net

The Social Safety Net is one of the programs issued by the government to overcome the problem of Covid19 which has been faced by Indonesia for almost a year. With this Social Safety Net, it is hoped that the community will be helped during difficult times during this pandemic.

B. Sentiment Analysis

Sentiment analysis is one of the fields of research in natural language processing, computational linguistics and text mining. Sentiment analysis or it can also be called opinion mining is a computational study of other people's opinions, appraisals, and emotions contained in entities, events and attributes they have [6]. Sentiment analysis can be categorized into three tasks, namely informative text detection, information extraction, and sentiment interestingness classification (emotional, polarity identification). Sentiment classification (negative or positive) is used to predict sentiment polarity based on sentiment data from users [7].

C. Data preprocessing

Preprocessing is a step that is needed to clean, eliminate, change the data source, both in the form of non-alphabet characters and words that are not needed. It is intended that the data used is more optimal when used [8]. The following are the preprocessing stages used in this study:

- Case folding, the process of converting data into a suitable format such as making all letters lowercase. This aims to reduce data redundancy that will be used in the classification process so that the calculation process becomes optimal.
- Noise Removal, a process to clean words from things that are not needed. Things that are not needed such as symbols, numbers, punctuation marks, words followed by numbers to URLs (Uniform Resource Locator).
- Tokenizing, the process of separating or cutting data in the form of phrases or clauses. Tokenizing is done to make it easier for the process to convert into vectors.
- Stopword Removal, the removal of words that have no relevant meaning.
- Word2vec, the algorithm was created by Mikolov et al. in 2013. Word2Vec is widely used in NLP research. This model is an unsupervised learning application using a neural network consisting of a hidden layer and a fully connected layer. Word2Vec relies on local information from the semantic language that is learned from a particular word influenced by the surrounding words. There are two Word2vec algorithms, namely Continuous Bag-of-Word (CBOW) and Skip-Gram [9].
- Imbalance Class, a condition where there is an imbalance in the dataset against the target feature. For example, the positive target is 70% while the negative target is only 30%. The imbalance class can usually be resolved by using the SMOTE method. SMOTE or Synthetic Minority Oversampling Technique is a derivative of oversampling. SMOTE is a popular method used to deal with class imbalances. This technique synthesizes a new sample from the existing minority class by sampling [10].

D. GRU

Gated Recurrent Unit (GRU) is a development of the RNN architecture. The GRU is a fairly recent development proposed by K. Cho in 2014. GRUs are much simpler in structure and perhaps more practical than LSTMs. Similar to the LSTM unit, where the GRU has a gate unit that modulates the flow of information within the unit, but without having a separate memory cell [12].

III. Research Method

A. Data Collection

Data collection is done by the online scrapping technique. The scrapping process is carried out by first looking for keywords, namely @Covid19 assistance, @social safety net on Twitter. Then the online scrapping process is carried out using the web cloud.scrapehero.com. Data retrieval from Twitter is filtered from February 2020 to December 2020. The February retrieval is because in February the Corona pandemic has started to become a major problem until last December 2020. From the scrapping results, 4061 tweets data were obtained which were divided into two columns, namely the 'content' column as the tweets sentence and the 'Sentiment' column which was stored in a *.csv file extension. Of the 4061 tweets, there were 2875 positive classes and 1186 negative classes.

B. Data Labeling

After the data collected from the scrapping results are stored in a *.csv file extension, then the data will be manually labeled positive and negative. Tweets are checked one by one and then labeled according to the sentence. Positive labels are given a score of 0 and negative labels are given a score of 1.

C. Preprocessing Data

This stage is a stage that has a major influence on the results of the quality of the data that will be processed in the next stage. Each sentence and word will be processed into several stages, namely noise removal, case folding, tokenizing, and stop removal

D. Distribution Of Training, Validation, and Testing Data

The data in this study will be grouped into training, validation, and testing data. Training data is used for creating and train the model to be used. The model will then be validated to reduce overfitting, and testing data is used to test the model that has been created. The data used in making the model amounted to 3654 tweets data.

E. Word2vec

Word2vec is a set of several interrelated models that are used to produce word embedding. This study uses the

Continuous Skip-gram model and built using the gensim library. This skip-gram model predicts the words that are around the target word.

F. GRU

This section describes the calculation process carried out by the GRU. The layer used is 1 and the activation function used is sigmoid and tanh.

IV. RESULTS

A. GRU SMOTE and GRU without SMOTE

Based on Fig. 1, it can be seen that the accuracy of the model when using SMOTE has decreased because the resulting instances have a high degree of similarity, thereby reducing the accuracy results. The precision value in the model with SMOTE has increased. Meanwhile, the recall value and f1 score of the model using SMOTE decreased due to a decrease in positive class instances (majority class) in the dataset which was predicted to be positive instance class due to the addition of negative class instances (minority).

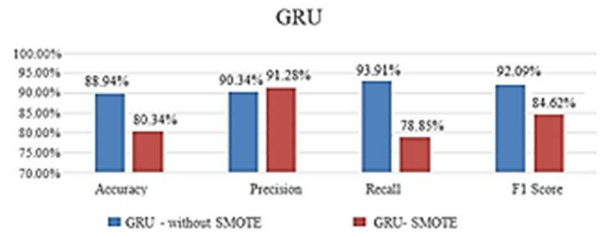


Fig. 1. Comparison visualization of GRU SMOTE and GRU without SMOTE

B. Hyperparameter Test

Hyperparameters are important in building a model. To produce a good model, optimal hyperparameters are needed, so hyperparameter testing is needed. In Word2vec, the hyperparameters tested are the dimensions and the window size.

TABLE I. TEST RESULTS DIMENSIONS ON WORD2VEC HYPERPARAMETERS

Test Order	Dimension Size	FI Score (%)
1	30	91,003
2	35	90,973
3	40	92,495
4	45	90,909
5	50	90,747

1) *Dimension size*: Based on the tests in Table I, it can be seen that the most optimal dimension size is dimension 40. This means that the dimension size 40 can represent the existing word sentiment. Meanwhile, if it is smaller or greater than 40, the f1 score will be smaller. Dimensions that are too large will result in additional vectors that do not represent sentiment, on the other hand, if the dimensions are smaller, a vector does not show the meaning of the existing sentiment data.

TABLE II. TEST RESULTS DIMENSIONS ON WORD2VEC HYPERPARAMETERS

Test Order	Dimension Size	FI Score (%)
1	15	92,089
2	20	91,929
3	25	91,570
4	30	90,876
5	35	90,747

2) *Window size*: Based on Table II, it can be seen that the window size of 15 gives optimal results. This is because the larger the window size does not produce a better f1 score because the larger the window size value, the more word contexts will cause the similarity value to be weak.

3) *Unit size*: Based on Table III, it can be seen that the unit size of 10 gives optimal results, whereas if the unit is smaller than 10 or larger than 10, it gives a smaller value.

TABLE III. UNIT TEST RESULTS ON GRU HYPERPARAMETERS

Test Order	Unit	FI Score (%)
1	5	89,692
2	10	91,710
3	15	90,322
4	20	90,125
5	30	89,891

4) *Dropout* size: The dropout sizes used in this test are 0.1, 0.2, 0.3, 0.4 and 0.5. Based on the purpose of the dropout which serves to improve the neural network by reducing overfitting and the results of the tests in Table 4, it can be seen that the most optimal results when using dropout 0.2. Because the dropout size is close to 0 it will produce an overfitting model, while if it is too close to 1 it will produce an underfitting model so that 0.2 produces a more optimal model.

TABLE IV. DROPOUT TEST RESULTS ON GRU HYPERPARAMETER

Test Order	Dropout	FI Score (%)
1	0,1	89,964
2	0,2	92,091
3	0,3	90,747
4	0,4	90,484
5	0,5	90,054

C. Result of Sentiment Regarding Social Safety Net

Sentiment analysis of the Social Safety Net program obtained 4061 data. From the 4061 data obtained 2875 data labeled positive and 1186 labeled negative whose data was taken from March to December 2020. It can be seen in Figure 2 is the percentage of positive and negative sentiment. From this percentage, it can be concluded that the community accepts this social safety net program very well.

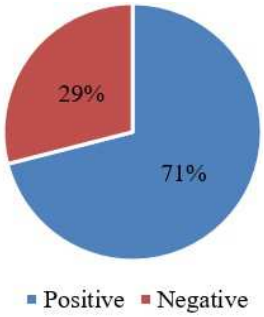


Fig. 2. Sentiment percentage

V. RESULTS

This study was conducted to analyze the sentiment of public opinion using data from social media Twitter on the Social Safety Net program totaling 3246 with 2596 positive tweets and 1058 negative tweets to build the Gated Recurrent Unit (GRU) model. This study preprocesses the data first, namely by noise removal, case folding, stopword removal and tokenization. Based on the research that has been done, there are several conclusions as follows:

a) When building the model, an imbalanced class problem occurs, so the SMOTE method is used. The model that did not use SMOTE turned out to give better results than the model that was built with SMOTE.
b) To build an optimal GRU model, the Word2Vec model must also be optimal. The Word2vec model used is the result of optimal hyperparameter testing so that the best Word2Vec model is obtained with dimensions of 40, window size 15, learning rate 0.002, min count 4, epoch 50, which produces the best f1 score with 92.089%.
c) The Gated Recurrent Unit (GRU) model in this study was built by first testing the hyperparameters so as to produce an optimal model. The best GRU model is using units of 10, sigmoid activation function, Adam optimizer, batch size 128, epoch 10, and dropout 0.2, which produces the best f1 score with 92.091%.
d) To be able to see people's opinions, using the optimal Gated Recurrent Unit (GRU) method and implemented in the form of a web-based application so that it can be viewed quickly and efficiently compared to manuals.

Based on the research that has been done, there are suggestions that can be made on the development of further research namely The addition of the dataset must be considered again in order to produce a better model because if you use a deep learning algorithm, it will be better if the dataset is large so that the resulting model is better. Also, at the preprocessing stage, it must be considered again to overcome the imbalanced class problem in addition to using the SMOTE method.

ACKNOWLEDGMENT

The Author thanks the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[2] Jamaruddin, R. B. Arshad, “Social Safety Nets To The Quality of Life in Indonesia”, International Journal of Humanities & Social Sciences Studies (IJHSSS), 96-108, 2017.

[3] N. T. Sihaloho, M. Silalahi, and B. Sujendra, “COVID19 : Policy Evalution to Protect Communities Through Social Safety Net”, Journal of Governance and Political Social UMA (JPPUMA), 124-133, 2020.

[4] M. Lango, "Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study", Foundations of Computing and Decision Sciences, 44(2), 151-178, 2019

[5] S. Pedipina1, R. Dhanalakshmi, “Sentiment Analysis of Twitter Political Data using GRU Neural Network”, International Journal of Advanced Science and Technology, 5307- 5320, 2020.

[6] H. Zou, X. Tang, B. Xie, B. Liu, "Sentiment classification using machine learning techniques with syntax features", 2015

[7] S. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature”, International World Wide Web Conference Committee, 751-760, 2010.

[8] M. Pejic-Bach,T. Bertoncel, M. Meško, abd Z. Krstić, "Text mining of industry 4.0 job advertisements", International journal of information management, 50, 416-431, 2020.

[9] S. Thavareesan, and S. Mahesan,"Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts", 2020 Moratuwa Engineering Research Conference (MERCon) (pp. 272-276), 2020.

[10] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary", Journal of artificial intelligence research, 61, 863-905, 2018.

[11] F. Wang, X. Zhang, R. Fu, and G. Sun, "Study of the home-auxiliary robot based on BCI" Sensors, 18(6), 1779, 2018.

[12] S. Biswas, E. Chadda, and F. Ahmad, “Sentiment Analysis with Gated Recurrent Units”, Advance in Computer Science and Information Technology (ACSIT), Volume 2, No 11, pp 59-63, 2015.

COVID-19 Detection In Chest X-Rays Using Inception Resnet-v2

Raihan Badrahadipura[ab1], Syachrul Qolbi Nur Septi [bc2], Julio Fachrel [ab3], Intan Nurma Yulita [bd4],
Anindya Apriliyanti Pravitasari [ab5], Dwi Agustian[e6]
[a]Department of Statistics, Universitas Padjadjaran
[b]Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran
cDepartment of Mathematics, Universitas Padjadjaran
[d]Department of Computer Science, Universitas Padjadjaran
[e]Deptartment of Public Health, Faculty of Medicine, Universitas Padjadjaran
Bandung, Indonesia
e-mail: [1]raihan18021@mail.unpad.ac.id, [2]syachrul17001@mail.unpad.ac.id, [3]julio19001@mail.unpad.ac.id, [4]intan.nurma@unpad.ac.id,
[5]anindya.apriliyanti@unpad.ac.id, [6]dwi.agustian@unpad.ac.id

***Abstract*— Screening for COVID-19 is a vital part of the triage process. The current COVID-19 gold standard, the RT-PCR test, is regarded to be costly and time consuming. Artificial intelligence can be utilized to identify COVID-19 in radiographic pictures to overcome the limitations of existing testing methods. This study describes how the Inception-ResNet-v2 architecture was used to categorize pictures into three categories using transfer learning (Normal, Viral Pneumonia, and COVID-19,). Despite only running for 29 epochs, the resultant model had an accuracy of 0.966. This demonstrates the utility of AI in the diagnosis of illnesses.**

Keywords—COVID-19, chest X-Ray, artificial intelligence, deep learning, convolutional neural networks, Inception-ResNet-v2

I. INTRODUCTION

The COVID-19 Pandemic has had unanticipated consequences around the world. As bed occupancy rates rise due to high infection rates and modest symptoms, hospitals and other healthcare facilities are becoming overburdened. Swiftly screening for the disease is one way to mitigate this. Real-time reverse transcription polymerase chain reaction (RT-PCR), benchmark of COVID-19 tests, is revered for its specificity, yet it is time, labor, and resource inefficient. Furthermore, RT-PCR sensitivity is quite variable, with early findings in China indicating low sensitivity. Because of how the sample was obtained and how the RT-PCR was performed, there was a lot of fluctuation in positivity rates due to how the sample was taken and time from symptom onset[1].

Chest x-rays might be used as an alternate screening approach to look for visual indications linked with severe acute respiratory syndrome coronavirus 2. (SARS-CoV-2). In chest x-ray images and computed tomography (CT) images, Ng et al. and Guan et al. discovered radiographic abnormalities such as ground-glass opacity, bilateral abnormalities, and interstitial abnormalities [2], [3].

The Inception-Resnet-v2 is a type of convolutional neural network (CNN) model that uses the inception module and residual connections to avoid overfitting, maintain image data integrity, and obtain higher accuracy, precision, sensitivity, and F1-scores. This study uses transfer learning techniques to develop a deep learning model using the Inception-Resnet-v2 [4].

The end goal of this study is to prove that diagnosis using machine learning is indeed effective and should be widely adopted to handle screening for an abundance of conditions. The authors of this paper hope to make a portable X-ray device that can diagnose patients on the spot.

II. LITERATURE REVIEW

A. Chest X-Rays

A chest x-ray, abbreviated as CXR is a radiographic projection of the chest, that is commonly used to diagnose ailments affecting the chest and other structures nearby. It employs ionizing radiation to create images of the chest[5]. An x-ray image is represented digitally in the form of a composite of pixels, each of which has a different light intensity. This intensity value can be represented by a two-dimensional matrix, with each pixel as an element of the matrix [6]. The images created in this process are typically greyscale images, but can be processed into RBG channels later on in the preprocessing stage.

B. Computer Vision

Computer vision is an interdisciplinary scientific field that answers the question on how to teach computers high level comprehension of digital images or video. Computer vision technology is one of the fields of Artificial Intelligence technology. A classic problem statement in computer vision is how to classify an image into two or multiple classes[7]. Other forms of computer vision might include object detection using bounding boxes, style-transfer and many other tasks. Computer vision is a growing field and the limits of its uses is not known yet.

The essence of computer vision technology is to duplicate the ability of human vision into electronic objects so that electronic objects can understand the meaning of the images entered [8]. Understanding the image on a computer is done by deciphering the information acquired from the image data using a variety of tools such as statistics, geometry, and learning theory as well as other methods [7]. In the discipline of technology, computer vision finds out how models and theories can be applied to system development in computer systems. Examples of some applications that use computer vision such as navigation

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

tools and controllers. Computer vision is now often used to detect faces in images, recognize facial expressions and in practice it is often used in tandem with artificial neural networks. In this paper we will attempt to produce a model that can classify chest x-ray images into three classes (COVID-19, Viral Pneumonia and Normal).

C. *Deep Learning*

The use of artificial neural networks in machine learning is commonly referred to as deep learning, which is a method of creating models to perform supervised, semi-supervised, or unsupervised learning. In this work, the approach utilized was supervised learning with labeled data [9].

Using deep learning enables the delivery of satisfactory results with end-to-end modelling without the need for human feature extraction from input data. Mimicking the human brain's method enables the interpretation of data like as pictures, sounds, and text. [10].

The idea of deep learning was proposed by Hinton [11] based largely on the deep belief network (DBN), in which an unsupervised greedy layer-by-layer training method promises to tackle the deep structure optimization problem. After that, a multi-layer automatic encoder's deep structure is presented. Furthermore, LeCun [12] presents the concept of a convolution neural network, which is the original learning technique with multiple layers that use relative space relations to minimize parameter count and improve training performance.

In general, deep learning architecture is made up of three layers. The input is made up of neurons that take data from variables. The hidden layer, which lies between the input and output levels and is where all the computations take place, is the second layer.

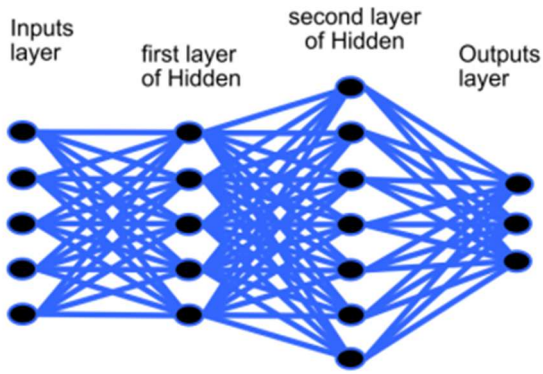


Fig. 1. Deep learning architecture

D. *Convolutional Neural Networks*

CNNs, or convolutional neural networks, are a form of artificial neural network used to analyze visual data. CNNs use convolutional layers or kernels, which are made up of filters that move the pixels of an input picture represented as a matrix. Pooling layers and thick layers are also utilized in addition to convolutional layers. [13].

A CNN's basic notion is that it has the ability to analyze complex features using local information [14]. The CNN architecture results in hierarchical feature extraction, namely filters that are trained for specific purposes, for example, the first layer is usually focused on identifying edges or color fluctuations, then the second layer is usually more focused on identifying shapes, and the next layer filters are usually more focused on studying the parts of the structure. Objects, both those that are seen a little or partly or those that are seen quite a lot and the last layer is used to identify objects [15]. Input, convolutional, pooling, fully connected (FC) layers, and output layers make up a CNN. Figure 2 illustrates the architecture.

The convolutional layer has a collection of kernels that use "stride(s)" to convolve all inputs such that the output quantity's dimensions become integers. [16]. After the convolutional layer is utilized to conduct the striding process, the input level scale is reduced. As a result, zero-padding [17] It's necessary to set the input level to zero and keep the input quantity dimension at a low stage.

The input matrix is I, a 2D filter of size m, n is K, and F is the output of a 2D feature map. The action of the convolutional layer is denoted by the letter I*K. In feature maps, the rectified linear unit (ReLU) layer is employed to increase nonlinearity.

To limit the quantity of parameters, the pooling layer downsamples a particular input dimension. The most frequent method is max pooling, which results in the maximum value of the targeted region. The FC layer is utilized as a classifier, making a choice based on features extracted from the convolutes.

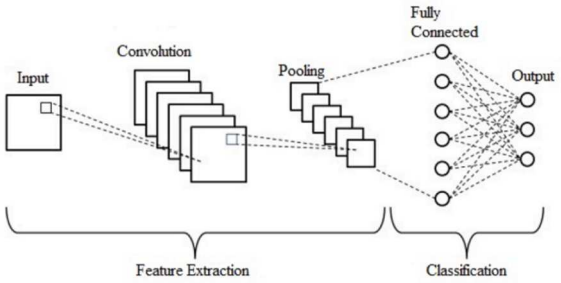


Fig. 2. Convolutional Neural Network architecture

E. *Inception-Resnet-v2*

Convolutional neural networks can take on many shapes and sizes, and one of them is Inception-Resnet-v2, an architecture that has been trained ImageNet database that consists of more than a million images belonging to a hundred classes. It uses the Inception module and residual connections to make for better training and higher accuracy [4]. The Inception-Resnet-v2 architecture was created as a combination between the Inception and ResNet architectures, by employing the Inception module and residual connections in its design.

Residual connections create shortcuts in the model to train the neural network more deeply which can improve model performance. It also significantly simplifies the inception block that was present in Inception v3. Inception Resnet v2 is more accurate than the previous model. Inception Resnet v2 has better performance than previous models including Inception Resnet v1 which has lower computational costs but also lower accuracy.

In 2016, Szegedy introduced the Inception-ResNet-v2 architecture, which combines inception blocks and residual layers. Residual connections are used to minimize the deterioration that deep networks cause while also lowering training time. The use the weights in these layers that have been pre-trained to help us detect COVID19 in X-Ray pictures. The 164 layers that make up the architecture include 20 Inception-Resnet blocks that heighten accuracy and speed up training [18].

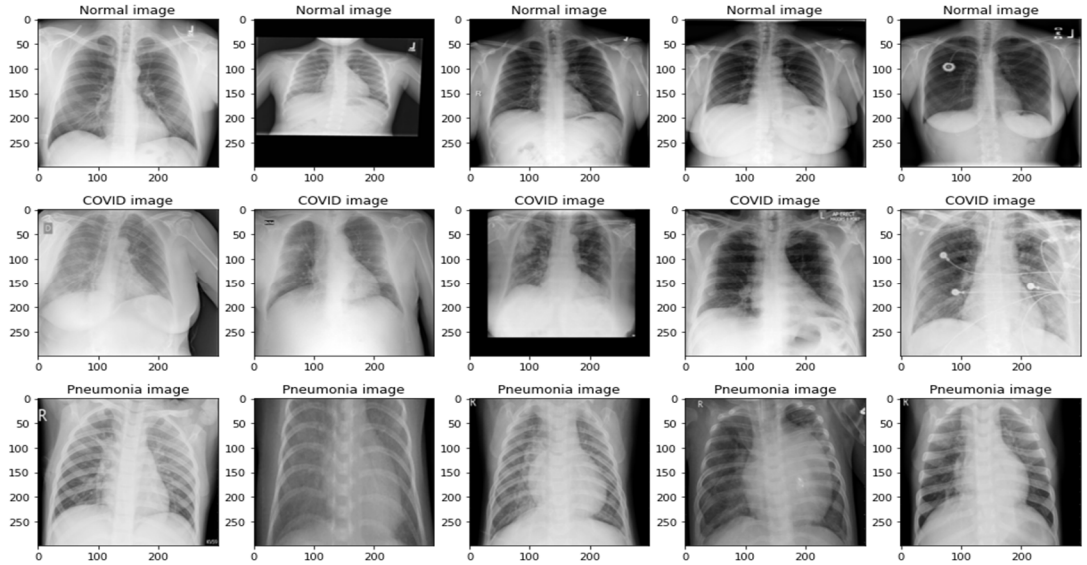


Fig. 3. The images in the frst, second, and third rows show 5 sample images of normal cases, COVID-19 cases, and pneumonia cases respectively

F. Transfer Learning

Transfer learning is concerned with retaining information gained from solving a problem and applying it to another similar problem. In this scenario, the weights and biases from each layer are saved. [19]. Building models using transfer learning allows researchers to bypass the common problems encountered during model building, such as the lack of data for training and insufficient computing power for training the deep learning models. This makes building models easier to train and also easier to deploy. But every method indeed has its imperfections. One such example is the differences between the image data from the pretrained model and the trained model.

III. MATERIALS AND METHODS

A. Dataset

The COVID-19 Radiography Dataset, which is an opensource dataset available on kaggle.com, is used in this work. There are 3,616 COVID-19, 10,192 Normal, 6012 Non-Covid lung infection, and 1,345 viral pneumonia pictures in total; however, this study only utilizes the COVID-19, normal, and viral pneumonia images, leaving out the non-Covid lung infection. All of the photos are in the Portable Network Graphics (PNG) file type and are 299 × 299 pixels in size. The pictures have three channels because they are in RGB format.

The training and testing portions of the dataset are then divided 80:20. In the table below, you'll find the following information:

TABLE I. TRAIN AND TEST SPLIT

Class	**Training Images**	**Testing Images**
Normal	8154	2038
COVID-19	2893	723
Pneumonia	1076	269

B. Preprocessing

In order to input the data into the model and make for better training on the targeted features, image preprocessing methods such as changing the image shape, rescaling, and image augmentation are implemented to the dataset.

The images are resized to 224 x 224 pixels and the values for each pixel is scaled to a minimum value of 0 and a maximum of 254. Image augmentation is also implemented by rotating the image randomly for 15 degrees, the resulting empty space caused by this transformation is filled with the nearest pixel.

The preprocessing is done using TensorFlow's ImageDataGenerator that automatically splits and preprocesses the data since editing manually will be quite laborious ad time consuming.

C. Model Architecture

Inception-ResNet-v2 combined with several additional layers to make for better accuracy is used in this paper. The Inception architecture is combined with residual connections in a convolutional neural network of 164 layers. The end layer that used to contain a dense layer with a SoftMax function to classify a hundred classes I the ImageNet dataset

is replaced with an end layer that classifies into the three classes encountered in the dataset. The following diagram illustrates the architecture used.

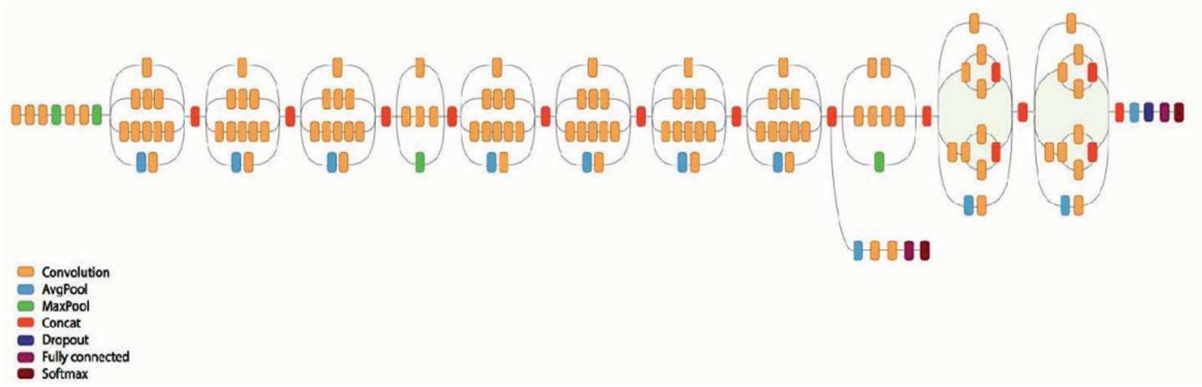


Fig. 4. Architecture of Inception-ResNet-v2 network [20]

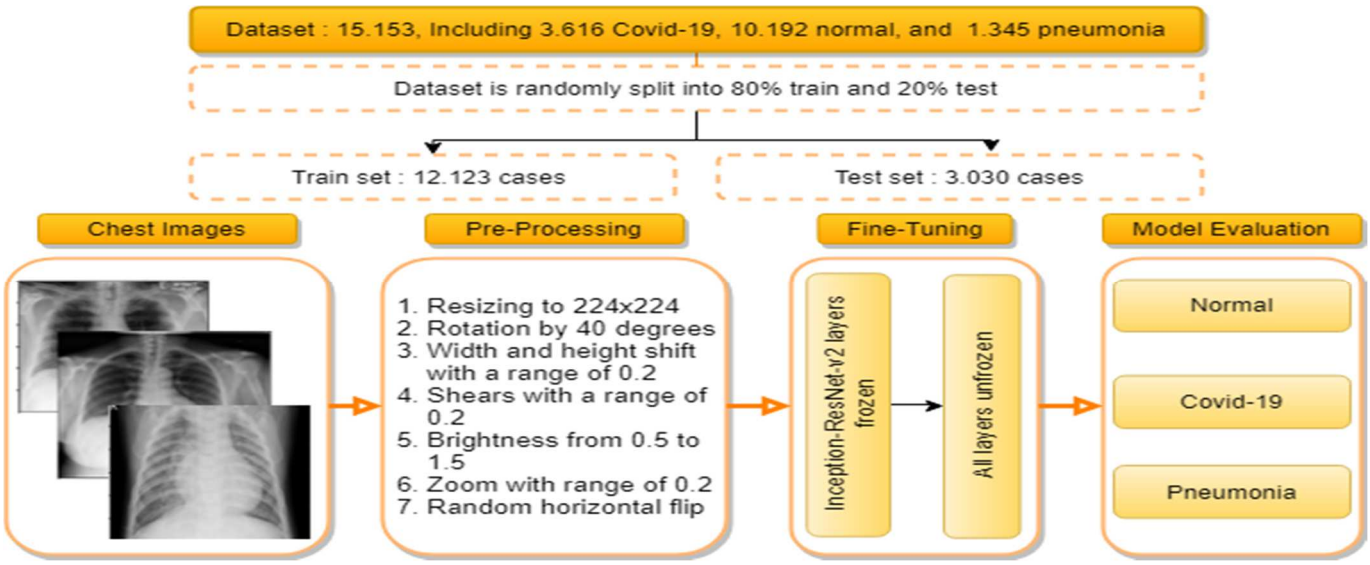


Fig. 5. Schematic representing training and testing phase of the proposed schem

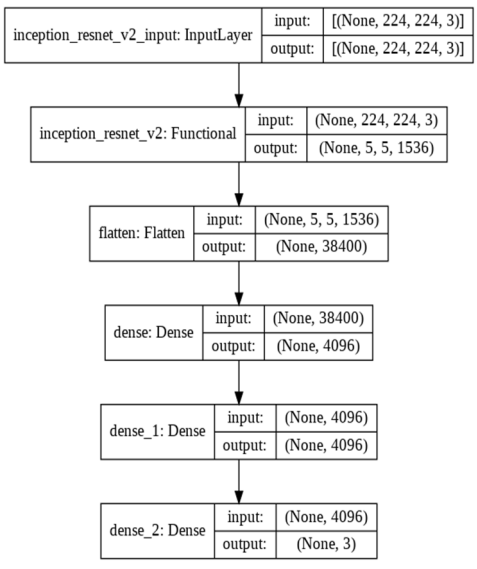


Fig. 6. Inception-ResNet-v2 Diagram

The Inception-ResNet-v2 architecture is depicted in Figure 6. The Inception-Resnet blocks employ multiple scale factors, enabling each block's output to be scaled in different ranges. The network accepts photos with size of 224x224x3.

D. Model Training and Fine Tuning

The model is then compiled with the Adam optimizer using a learning rate of $1e^{-5}$, batch size of 128, and decay of $2e^{-7}$. With categorical cross-entropy as the model's loss function and accuracy set as the model's evaluation metric for training. The model's training is set to 50 epochs, but uses an early stopping monitor to monitor validation data, using a minimum delta of $1e^{-3}$ with a patience of 5 epochs. Using this method enables faster training time and ensures that the best weights are restored. The training for this model is done twice, first with the Inception-ResNet-v2 layers frozen so that it may maintain the weights and biases stored from the ImageNet dataset, the only layers trained are the layers added after the Inception-ResNet-v2. The second training is done with all layers unfrozen. Figure 5 shows a

schematic diagram that illustrates the training, and testing phase.

E. Model Evaluation

The model is then assessed using a confusion matrix, which yields assessment metrics including accuracy, sensitivity, precision, recall, and F1-scores. These metrics are used to evaluate the quality of the model and if the model is viable to use.

IV. Results and Discussion

A. Model Training

With a maximum validation accuracy of 0.9259, the first round of training with frozen layers was conducted for 19 epochs. The second test had ten epochs and a maximum validation accuracy of 0.9658. The diagrams below depict the second run.

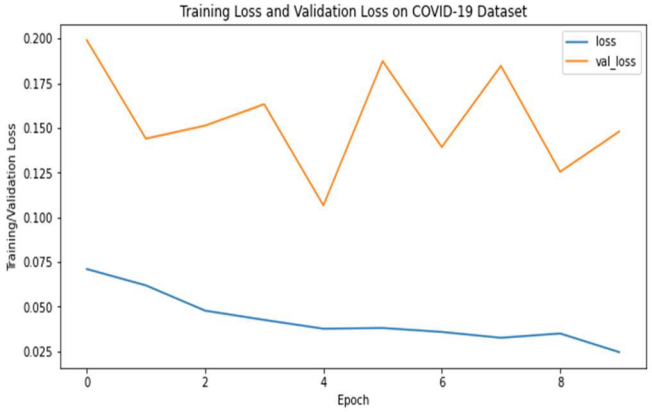


Fig. 7. Training and Validation Loss

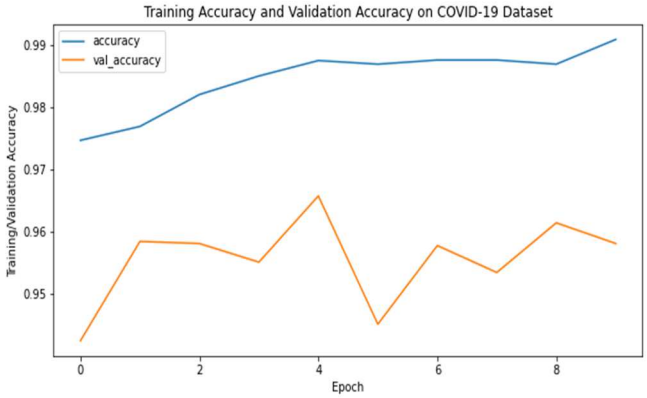


Fig. 8. Training and validation accuracy

Even though the model was run for only 19 and 10 epochs decent accuracy was able to be attained

B. Model Evaluation

The following is a confusion matrix of the model's performance on the test dataset.

TABLE II. Confusion Matrix

Actual Class	Predicted Class		
	COVID-19	***Normal***	***Viral Pneumonia***
COVID-19	679	41	3
Normal	48	1990	0
Viral Pneumonia	0	11	258

With the information gained from the confusion matrix, it is observed that the model most often confuses between the COVID-19 and Normal Classes, while the Viral Pneumonia class is rarely misclassified, although there are still instances of misclassification.

TABLE III. Classification Report

	Precision	**Recall**	**F1-Score**	**Support**
COVID-19	0.93	0.94	0.94	723
Normal	0.97	0.98	0.98	2038
Viral Pneumonia	0.99	0.96	0.96	269
Accuracy	-	-	0.97	3030
Macro Average	0.97	0.96	0.96	3030
Weighted Average	0.97	0.97	0.97	3030

It is observed from the table above, that the model is better at classifying images belonging to the Viral Pneumonia and Normal classes than the COVID-19 class as inferred from the precision, recall and F1-Scores. The model's overall accuracy is 0.966. For the macro average ad weighted averages, the model has obtained a score of 0.97 except for the recall and F1-score macro averages.

C. Conclusion

Transfer learning is a very powerful technique for training and deploying models quickly. Diagnosis using machine learning is proven here to be accurate but requires further development using higher capacity CPUs. But this idea remains to be tested as there hasn't been any real-world experiments carried out yet to the author's knowledge. Further research can be done using this model and creating a portable X-ray device that can feed directly into the model that can be stored in an IoT device such as a Raspberry-Pi. The author of this paper also recommends using other techniques such as using Visual Transformers or other types of convolutional neural networks to create a better and more accurate model.

ACKNOWLEDGMENT

The Author thanks the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] Y. Yang et al., "Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections," medRxiv, p. 2020.02.11.20021493, Feb. 2020.

[2] M.-Y. Ng et al., "Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review," https://doi.org/10.1148/ryct.2020200034, vol. 2, no. 1, p. e200034, Feb. 2020.

[3] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," Lancet, vol. 395, no. 10223, pp. 497–506, Feb. 2020.

[4] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor," IEEE Access, vol. 8, pp. 55135–55144, 2020.

[5] M. FA, H. W, Y. TT, and M. M, "Effective doses in radiology and diagnostic nuclear medicine: a catalog," Radiology, vol. 248, no. 1, pp. 254–263, Jul. 2008.

[6] Blinn, J. F, What is a pixel? IEEE Computer Graphics and Applications, 25(5), 82–87. https://doi.org/10.1109/MCG.2005.119, 2005

[7] D. Forsyth, J. Ponce, Computer vision: a modern approach, 2nd Editio. Pearson, 2012.

[8] SONKA, Milan; HLAVAC, Vaclav; BOYLE, Roger. Image processing, analysis, and machine vision. Cengage Learning, 2014.

[9] C. Ranjan, "Understanding Deep Learning: Application in Rare Event Prediction," J. Chem. Inf. Model., vol. 1, no. 1, pp. 1–426, 2021.

[10] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[11] G. E. Hinton, "Deep belief networks," Scholarpedia, vol. 4, no. 6, p. 5947, 2009.

[12] Y. Lécun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 2001.

[13] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," Proc. - IEEE Int. Symp. Circuits Syst., vol. 2018-May, Apr. 2018.

[14] Islam, Md Zabirul, Md Milon Islam, and Amanullah Asraf. "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images." Informatics in medicine unlocked 20: 100412. 2020

[15] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[16] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, Chen T. Recent advances in convolutional neural networks. Pattern Recogn 2018

[17] K. Avcı, A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-term memory networks. Sensors 2019;19:1992.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. "Inception-V4 Inception-Resnet And The Impact Of Residual Connections On Learning", 2017

[19] B. A. M. Ashqar, S. S. Abu-Naser, "Identifying Images of Invasive Hydrangea Using Pre-Trained Deep Convolutional Neural Networks," Int. J. Control Autom., vol. 12, no. 4, pp. 15–28, Apr. 2019.

[20] A. DEMİR, F. YILMAZ, "Inception-ResNet-v2 with Leakyrelu and Averagepooling for More Reliable and Accurate Classification of Chest X-ray Images" In 2020 Medical Technologies Congress (TIPTEKNO) (pp. 1-4). IEEE.

[21] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, C. Wang. Machine learning and deep learning methods for cybersecurity. Ieee access, 6, 35365-35381. 2018.

Preprocessing Application for Car Insurance Claim Classification Model

Farhan Gunadi
Research Center for Artificial Intelligence and Big Data
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
farhan19004@mail.unpad.ac.id

Muhammad Fauzi
Research Center for Artificial Intelligence and Big Data
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
muhammad19246@mail.unpad.ac.id

Bagas Firdaus
Research Center for Artificial Intelligence and Big Data
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
bagas19005@mail.unpad.ac.id

Afrida Helen
Research Center for Artificial Intelligence and Big Data
Department of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
helen@unpad.ac.id

***Abstract*—The use of data in various industry sectors becomes a necessity in predicting or deciding on one of them is the use of car insurance data related to claims made by vehicle owners. The data can later be used by insurance companies to be analyzed related to car insurance claims both from the owner's side of the car and from the condition of the car. This paper will discuss the preprocessing data on auto insurance claim data in America with the aim of later making a data model so that it can be used next to see the accuracy of the data processed classification. The results of processing data and data classification can help car insurance companies in deciding a policy or problem that occurs accurately and measurable. The data used in this paper is data that still has a missing value. Therefore, data cleaning is done by cleaning, filtering, and combining these data. The study used the Car Insurance Claim Data dataset downloaded on kaggle's website. The results showed that the JRIP algorithm had the best accuracy of 83.09 percent (before the preprocessing dataset) and 83.14 percent (after the preprocessing dataset was applied) in the 10 fold cross-validation test mode. With an increased level of accuracy, the data can be better used again as an example for forecasting tau as a reference company to trigger something related to the data.**

Keywords— data, model, preprocessing

I. INTRODUCTION

Data mining is the process of manipulating data by extracting previously unknown information from large datasets [1]. These days, data mining is often used in several industries including insurance. This is based on the growth in the number of users of four-wheeled vehicles from year to year which experienced a very rapid increase which was influenced by the high level of mobility of the community in their daily lives [2]. The high number of car users also affects the number of car insurance users. Therefore, we are researching to process the data of insurance users who claim to the insurance concerned which can later be used for the company's purposes in determining company policies or deciding something related to car insurance [3]. However, the car insurance claim data used is data that still contains a lot of missing values, therefore it is necessary to do preprocessing to improve data accuracy and clean data from missing values [4].

The author preprocesses to clean and repair the data. In addition, in solving classification problems, the use of methods or techniques will facilitate the classification process [5]. Classification is one of the data mining techniques used to create models of unclassified data samples for use in classifying new data samples into similar classes [6]. The process of data classification consists of learning and classification. In learning data training is analyzed using classification algorithms, then in the classification used data testing to ensure the accuracy of the classification rule used. Classification techniques are divided into five categories based on differences in mathematical concepts, namely statistical-based, distance-based, decision tree-based, neural network-based, and rule-based [7]. This research proposes comparing models in classification techniques in data mining. The classification techniques used are JRIP, Decision Table, One-R, Lazy-IBK with the aim of comparing the accuracy of correct data and data errors from the methods used.

One-R is the simplest classification method that is on target and ignores all predictors [8]. One-R classifiers only predict the majority (class) category. Although there is no predictive power in One-R, it is useful for determining basic performance as a benchmark for other classification methods. Lazy-Ibk is the classifier of K's nearest neighbor. You can also do distance weighting. JRIP applies proportional rule learning. Repeated incremental pruning to result in error reduction (RIPPER) [9]. The decision table is a method used to explain and describe the flow of data logically stored in it that can be used to solve a problem. The decision table works by combining all existing conditions where these conditions contain rules that are stored in the form of tables on a problem so that it can be ascertained that no possibility is passed in the logical analysis of the problem.

In this paper, the authors have done data processing on the data used. By conducting several tests from each stage of data processing with the aim of knowing the accuracy of the data from before the processing data to after data processing with some filtering. The authors use preprocessing and classification processed car insurance data to obtain clean data and data accuracy that can later be used to create data models that can be used for forecasting, business strategies, and decision-making tools.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

II. LITERATURE STUDY

The data source in this test is car insurance claim data by Xiaomengsun which can be obtained easily on Kaggle's website and uploaded on October 15, 2018. The data used in this study were 10.303 data. The study used the Car Insurance Claim Data dataset because in this data there are missing values so we will do preprocessing to find the best algorithm that can improve accuracy, the training data set contains 8161 observations of 26 variables (one index, two response variables, and 23 predictor variables). This dataset is based on a customer dataset of an auto insurance company. Build two predictive models that estimate:

- the probability of a customer having a car accident
- the monetary amount of insurance claims in the event of an accident. After an initial inspection of the variables, three logistic regression models and two multiple linear regression models were prepared and compared with the test data.

Based on the classification performance metrics, the best model is suggested and applied to this dataset. The data in this is estimated to be taken in the United States in the California section because California has a Transport Improvement Fee (TIF) remember given the dollar monetary values, and MVR and possible TIF definitions.

III. METHODOLOGY

In this research process, several stages are carried out. The initiation stages of several activities range from problem identification and literature studies to finding alternative solutions to existing problems. The next stage, which is the stage of model development by preparing datasets with missing values for research and the selection of algorithms from a number of suitable alternatives for dataset processing. We do some stages. Here are the steps we do, namely:

1) Stages of Data Collection

The data used for this study comes from car insurance claims data uploaded on October 15, 2018 by Xiaomengsun. Such data includes secondary and public data obtained from the kaggle website. The data obtained is raw data amounting to 10.303 data. The data consists of 27 attributes and have 26 variable, The variables that describe the various aspects of income are positively correlated with each other and with age and years in the same job: INCOME, HOME_VAL (value of the house), BLUEBOOK (value of the car), AGE, YOJ (years in the same job), HOMEKIDS (number of children in the household) is negatively correlated with age, income and age of the car, CLM_FREQ (frequency of claims in the last 5 years), as well as OLDCLAIM (the total amount claimed) and MVRPTS (the number of Vehicle Recors Points) are weakly positively linked to TARGET_AMT (the payment in case of a car accident). The data must go through the selection and preprocessing stages before being used in research.

2) Stages of Algorithm Selection

In this study, we ensured that the data that exists is more than 10.000 instances and from the data has missing value and must overcome it, because not all data is filled all well, much data is lost or does not exist because it is difficult to find or others. Then it is necessary to clean up the data so that the classification of datasets is more relevant and accuracy is good.

- Integrating Data
  Data integration is the merging of data from various databases into one new database. Here is the process of changing data or converting data in numerical form. Just like the previous stage, at this stage, the process of integrating data uses two tools that support each other.
- Target data
  This stage is an explanation of the data that is ready to be used in the data mining process, where the data target becomes the input for subsequent processes.

Algorithm selection in this study we selected 4 algorithms to classify data, namely:

a) JRIP

This class implements a proportional rule learner, Repeated Repeated Pruning to Result in Error Reduction (RIPPER)

b) Decision Table

an accurate method for numerical prediction of the decision tree and is a sequenced set of If-Then rules that has the potential to become more compact and therefore more understandable than a decision tree. Selection to explore decision tables because simpler, computational algorithms are less intensive than the Tree-based decision approach.

c) One-R

A simple classification algorithm that builds a one-level decision tree. The One-R algorithm was first created by Robert C. Holte in 1993. The One-R algorithm ranks average error attributes in training data such as entropy measurements in C4.5 (Holte 1993) [10]. One-R treats the numeric value of an attribute as a continuous value and uses a simple method to divide the value coverage into several separate intervals. One-R creates one rule for each attribute in the training data then selects the rule with the smallest error calledonerule (Buddhinath & Derry 2005) [11].

d) Lazy-IBK

Use distance sizes to find "close" instances in training data for each test instance and use those selected instances to make predictions.

The reason we use these four algorithms is that the best algorithm to classify the data that we will examine is the fourth and only algorithm that has improved the accuracy of the data.

3) Stages of Preprocessing

Before the implementation stage is applied, the preprocessing stage is first carried out. The amount of initial data that can be obtained from data collection is as much as 10.303 data, but not all data is used and not all attributes are used because the data must go through the initial processing stage of data or called data preparation.

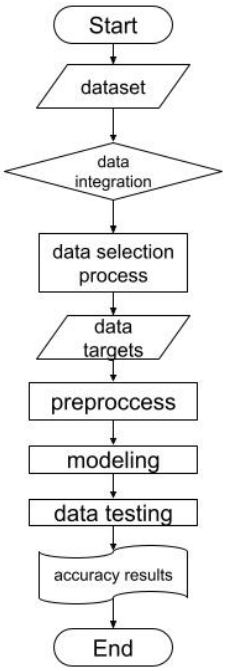


Fig. 1. Research flowchart

The earliest stage to do this research is to start. The second stage uses 10.303 car insurance claim datasets to be tested for accuracy to be selected in the next stage. Data that is not qualified will produce data mining that is not qualified. Quality decisions should be based on quality data (double data or data that has been lost can lead to untruths). that is by correcting the inconsistency of data filling the value that has been lost in incomplete data (missing value). In the data selection process stage selects the column to be tested is URBANCITY (nom) totaling 10.303 instances. The target data step is an explanation of the data that is ready to be used in the data mining process, where the target data becomes the input for subsequent processes.

Before processing data mining is often required preprocessing. Data preprocessing describes the types of processes that carry out raw data to prepare for other procedural processes [12]. The purpose of preprocessing in data mining is to transform data into a format whose process is easier and more effective for the needs of the user, with the following indicators: Getting more accurate results Reduction of computing time for large scale problems Makes the value of data smaller without changing the information it contains. There are several methods used for preprocessing such as: Sampling selects a representative subset of large data populations, discrete (part of the reduction data but has its own significance, especially for numerical data), deleting lost, impute, and feature selection.

The modeling stage is the best filter search stage and the best algorithm, which is the best algorithm we found for this data is JRIP, Decision Table, One-R, Lazy-IBK. After getting modeling for the dataset we tested, the data we tested to see the results of accuracy. We are looking for an improved accuracy result. Then, after the data in the test will come out the results of accuracy and we look for the results of accuracy increased from before.

IV. RESULT AND ANALYSIS

This section represents the results obtained after testing four data mining classification models. There are two phases of the experiment, the first phase is without the application of dataset preprocessing and the second phase is after the application of the dataset preprocessing. The results of the classification of models in data mining for car insurance claims without the application of dataset preprocessing are shown in Table I.

TABLE I. EXPERIMENTAL RESULTS ON CLASSIFICATION MODELS FOR CAR INSURANCE CLAIMS WITHOUT DATASET PREPROCESSING

No	Method	Accuracy (%)
1	JRIP	83.09%
2	Decision Table	82.80%
3	One-R	81.14%
4	Lazy-IBK	78.30%

In Table I, the accuracy value is the number of instances that are classified correctly divided by the number of all instances. The results of testing the classification model without the application of the dataset preprocessing showed the best accuracy value in the JRIP method with an accuracy value of 83.09%. Then the results of testing with other methods showed that the Decision Table method had an accuracy value of 82.80%. Followed by the One-R method with an accuracy value of 81.14% and the Lazy-IBK method with an accuracy value of 78.30%.

After experimenting on the classification model for car insurance claims without dataset preprocessing, the authors conducted a number of experiments in the form of preprocessing applications to each method used using different parameters. This experiment was conducted to determine whether there will be a difference in classification accuracy after the application of dataset preprocessing or not. The results of the classification of models in data mining against targets for car insurance claims after the application of preprocessing to the dataset are shown in Table II.

In Table II, variable 1 and variable 2 are the used preprocessing parameters. The results of testing the classification model after applying the dataset preprocessing showed the best accuracy value in the JRIP method after being applied to standardize and merge in frequent nominal values preprocessing on the dataset with an accuracy value of 83.14%. It showed that there was an increase in accuracy of 0.05% from the accuracy of the JRIP method without the application of the dataset preprocessing, which was 83.09%.

TABEL II. EXPERIMENTAL RESULTS ON CLASSIFICATION MODEL FOR CAR INSURANCE CLAIM AFTER DATASET PREPROCESSING APPLICATION

No	Method	Variable 1	Variable 2	Accuracy (%)
1	JRIP	Normalize	Numeric To Nominal	82.64%
2	JRIP	Standardize	Merge In Frequent Nominal Values	83.14%
3	Decision Table	Normalize	Numeric To Nominal	80.51%
4	Decision Table	Cartesian Product	PKI-Discretize	82.80%
5	One-R	Normalize	Numeric To Nominal	80.68%
6	One-R	Cartesian Product	PKI-Discretize	81.14%
7	Lazy-IBK	Standardize	Merge In Frequent Nominal Values	78.30%
8	Lazy-IBK	Supervised Discretize		82.39%

Then the test results showed that the highest accuracy value increase was in the Lazy-IBK method after being applied supervised discretize preprocessing to the dataset with an accuracy value of 82.39%. It showed that there was an increase in accuracy of 4.09% from the accuracy of the lazy-IBK method classification without the application of the dataset preprocessing, which was 78.30%. The results also showed that the highest accuracy value decrease was in the Decision Table method after being applied normalized and numeric to nominal preprocessing to the dataset with an accuracy value of 80.51%. It showed that there was a decrease in accuracy of 2.29% from the accuracy of the Decision Table method classification without the application of the dataset preprocessing, which was 82.80%.

Meanwhile, the results of testing with other methods or other preprocessing parameters showed that the JRIP method after being applied normalized and numeric to nominal preprocessing had an accuracy value of 82.64%. It showed that there was a decrease in accuracy of 0.45% from the accuracy of the JRIP method without the application of the dataset preprocessing, which was 83.09%. The Decision Table method after being applied cartesian product and PKI-discretize preprocessing in the dataset had an accuracy value of 82.80%. It showed that there was no increase or decrease in accuracy from the accuracy of the Decision Table method without the application of the dataset preprocessing, which was 82.80% too. The One-R method after being applied normalize and numeric to nominal preprocessing in the dataset had an accuracy value of 80.68%. It showed that there was a decrease in accuracy of 0.46% from the accuracy of the One-R method without the application of the dataset preprocessing, which was 81.14%. The One-R method after being applied cartesian product and PKI-discretize preprocessing in the dataset had an accuracy value of 81.14%. It showed that there was no increase or decrease in accuracy from the accuracy of the One-R method without the application of the dataset preprocessing, which was 81.14% too. Then the Lazy-IBK method after being applied standardize and merge in frequent nominal values preprocessing in the dataset had an accuracy value of 78.30%. It showed that there was no increase or decrease in accuracy from the accuracy of the Lazy-IBK method without the application of the dataset preprocessing, which was 78.30% too.

From the experiments that have been done in table I and table II, it was shown that there was a change in the accuracy value on some models after certain preprocessing parameters were applied and there were some models that did not show a change in accuracy after certain preprocessing parameters were applied. The changes in accuracy value also vary for each method and its preprocessing parameters, it can be increased or decreased. Based on that, it was concluded that the dataset preprocessing can increase the accuracy value of model classification for car insurance claim when using the optimal preprocessing parameters for each method used. It can be said that the dataset preprocessing leads to a performance increase [3]. Otherwise, the dataset preprocessing can also decrease the accuracy value of model classification for car insurance claim when using the wrong preprocessing parameters for each method used. Meanwhile, there are some preprocessing parameters that have no effect on the accuracy value after being applied to certain models.

V. CONCLUSION

Many algorithms, as well as preprocessing parameters, can be proposed for classification on the car insurance claim dataset. But in this study proposed classification using the methods JRIP, Decision Table, One-R, and Lazy-IBK. After testing using tools, the highest classification accuracy results were obtained by the JRIP model followed by Decision Table, One-R, and Lazy-IBK models.

Of the four models, experiments were conducted again to find optimal preprocessing parameters for each model. For JRIP model, the optimal parameters obtained are standardized and merged in frequent nominal preprocessing values with an increase in accuracy value of 0.05%. Then for the Decision Table model, the optimal parameters obtained are cartesian product and PKI-discretize preprocessing in the absence of an increase or decrease in the value of accuracy. Similar to One-R model, the optimal parameters obtained are cartesian products and PKI-discretize preprocessing in the absence of an increase or decrease in the value of accuracy. Then for Lazy-IBK model, the optimal parameter is supervised discretize preprocessing with an increase in accuracy value of 4.09% which is the highest accuracy value increase.

From the experiments that have been done on each model, it was found that the dataset preprocessing can increase the accuracy value of model classification for car insurance claim when using the optimal preprocessing parameters for each method used. Otherwise, the dataset preprocessing can also decrease the accuracy value of model classification for car insurance claim when using the wrong preprocessing parameters for each method used.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021

REFERENCES

[1] S. Vijayarani, M. Muthulakshmi, “Comparative Analysis of Bayes and Lazy Classification Algorithms”, vol 3, issue 8, 2013.

[2] A. S. Rani, S. Jyothi, "Performance analysis of classification algorithms under different datasets", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1584-1589.

[3] N. Eén, A. Biere, “Effective preprocessing in SAT through variable and clause elimination”, Theory and Applications of Satisfiability Testing, vol 3569, pp. 61-75, 2005.

[4] V. Rajeswari, K. Arunesh, “Analysing soil data using data mining classification techniques”, Indian Journal of Science and Technology, 9(19), 2016. https://doi.org/10.17485/ijst/2016/v9i19/93873

[5] A. Gupta, A. Mohammad, A. Syed, and N., M. “A comparative study of classification algorithms using data mining: Crime and accidents in Denver city the USA”, International Journal of Advanced Computer Science and Applications”, 7(7). https://doi.org/10.14569/ijacsa.2016.070753

[6] V. S. Parsania, N. N. Jani, and N. H. Bhalodiya, “Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis”, vol 3, 2014.

[7] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, “Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms”, Knowledge Discovery and Data Mining, pp. 847-855, 2013.

[8] A. W. Moore, D. Zuev, “Internet Traffic Classification Using Bayesian Analysis Techniques”, Measurement and Modeling of Computer Systems, vol 33, no. 1, pp. 50-60, 2005.

[9] D. Lu, Q. Weng, “A Survey of Image Classification Methods and Techniques For Improving Classification Performance”, Journal of remote sensing, vol. 28, no. 5, pp. 823-870, 2007.

[10] S. R. Kalmegh, “Comparative Analysis of the WEKA Classifiers Rules Conjunctiverule & Decision table on Indian News Dataset by Using Different Test Mode”, Volume 7 Issue 2Ver III, February 2018.

[11] V. Vijayakumar, R. Nedunchezhian, “A Study on Video Data Mining”, International Journal of Multimedia Information Retrieval, vol 1, issue 3, pp 153-172, 2012.

[12] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

Market Basket Analysis on Sales Transactions for Micro, Small and Medium Enterprises Using Apriori Algorithm to Support Business Promotion Strategy in RDA Hijab

Aulia Ghassani Nabila[a1], Intan Nurma Yulita[ab2], Ino Suryana[a3], Mira Suryani[a4]
[a]*Department of Computer Science, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Bandung, Indonesia
e-mail: [1]auliaghassaninabila@gmail.com, [2]intan.nurma@unpad.ac.id, [3]ino.suryana@unpad.ac.id, [4]mira.suryani@unpad.ac.id

***Abstract*—The increasingly fierce business competition requires business owners to always innovate to create effective and efficient business promotion strategies to develop their business. One way to get recommendations for business promotion strategies is to process sales transaction data owned by the business. With data mining, the data can be processed so that new knowledge is obtained to support the creation of new business promotion strategies. The method used is the method of market basket analysis association rules with a priori algorithm. The results of the analysis in this study can be concluded for the rule with a combination of 2 items that have the highest confidence value of 41.61%, the lift ratio value of 12.10, and the frequency of the rule in the highest testing data is 66, that is, if consumers buy navy neck cuffs, then he will also buy white neck cuffs. And for the rule with a combination of 3 items that have the highest confidence value of 50.00%, the lift ratio value of 11.94, and the frequency of the rule in the highest testing data is 13, that is, if the consumer buys black neck cuffs and navy neck cuffs then he will also buy a light khaki neck cuff.**

Keywords— Data Mining, Market Basket Analysis, Apriori, Confidence, Lift Ratio, Frequency

I. INTRODUCTION

The development of technology today is very fast. This makes business competition in trade and technological advancements today very tight. Today's businesses must be able to adapt to rapid and complex changes. This requires developers and business people to always be innovative in looking for ideas to develop their business with a business strategy that ensures the sustainability of their business.

RDA Hijab is one of the micro, small and medium enterprises that was founded in 2016. RDA Hijab is engaged in clothing and accessories, especially inner hijab such as ciput, cufflinks, leggings, tank tops, scrunchies, and other basic essentials moslem wear products. RDA Hijab has served various types of transactions to have sales transaction data that can be processed to determine customer behavior regarding which products are often purchased simultaneously. To process this, data mining is needed.

Data Mining can turn mountains of data into valuable information or knowledge by using certain methods or algorithms. In the retail business this is known as the shopping basket method or market basket analysis. Market basket analysis is a method of finding out information by analyzing consumer spending behavior patterns specifically from a particular group or group [1]. If discussed in more depth there is the term association rules. Association rules are one of the techniques in market basket analysis to find relationships between items in a dataset. The association rule consists of two stages, namely the search for frequent itemset (item combinations) and the formation of association rules. In this case, the a priori algorithm will be used to find a combination of items based on items purchased by customers simultaneously from sales transaction data so that information or knowledge is obtained [2] to be utilized by the Hijab RDA team to form new business strategies

II. LITERATURE STUDY

A. Data Mining and Association Rule

Data mining is one of the processes in knowledge discovery in database (KDD), namely the process of processing a collection of data so that knowledge is found in the form of interesting patterns or information using certain techniques and algorithms [3]. There are five stages in KDD, namely data selection, preprocessing/cleaning, transformation, data mining, interpretation/evaluation.

Association rule is a data mining method that aims to find associations between a combination of items [3]. There are two stages in the association rules, namely the analysis of high-frequency patterns and the formation of association rules [4]. In the high-frequency pattern stage, an a priori algorithm is usually used as a technique to determine frequent itemset [5]. Furthermore, in the stage of forming association rules, interesting patterns will be generated (interesting). The association rules will be said to be interesting if the support value in the association rules exceeds the minimum support value determined at the beginning by the researcher and the confidence value in the association rules exceeds the minimum confidence value determined at the beginning by the researcher [6]. This is commonly referred to as an interestingness measure [7], here is the formula for calculating support and confidence values:

Support value for one item:

$$Support(A) = \frac{(transaction\ amount\ contains\ A)}{(total\ transaction)} \quad (1)$$

Transaction amount contains A:

$$Support(A,B) = \frac{(transaction\ amount\ contains\ A\ and\ B)}{(total\ transaction)} \quad (2)$$

The value of confidence in the formation of association rules:

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

$$Confidence\ P(A|B) = \frac{(transaction\ amount\ contains\ A\ and\ B)}{(total\ transaction\ contains\ A)} \quad (3)$$

B. Market Basket Analysis with Apriori Algorithm

In the association rules, the term Market Basket Analysis is often heard. Market basket analysis is the process of analyzing buyer habits by finding relationships between items purchased at the same time and in the same shopping basket [8].

A priori algorithm is one way to find frequent itemsets in association rules proposed by Agrawal and Srikan in 1994 [9]. One of the steps in the association rule is the analysis of high-frequency patterns that can be carried out by the a priori algorithm for that the a priori algorithm is included in the association rule method [10]. The following are the stages of performing the a priori algorithm [11]:

- Formation of candidate itemset
- Calculation of support for each candidate k-itemset
- Set high-frequency pattern
- If no new high-frequency pattern is found, the whole process is stopped

C. Lift Ratio

Lift Ratio is a parameter to measure how strong the result of the rule resulting from the association rule process [12]. The lift ratio will provide information on how much the validity of the relationship between the products from the results of the rules formed so that it can be used as an evaluation parameter in this study [13]. The following is the formula for lift ratio [14]:

$$Lift\ Ratio = \frac{Support\ (A + B)}{Support\ (A) \times Support\ (B)} \quad (4)$$

III. Research Methodology

A. System Design

The steps used in this study are depicted in Fig 1. :

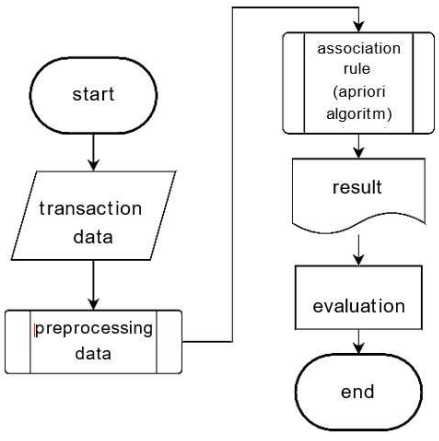


Fig. 1. Research flow

- Transaction data. The sales transaction data used in this study was for twelve months starting from December 1, 2019 to November 30, 2020 with a total of 10,182 sales transaction data. The dataset used in the study will be divided into two parts, namely training data (9 months) and testing data (3 months).
- Preprocessing Data. The nine attributes of the transaction data need to be pre-processed so that the data is cleaner and more selected so that only two attributes are used in this study, namely the date and product attributes.
- Association Rule-Algoritma Apriori. To analyze the sales transaction data in this study, the process of association rules for market basket analysis will be carried out using the apriori algorithm. The a priori algorithm will help the frequent itemset search process.
- Evaluation. After doing data mining, the next step is to evaluate the results of the analysis process. The analysis of the results will be a reference for drawing conclusions and suggestions for this study. The evaluation parameters used in this study are the lift ratio value and checking the frequency of association rules in data testing.

IV. Results and Discussion

A. System Implementation

Some examples of RDA Hijab website development pages with data mining can be seen as follows.

Fig. 2. Apriori process page

Fig. 2 is the a priori process page that contains a form to input the data used in the a priori process, namely start date, end date, minimum support, and minimum confidence. These data are supporting data to carry out iterations so that the results of association rules are obtained. After the data is inputted, the iteration results will be displayed in the form of a table below the data input form. The table shown is the result of iteration 1, the results of iteration 1 that pass, the results of iteration 2, the results of iteration 2 that pass, the results of iteration 3, and the results of iteration 3 that pass.

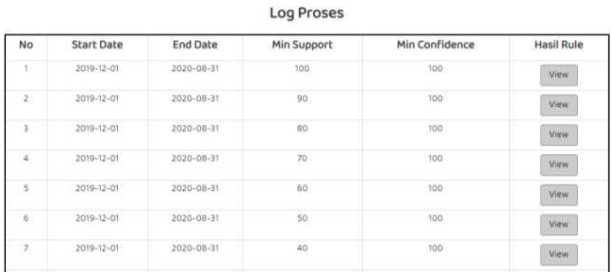


Fig. 3. Rule results page

Fig. 3 is the rule result page in which there is a table containing the history of a priori processes that have been carried out, the table column contains data that has previously been inputted to perform a priori processes, namely start date, end date, minimum support, and minimum confidence. In addition, there is also an action column called the view rule where the manager can see the results of the association rules that are formed. There are three menus that the manager can choose from, namely the results of the confidence itemset, the results of the lift test, and the results of the analysis

Hasil Confidence Itemset

Confidence Itemset 3

No	X -> Y	Support X	Support X dan Y	Confidence	Status

Confidence Itemset 2

No	X -> Y	Support X	Support X dan Y	Confidence	Status
1	Manset Leher Light Khaki->Manset Leher Putih	1,03	4,19	24,66	Lolos
2	Manset Leher Putih->Manset Leher Light Khaki	1,03	3,44	30,04	Lolos
3	Manset Leher Light Khaki->Manset Leher Hitam	1,56	4,19	37,16	Lolos
4	Manset Leher Hitam->Manset Leher Light Khaki	1,56	5,96	26,13	Lolos
5	Ciput Bandana Cream->Ciput Bandana Abu Muda	1,20	3,91	30,80	Lolos
6	Ciput Bandana Abu Muda->Ciput Bandana Cream	1,20	4,43	27,16	Lolos

Fig. 4. Confidence results page

Fig. 4 is the confidence itemset result page in which there are the results of the confidence values formed in each itemset. The confidence value is obtained from the calculation of the number of transactions containing (xUy) divided by the number of transactions containing x. The results of the calculation of the confidence value are displayed in the form of a table with the item combination column formed (combination 1 (x) → combination 2 (y)), the support value x, the support value from xUy, the confidence value, and the status of the itemset formed whether it passes. or not pass.

Hasil Rule Asosiasi (Uji Lift)

No	X -> Y	Confidence	Nilai Uji Lift
1	Manset Leher Light Khaki->Manset Leher Hitam	37,16	6,24
2	Ciput Bandana Abu Muda->Ciput Bandana Light Khaki	32,59	4,60
3	Manset Leher Putih->Manset Leher Hitam	41,98	7,05
4	Ciput Bandana Deep Khaki->Ciput Bandana Hitam	34,05	4,08

Fig. 5. Lift test results page

Fig. 5 is the lift test results page in which there is a table of lift test values formed in each itemset. The lift test value is obtained from the calculation of the support value (xUy) divided by the value of support x which was previously multiplied by the value of support y. The results of the calculation of the lift test value are displayed in tabular form with the combination of items formed (combination 1 (x)→ combination 2 (y)), confidence value (xUy), and lift test value.

Hasil Analisa

No	Hasil
1	Jika konsumen membeli Manset Leher Light Khaki konsumen juga akan membeli Manset Leher Putih
2	Jika konsumen membeli Manset Leher Putih konsumen juga akan membeli Manset Leher Light Khaki
3	Jika konsumen membeli Manset Leher Light Khaki konsumen juga akan membeli Manset Leher Hitam
4	Jika konsumen membeli Manset Leher Hitam konsumen juga akan membeli Manset Leher Light Khaki
5	Jika konsumen membeli Ciput Bandana Cream konsumen juga akan membeli Ciput Bandana Abu Muda

Fig. 6 is the analysis results page which contains a table of analysis results from the process that has been carried out in the form of sentences to make it easier for managers to read and understand. The results of this analysis are formed from the results of the lift test carried out. The results of this analysis are displayed in the form of a table consisting of a column number and the results of the analysis formed.

B. Experiment Results

First Trial Results. In the first experiment, research was conducted on training data from sales transaction data by providing a minimum value of 50% confidence and a minimum value of support of 20. In iteration one, 293 item sets were produced, 147 item sets were passed so that the item sets were reused in the second iteration process and 146 item sets had status. does not pass so that the itemset is not reused in the next iteration process. In the second iteration, 10731 item sets were produced, 92 item sets were passed so that the item sets were reused in the third iteration process and 10639 item sets were not passed so that the item sets were not reused in the next iteration process. In the third iteration, 3509 item sets were produced, 23 item sets were passed and 3486 item sets were not passed. From the process of the first experiment, 11 rules were formed.

Second Trail Results. In the first experiment, research was conducted on training data from sales transaction data by providing a minimum value of 40% confidence and a minimum support value of 60. In iteration one, 293 item sets were produced, 60 item sets were passed so that the item sets were reused in the second iteration process and 233 item sets had status. does not pass so that the itemset is not reused in the next iteration process. In the second iteration, 1770 item sets were produced, 19 item sets were passed so that the item sets were reused in the third iteration process and 1751 item sets were not passed so the item sets were not reused in the next iteration process. In the third iteration, 215 item sets were produced, 0 item sets were passed and 215 item sets were not passed. From the process of the first experiment, 2 rules were formed.

Third Trail Results. In the first experiment, research was conducted on training data from sales transaction data by providing a minimum value of 30% confidence and a minimum value of support of 100. In iteration one produced 293 item sets, 36 item sets had passed status so that the item sets were reused in the second iteration process and 257 item sets had status. does not pass so that the itemset is not reused in the next iteration process. In the second iteration, 630 item sets were produced, 7 item sets were passed so that the item sets were reused in the third iteration process and 623 item sets were not passed so that the item sets were not reused in

the next iteration process. In iteration three it produces 52 item sets, 0 item sets are passed and 52 item sets are not passed. From the process of the first experiment, 4 rules were formed

C. Analysis Results of Minimum Support and Minimum Confidence to the Rule of Combination of 2 Items

Table 1 shows the results of the analysis of the minimum support with a fixed minimum confidence of 40%:

TABLE I. SUPPORT ANALYSIS OF 2 ITEM COMBINATIONS

No	Min Sup	Min Conf	Lift Ratio	Conf	Detail Rule	Freq
1	100	40%	7,05	41,98	Manset Leher Putih->Manset Leher Hitam	49
2	60	40%	12,10	41,61	Manset Leher Navy->Manset Leher Putih	66
3	20	40%	8,24	49,09	Manset Leher Cream->Manset Leher Hitam	14

It can be seen from the three rules that the highest frequency is owned by the rule "navy neck cuff -> white neck cuff" with a total frequency of 66 with a confidence of 41.6%. Frequency shows the number of sales transaction data containing items in the rule obtained from testing data. So that the rule chosen in the research analysis of the minimum support for the combination of 2 items is "navy neck cuff -> white neck cuff". This rule will be used as one of the recommendations for making the RDA Hijab business promotion strategy.

Table 2 shows the results of the analysis of the minimum confidence with a fixed minimum support of 100.

TABLE II. EXAMPLE OF COMBINATION OF 2 ITEM CONFIDENCE ANALYSIS

No	Min Sup	Min Conf	Lift Ratio	Conf	Detail Rule	Freq
1	100	40%	7,05	41,98	Manset Leher Putih->Manset Leher Hitam	49
2	100	30%	6,24	37,16	Manset Leher Light Khaki->Manset Leher Hitam	66
3	100	30%	4,60	32,59	Ciput Bandana Abu Muda->Ciput Bandana Light Khaki	14
4	100	30%	4,08	34,05	Ciput Bandana Deep Khaki->Ciput Bandana Hitam	51
5	100	20%	1,50	22,78	Scrunchie Cream->Scrunchie Hitam	53
…	…	…	…	…	…	…

The highest frequency was formed by the rule "light khaki neck cuff -> black neck cuff" with a total frequency of 66 and the highest confidence of 37.1%. Frequency shows the number of sales transaction data containing items in the rule obtained from testing data. So the rule chosen in the analysis of the minimum confidence in the rule of the combination of 2 items is "light khaki neckcuff -> black neck cuff". This rule will be used as one of the recommendations for making the RDA Hijab business promotion strategy.

D. Results of Analysis of Minimum Support and Minimum Confidence to the Rule of Combination of 3 Items

The Table 3 shows the results of the analysis of the minimum support with a fixed minimum confidence of 50%:

TABLE III. ITEM COMBINATION SUPPORT ANALYSIS

No	Min Sup	Min Conf	Lift Ratio	Conf	Detail Rule	Freq
1	30	50%	16,54	56,90	Manset Leher Light Khaki, Manset Leher Navy->Manset LeherPutih	4
2	30	50%	11,83	52,38	Ciput Bandana Sandalwood, Ciput Bandana Cream->Ciput Bandana Abu Muda	1
3	20	50%	11,94	50,00	Manset Leher Hitam ,Manset Leher Navy->Manset Leher LightKhaki	13
4	20	50%	17,07	58,70	Manset Leher Hitam ,Manset Leher Navy->Manset Leher Putih	6
5	20	50%	13,13	55,00	Manset Leher Navy ,Manset Leher Abu Muda->Manset Leher Light Khaki	4
6	20	50%	13,26	55,56	Manset Leher Hitam ,Manset Leher Abu Muda->Manset Leher Light Khaki	4
7	20	50%	13,49	56,52	Manset Leher Putih ,Manset Leher Abu Muda->Manset Leher Light Khaki	2
8	20	50%	16,80	57,78	Manset Leher Abu Muda , Manset Leher Light Khaki->MansetLeher Putih	2
…	…	…	…	…	…	…

The highest frequency is formed by the rule "black neck cuff, navy neck cuff -> light khaki neck cuff" with a total frequency of 13 and a confidence of 50%. Frequency shows the number of sales transaction data containing items in the rule obtained from testing data. So the rule chosen in the analysis of the minimum support for the combination of 3 items is the rule "black neck cuff, navy neck cuff -> light khaki neck cuff". This rule will be used as one of the recommendations for making the RDA Hijab business promotion strategy.

Table 4 is the result of the analysis of minimum confidence with a fixed minimum support of 20:

TABLE IV. EXAMPLE OF 3 ITEM COMBINATION CONFIDENCE ANALYSIS

No	Min Sup	Min Conf	Lift Ratio	Conf	Detail Rule	Freq
1	20	50%	18,17	62,50	Manset Leher Navy,Manset Leher Abu Muda->Manset Leher Putih	1
2	20	50%	13,13	55,00	Manset Leher Navy , Manset Leher Abu Muda->Manset Leher Light Khaki	4
3	20	50%	13,49	56,52	Manset Leher Putih , Manset Leher Abu Muda->Manset Leher Light Khaki	2
4	20	50%	17,07	58,70	Manset Leher Hitam,Manset LeherNavy->Manset Leher Putih	6
5	20	50%	11,94	50,00	Manset Leher Hitam,Manset LeherNavy->Manset Leher Light Khaki	13
6	20	50%	13,26	55,56	Manset Leher Hitam , Manset LeherAbuMuda->Manset Leher Light Khaki	4
7	20	50%	23,85	54,35	Manset Leher Abu Muda , Manset LeherPutih->Manset Leher Navy	1
8	20	50%	16,80	57,78	Manset Leher Abu Muda , Manset Leher Light Khaki->Manset Leher Putih	2
9	20	50%	7,84	55,56	Ciput Bandana Putih,CiputBandana Sandalwood->Ciput Bandana LightKhaki	2
10	20	40%	6,76	40,30	Manset Leher Navy , Manset LeherPutih->Manset Leher Hitam	6

The highest frequency is owned by the rule "black neck cuff, navy neck cuff -> light khaki neck cuff" with a total frequency of 13 and a confidence of 50%. Frequency shows the number of sales transaction data containing items in the rule obtained from testing data. So, the rule chosen in the analysis of the minimum confidence rule for the combination of 3 items is "black neck cuff, navy neck cuff -> light khaki neck cuff". This rule will be used as one of the recommendations for making the RDA Hijab business promotion strategy.

V. CONCLUSION

The a priori algorithm can analyze sales transaction data for RDA Hijab products for 12 months, from December 1, 2019 to November 30, 2020 by finding the association rule. The RDA Hijab website is built using HTML, PHP, and CSS programming languages. There are 6 pages including the login page, the list page, the home page, the dataset page, the a priori process page, and the rule results page.

From the results of the analysis in this study, it can be concluded for the rule with a combination of 2 items that have the highest confidence value of 41.6%, the lift ratio value of 12.1, and the frequency of the rule in the highest testing data is 66, that is, if consumers buy navy neck cuffs then he will also buy white neck cuffs. And for the rule with a combination of 3 items that have the highest confidence value of 50%, the lift ratio value of 11.9, and the frequency of the rule in the highest testing data is 13, that is, if the consumer buys black neck cuffs and navy neck cuffs, he will also bought a light khaki neck cuff.

The results of the rules formed can be a recommendation for making new business promotion strategies for RDA Hijab. One of them is by making promotions in the form of giving bonuses for unsold products to consumers who buy certain product packages. From the research that has been done, it is recommended that 16 forms of product packages are recommended, one of which is to give product bonuses to consumers who buy navy neck cuffs and white neck cuffs.

The author realizes that the research carried out still has many shortcomings and limitations, therefore the author will suggest further research to creating an automatic data processing system (data preprocessing) so that the process is more efficient and the data processing process is shorter. In addition, creating a priori process with python language so that the execution process is faster.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] L. A. Dewi, Listriani, F. H. Setyaningrum, F, "Application of Association Method Using Apriori Algorithm in Consumer Shopping Pattern Analysis Application (Case Study of Bintaro Gramedia Bookstore)." Journal of Informatics Engineering, 120- 127, 2016

[2] E. Buulolo, Data Mining For Universities. Sleman: Depublish. 2020.

[3] G. Gunadi, D.I Senses, "Application of Data Mining Market Basket Analysis Method to Sales Data of Book Products Using Apriori Algorithm and Frequent Pattern Growth (FP-Growth) Case Study: PT.Gramedia Printing." Telematics MKOM Vol.4 No.1, 1, 2012.

[4] R.R Rerung, "Implementation of Data Mining by Utilizing the Association Rule Method for Product Promotion." Journal of Engineering Technology, Vol. 3, No. 1, 92, 2018.

[5] F.A Sianturi, "Application of Apriori Algorithm for Order Level Determination." Penusa Mantik Journal, 51, 2018.

[6] S. Hadi, Implementation of Data Mining with Association Rule in Decision Making for Correlation of Product Purchases Using Apriori

Algorithm (Case Study: ASR JEANS Jember). Jember: Muhammadiyah University of Jember, 2016.

[7] D. Widiastuti, N. Sofi, "Comparative Analysis of Apriori Algorithm and FP-Growth in Cooperative Transactions." UG Journal, 21, 2014.

[8] R. Umar, R.I. Putri, "Integrated E-Commerce Development With Market Basket Analysis". Proceedings of Sentrinov Vol. 001, 295, 2015.

[9] M. Badrul, "Association Algorithm With Apriori Algorithm For Sales Data Analysis." Journal of Pilar Nusa Mandiri Vol. XII, No. 2, 125, 2016.

[10] G.A Saputro, Application of Apriori Algorithm to Find Sales Patterns in Cafe Case Study: Jurney Coffee. Yogyakarta: Sanata Dharma University, 2017.

[11] K. Tampubolon, H. Saragih, B. Reza, "Implementation of Apriori Data Mining Algorithm in Medical Device Inventory System." Scientific Information and Technology Scientific Magazine (INTI), 99, 2013.

[12] D. D. Wicaksono, Deployment of Data Mining With Apriori Algorithm For Information Query Suggestions for Goods (Case Study: Site 'Inkuiri.com'). Yogyakarta, 2015.

[13] M. Fauzy, K. R. Saleh, and I. Asror, "Application of Association Rule Method Using Apriori Algorithm in Rain Prediction Simulation in Bandung City." Scientific Journal of Applied Information Technology, 225, 2016.

[14] Khairatin, Ship Selection Using Apriori Algorithm in Web- Based Ticket Reservation Application (Case Study: Tidung Island - Kali Adem). Jakarta, 2019.

Indonesian Abusive Tweet Classification based on Convolutional Neural Network and Long Short Term Memory Method

Reinaldo Yosafat Gultom
Research Center for Artificial Intelligence and Big Data
Universitas Padjadjaran
Bandung, Indonesia
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
reinaldogultom@gmail.com

Fahri Izzuddin Zulkarnaen
Research Center for Artificial Intelligence and Big Data
Universitas Padjadjaran
Bandung, Indonesia
Informatics Engineering Department
UPN Veteran Jawa Timur
Surabaya, Indonesia
Fahrizulkarnaen1@gmail.com

Yahma Nurhasanah
Research Center for Artificial Intelligence and Big Data
Universitas Padjadjaran
Bandung, Indonesia
Department of Statistics
Universitas Padjadjaran
Bandung, Indonesia
yahma2204@gmail.com

Asep Sholahuddin
Research Center for Artificial Intelligence and Big Data
Universitas Padjadjaran
Bandung, Indonesia
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
asep.sholahuddin@unpad.ac.id

***Abstract*— Twitter is one of the famous social media applications in Indonesia. The study will classify data to detect hate speech on Twitter into two classes: abusive and non-abusive. The abusive class will contain tweets in the form of words that contain negative elements or hate speech, while the non-abusive class contains tweets in the form of words that do not contain negative elements. The desired purpose of this research is to help those who need this research to be able to use it to supervise others from hate speech and find the perpetrators who do it. The data used is secondary data taken from the Kaggle website. The data contains tweets that have been classified based on connotations of abusive or non-abusive. Several processes are carried out to achieve the desired goal, including data acquisition, preprocessing, classification, evaluation, and content detection. The algorithms used in the study were LSTM and CNN. Both algorithms produce the same performance relatively. It's just that CNN is slightly superior. The accuracy value of CNN is 0.9131 or 91.31%, the precision is 0.9131 or 91.31%, the recall is 0.9131 or 91.31%, and the F1-Score of 0.9131 or 91.31%. Meanwhile, for LSTM, the accuracy value is 0.9104 or 91.04%, the precision is 0.9104 or 91.04%, the recall is 0.9104 or 91.04%, and the F1-Score is 0.9104 or 91.04%**

Keywords—Twitter, CNN, LSTM, Abusive

I. INTRODUCTION

Humans are living beings that always interact with each other. Humans as social beings cannot live alone, but desperately need the role of others. Because we live in this world, we need each other. Humans, as social beings, can never live alone. Whenever and wherever humans are, humans are always required to relate to other people.

Social media is a place for internet-based communication and a means for humans to connect with other people. Social media can be one of the media tools that allow anyone (the user) to be able to represent themselves, interact, share, cooperate, communicate with others, and form virtual social bonds. So, with social media, we can achieve many advantages, such as communicating with other people, obtaining information or news, or even as a means of entertainment. Several types of social media that are generally owned by Indonesian people include Facebook, Tik Tok, Instagram, and Twitter. Almost all Indonesian people from various circles have social media accounts, even more than one.

Twitter, which is one of the most popular social media applications in Indonesia, can be used freely, so comments or tweets on Twitter can be given either positively or negatively. Negative comments such as hate speech, racism, spreading stories, and fraud are often seen on social media, including Twitter. These things show that hate on social media is one

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

of the most used on Twitter [1]. The impact can not only harm individuals but the impact of this can also be felt by the international community. the spreaders of hate speech on social media mostly use anonymous names for their accounts to avoid knowing their true identity. In addition, they generally target accounts that have a large number of followers or accounts that have a high level of activity.

Therefore, the purpose of this research is to prove that deep learning can be used to create a healthy social media environment so that users can be comfortable using/accessing various social media on the internet. This research can be used as a basis for describing the level of awareness of internet users in Indonesia regarding the existence of hate speech on Twitter. This goal can be achieved by building a model that implements the Long Short-Term Memory (LSTM) method and the Convolutional Neural Network (CNN) method to identify hate speech. In the case of classification, LSTM can be used to make accurate predictions of various variables. The best forecasting is based on the prediction error rate, where the smaller the error generated, the more accurate a method is in predicting. While the CNN method, although generally used in digital image processing, can also be used in text processing. Thus, the CNN Model used with word vectors extracted from various features can be used to detect hate speech.

II. METHODOLOGY

Some of the steps that will be used in the research are data acquisition, data preprocessing, classification, evaluation, and then content detection, is illustrated in Fig. 1.

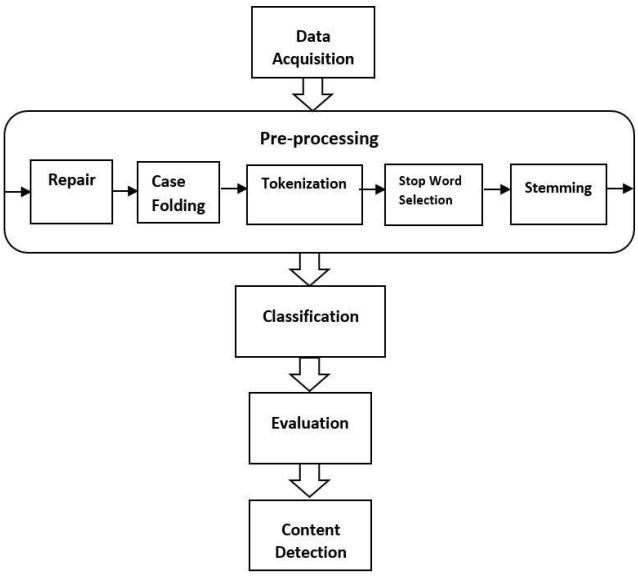


Fig. 1. Proposed Abusive Tweet Classification Model.

A. Data Acquisition

In carrying out this research, we took and processed datasets from the Kaggle website [2]. This dataset contains Twitter tweets that have been classified based on the connotation of hate speech, abusive, or not. The dataset that we take is still raw data in which it still has 13 attributes. Then we do data pre-processing so that the data we get is suitable and easily understood by the programming language later, besides that the pre-processing data is useful for removing anomalies and dataset oddities that occur.

B. Pre-processing

Pre-processing is preparing unstructured data as a result of data into data that is ready for use. This process is carried out to obtain better data quality and improve classification accuracy [3]. In the pre-processing stage, several steps must be carried out: repair, case folding, tokenization, stop word deletion, and stemming.

There are six stages of data pre-processing that we do:

1) Make the entire tweet lowercase.
2) Remove all non-alphanumeric characters from each tweet
3) Remove all unusual characters from each tweet.
4) Make the slang term more common.
5) Convert all affixed words into basic words.
6) Remove all pre-existing stop words.

Then after pre-processing the data, we have left as many as 13,169 data. With details of 5043 data containing abusive words (38.3%) and 8126 data that do not contain abusive words (61.7%).

C. Classification

The data that has previously been preprocessed in the form of strings is classified, namely as "abusive" or "non-abusive" using both algorithms, because later it will produce two models where the accuracy level will be compared. In this research, the classification of tweet data will be split according to 2 classes, namely abusive or non-abusive class. In this step, we will explain the classification process of Tweet/comment data using several algorithms for deep learning, namely Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) Methods [4].

1) Long Short Term Memory (LSTM)

LSTM is the result of a modification of the Recurrent Neural Network (RNN) after the addition of a memory cell that is useful for long-term information storage. In addition, LSTM can also overcome the vanishing gradient problem contained in the RNN when processing long sequential data by using a set of gates that are useful for controlling the information that goes into memory [5]. The hidden layer consists of memory cells. In the memory cells, there is one input gate, one forget gate, and one output gate. The input gate controls how much information should be stored in the cell state. This prevents cells from storing unnecessary data. Forget gate serves to control the extent to which values remain in the memory cells. The output gate decides how much content or value is in a memory cell that is used to calculate the output. Memory cells in LSTM are useful for storing a value or state (cell state) for long or short periods. The following is an explanation for the gates in a long-short term memory (LSTM) memory cell:

a) Input Gate (i_t)

The input gate plays a role in taking the previous output and new input, then passing them through the sigmoid layer. This gate returns a value of 0 or 1. The following is the formula of i_t:

$$i_t = \sigma(W_i S_{t-1} + W_i X_t) \quad (1)$$

W_i : Weight of the input gate.

S_{t-1} : Previous state or at t-1.

X_t : Input at t.

σ : Sigmoid activation function.

The input gate multiplied the output of the candidate layer. The following is the formula of the $\underline{C}$:

$$\underline{C} = tanh\ tanh\ (W_c S_{t-1} + W_c X_t) \quad (2)$$

$$c_t = i_t * f_t + \underline{C}_t * c_{t-1} \quad (3)$$

$\underline{C}$: Intermediate cell state.

W_c : Weight of cell state.

S_{t-1} : Previous state or at t-1

X_t : Input at t.

The previous state multiplied the forget gate and then added it to the new candidate function allowed by the output gate.

b) *Forget Gate* (f_t)

Forget gate is a sigmoid layer that takes output at t-1 and input at t. Then, it combines and implements sigmoid activation functions. The value of this gate is 0 or 1. If $f_t = 0$, then the previous state would be forgotten, while the previous state does not change. The following is the formula of the f_t:

$$f_t = \sigma(W_f S_{t-1} + W_f X_t) \quad (4)$$

W_f : Weight of the forget gate.

S_{t-1} : Previous state or at t-1

X_t : Input at t.

σ : Sigmoid activation function.

This layer applies a hyperbolic tangent to the previous input and output mixture. Then, return the candidate vector to be added to the state.

c) *Output Gate* (o_t)

Output gate plays a role in controlling how much state passes to the output and works in the same way as other gates. And finally, produce a new cell state (h_t). The following is the formula of the o_t and h_t :

$$o_t = \sigma(W_o S_{t-1} + W_o X_t \quad (5)$$

$$h_t = o_t * tanh\ tanh\ (c_t) \quad (6)$$

W_o : Weight of the output gate.

S_{t-1} : Previous state or at t-1

X_t : *Input at t.*

σ : Sigmoid activation function.

The predicted accuracy is obtained from the data that has been trained. The key to success is the number of hidden layers [6].

In this case, we used a stacked LSTM method. The stacked LSTM is a model that has multiple hidden LSTM layers where each layer contains multiple memory cells. Stacking LSTM hidden layers makes the model deeper. We use this stacked LSTM method because it is generally attributed to the success of the approach on a wide range of challenging prediction problems. [7]

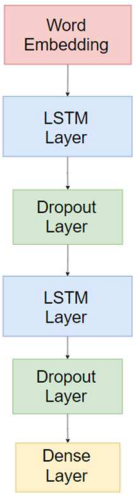


Fig. 2. Stacked LSTM Architecture

Fig. 2 presents the stacked LSTM architecture. The first stage of the model training on the LSTM method in this paper is Word Embedding. Word Embedding is useful for changing the shape of each input word into vector form. After that, we use stacked LSTM that consist of two LSTM layers. Every LSTM layer is followed by a dropout layer. Dropout layer is used to prevent overfitting. It is a neural network regularization technique where several neurons will be randomly selected and not used during training. The last stage is the dense layer which is one of the neural network models that has a function to classify according to the class in the output. A dense layer is a layer that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer.

2) Convolutional Neural Network

The Convolutional Neural Network (CNN) is the development of Multilayer Perceptron (MLP) designed to process two-dimensional data for image processing. Nevertheless, for the Natural Language Processing task that applied text data instead of images, we have a one-dimensional array that represents the text. Therefore, the architecture of the CNN is changed into 1D convolutional and pooling operations. [8] In 1D CNN, the kernel moves in 1 direction. Input and output data of 1D CNN is 2 dimensional [9]. Convolution Neural Network (CNN) involves a group of filters of different shapes and sizes which convolve the original sentence matrix to reduce it into further low dimension matrices. The following is an explanation for the process of CNN text classification.

a) Word Embedding

Word embedding is the feature learning technique where words from the vocabulary are mapped to vectors of real numbers capturing the contextual hierarchy. [10]

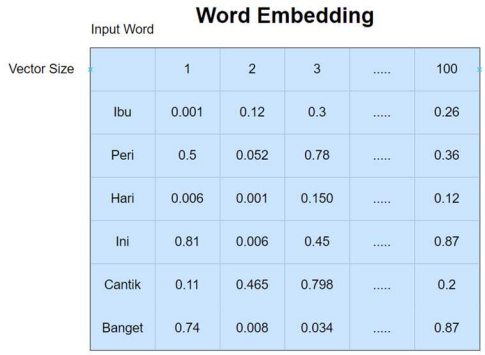


Fig. 3. The Diagram for Word Embedding

As we can see in Figure 4, each word is mapped into vectors. For doing word embedding, we use Keras library from TensorFlow. The parameters in the embedding layer are the size of the vocabulary, the number of the dimensions for each word vector, and the maximum length of the input sentence.

b) Convolutional Layer

The convolution layer uses a kernel or filter to extract objects from input. This kernel contains weights that are useful for detecting the character of the objects. Convolution will result in a linear transformation of the input data that corresponds to the spatial information on the data. The kernel is applied repeatedly resulting in a series of receptive fields. In addition, there are parameters that can be changed to modify the properties of each layer, namely filter size, stride, and padding. Stride is useful for controlling how filters are applied to input data by moving along a predetermined pixel size. Padding is useful for increasing pixel size with a certain value around the input data so that the results of the receptive field are not too small so there is not much information missing. This value is usually worth zero, so it is called zero padding. For a one-dimensional convolutional layer, the kernel slides along one dimension.

As we can see in Figure 4, the kernel slides down with one stride and makes the output from word embedding become an input. Then, the input is calculated with some filters.

c) Global Max Pooling

Because the number of parameters is too large, we need to do a reduction with adding a pooling layer. Pooling or subsampling is a reduction in the size of a matrix. In this case, we use global max pooling.

d) Dropout Layer

Dropout is usually applied to prevent overfitting and speed up the training process. During the training phase of a neural network, a fraction of the hidden units is randomly dropped at every iteration with probability p_{drop} (or the keep probability p_{keep} =1− p_{drop}). The hidden units randomly dropped during training only, while for the evaluation phase, all the hidden units must be active [8].

e) Dense Layer

The dense layer (Fully Connected layer) takes an input from the flatten layer and gives out an N-dimensional vector where N is the number of classes. The function of the fully connected layer is to use these features for classifying the input text into various classes based on the loss function on the training dataset. In this case, we use a two-dimensional vector because we will classify two classes that consist of abusive and not abusive.

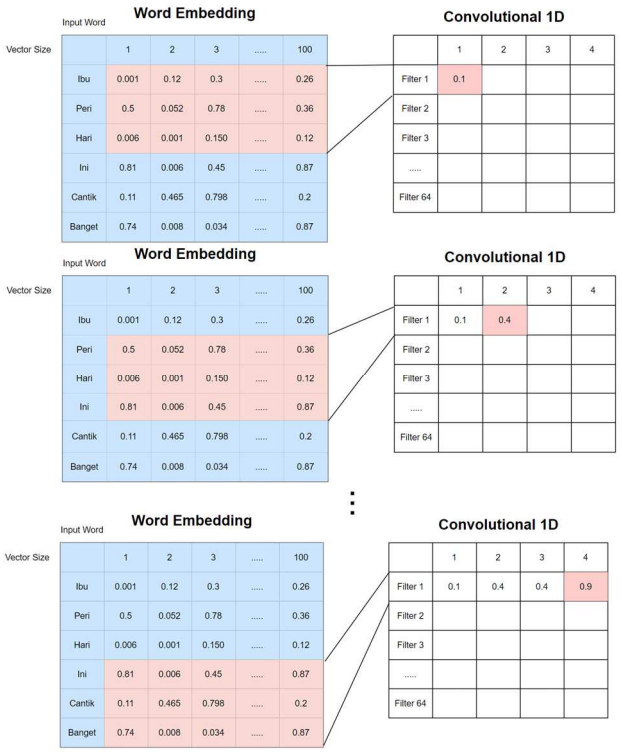


Fig. 4. The Visualization of Convolution Layer

D. Evaluation

The Confusion Matrix is a table with four different combinations of predicted and actual values. Four terms represent the results of the classification process in the confusion matrix, namely True Positive, True Negative, False Positive, and False Negative. The matrix stores information to determine the performance of the model used and is used as a reference for the classification performance of the algorithm used later.

The confusion matrix will tell us how well the model we have made works. In particular, the confusion matrix also provides information about TP, FP, TN, and FN [11]. The confusion matrix is advantageous because the results of the classification generally cannot be expressed properly in a single number. The explanation is as follows:

1) True Positive (TP): predicting a positive result when the actual result is positive.

2) True Negative (TN): predicting a positive result when the actual result is negative.

3) False Positive (FP) (Error Type 1): predicting a positive result when the actual result is negative.

4) False Negative (FN) (Error Type 2): Predicting a negative result when the actual result is positive.

To determine the performance level of a deep learning model, it can calculate accuracy, precision, recall, and F1-Score. The formula for the calculation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$Precission = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1 - Score = \frac{2 \times Precission \times Recall}{Precision + Recall} \quad (14)$$

E. Content Prediction

The next step is to detect whether the sentence is abusive or non-abusive. At this stage, the model with the best accuracy is used from the scenarios that have been made from the various models made. So, it can be interpreted, at this stage, the sentence input is carried out on the existing model and will then make predictions.

III. Results and Analysis

A. Scenario

The scenario that we have done is by pre-processing the datasets first. Then the results of pre-processed datasets that have been done can be tested by two methods, the two methods that we use are Long Short Term Memory and Convolutional Neural Networks.

B. Model Performance

Dataset that has gone through this pre-processing will be divided into two data model, namely testing data and training data. We conducted experiments with different splitting and epochs calculations to compare the best results.

We did several time testings for every each hyperparameter to get the best accuracy. For the Long Short Term Memory Algorithm we did several testing for every different Vector Size and Filters. So, the results are in Table 1.

TABLE I. Different Hyperparameter Test for LSTM

	32 Units	64 Units	128 Units
1100 Vectors	0.8576	0.9104	0.8652
1500 Vectors	0.9058	0.9103	0.8333
2000 Vectors	0.9051	0.7692	0.8937

From the data above we got the best accuracy when we train model with 1100 vectors and 64 filters size. For the Convolutional Neural Network the hyperparameter that we different test are Vector Size for word embedding and the number of filters that we use in Convolutional Layers, the results are in Table 2.

TABLE II. Different Hyperparameter Test for CNN

	32 Filters	64 Filters	128 Filters
100 Vectors	0.9116	0.9131	0.9051
200 Vectors	0.9124	0.9090	0.9086
300 Vectors	0.9070	0.9097	0.9070

From Table 2, we can know the results of a several hyperparameter testing for CNN Algorithm is we got a best accuracy when we pick 100 vectors in word embedding and 64 filters for Convolutional layer for training data. After we got hyperparameter value that has the highest Accuracy, the next step is training data. For the ratio of splitting data we do three different ratios, categorized by a splitting ratio of 80% for training data and 20% for testing data and a splitting ratio of 70% for training data and 30% for testing data, and also a splitting ratio of 60% for training data and 20% for test data. We train data in 10 epochs and 25 epochs for each splitting data ratio. We use Relu and Softmax for the activation function and Adam for Optimizer. The results that we had are in Table 3-5.

1) Splitting data 80:20

a) Results of Splitting data 80:20 ratio:

TABLE III. Splitting data 80:20

Algorithm	Accuracy	Precision	Recall	F1-Score
10 Epochs				
CNN	**0.9131**	**0.9131**	**0.9131**	**0.9131**
LSTM	0.9104	0.9104	0.9104	0.9104
25 Epochs				
CNN	**0.8899**	**0.8899**	**0.8899**	**0.8899**
LSTM	0.4085	0.4081	0.4077	0.4079

2) Splitting data 70:30:

a) Results of Splitting data 70:30 ratio:

TABLE IV. Splitting data 70:30 Ratio

Algorithm	Accuracy	Precision	Recall	F1-Score
10 Epochs				
CNN	**0.9035**	**0.9035**	**0.9035**	**0.9035**
LSTM	0.8924	0.8924	0.8924	0.8924
25 Epochs				
CNN	**0.9027**	**0.9027**	**0.9027**	**0.9027**
LSTM	0.5270	0.5270	0.5270	0.5270

3) Splitting data 60:40:

a) Results of Splitting data 60:40 ratio

TABLE V. Splitting data 60:40 Ratio

Algorithm	Accuracy	Precision	Recall	F1-Score
10 Epochs				
CNN	**0.9025**	**0.9025**	**0.9025**	**0.9025**
LSTM	0.8724	0.8724	0.8724	0.8724
25 Epochs				
CNN	**0.8943**	**0.8943**	**0.8943**	**0.8943**
LSTM	0.8512	0.8512	0.8512	0.8512

We decided to do different training with many epochs and data splitting ratio variables to get the best hyperparameter

for each method. For the dataset that we used, some conclusions that we got are:

- By using the LSTM method, the best performances that we got are when training with a splitting ratio of 80:20 and 10 epochs. The value of accuracy, precision, recall, and F1-Score values are 0.9131 or 91.31%.
- For the CNN method, we got the maximum results when training with a splitting ratio of 80:20 and 10 epochs. The result is accuracy, precision, recall, and F1-Score values are 0.9104 or 91.04%.

IV. Conclusion and Discussion

A. Conclusion

Based on the results above, our research has succeeded in identifying tweets containing abusive words on Twitter social media. After testing and evaluating the analysis of tweets that contain abusive words, there are several points that we can conclude:

- The difference in the number of datasets in the distribution of training and testing data greatly affects the final results of the prediction accuracy because the quality and number of datasets that we use greatly affect the results obtained.
- For the different hyperparameter training in Convolutional Neural Network and Long Short Term Memory also greatly affect the results that we got.
- Based on differences in the results achieved by using two different methods, namely the LSTM and CNN methods. For the dataset that we use, it is proven that the CNN method can perform better prediction results with an accuracy value of 0.9131, precision 0.9131, recall 0.9131, and F1-Score of 0.9131.
- Although the results of this LSTM method are slightly lower than CNN with accuracy, precision, recall, and F1-Score values are 0.9104 or 91.04%. However, the LSTM method cannot be underestimated for testing our dataset.

B. Discussion

We hope that the results of our research can be used by Social Media Developer to detect some abusive word content. An example of its application is when a user uploading an abusive words content so it can be deleted automatically by the system.

Acknowledgment

On this occasion, the Author would like to thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This paper also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

Reference

[1] F. D Vigna, et al, "Hate me, hate me not: Hate speech detection on facebook", First Italian Conference on Cybersecurity (ITASEC17) (pp. 86-95), 2017

[2] I.F Putra, "Indonesian Abusive and Hate Speech Twitter Text", kaggle.com, https://www.kaggle.com/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text, 2020. (accessed Sept, 11, 2021)

[3] W. Jenq-Haur, An LSTM Approach to Short Text Sentiment Classification with Word Embeddings, 2018 pp 2-6.

[4] A. Hassan, A. Mahmood, Convolutional Recurrent Deep Learning Model for Sentence Classification, IEEE Access, 6, pp. 13949-13957, 2018.

[5] L. Deng,and D. Yu, "Deep learning: methods and applications" Foundations and trends in signal processing, 7(3–4), 197-387, 2014

[6] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[7] J. Brownlee, Long Short-Term Memory Networks With Python: Develop Sequence Prediction Models With Deep Learning, Pp. 78-79, 2017.

[8] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network", 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6), 2017

[9] U. R. Acharya, et al, "A deep convolutional neural network model to classify heartbeats", Computers in biology and medicine, 89, 389-396, 2017

[10] A. Kulkarni, A. Shivananda, Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. pp. 82-83, 2019

[11] S. Raschka, V. Mirjalili, Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. 2017. pp. 512-513.

Prediction New Cases of COVID-19 in Indonesia Using Vector Autoregression (VAR) and Long-Short Term Memory (LSTM) Methods

Daffa Rahmansyah Danistya[ab1], Fahira Qaulifa[ac2], Yovi Adhi Ramadani[ad3], Intan Nurma Yulita[ae4], Mulya Nurmansyah Ardisasmita[f5], Dwi Agustian[f6]
[a]*Research Center for Artificial Intelligence and Big Data Padjadjaran University*
Bandung, Indonesia
[b]*Department of Electrical Engineering Brawijaya University*
Malang, Indonesia
[c]*Department of Electrical Engineering Telkom University*
Bandung, Indonesia
[d]*Department of Informatics Engineering, Semarang University*
Semarang, Indonesia
[e]*Department of Computer Science Padjadjaran University*
Bandung, Indonesia
[f]*Department of Public Health, Padjadjaran University*
Bandung, Indonesia
e-mail: [1]daffarahmansyahx1@gmail.com, [2]vahira39@gmail.com, [3]yoviadhi48@gmail.com, [4]intan.nurma@unpad.ac.id, [5]mulya@unpad.ac.id, [6]dwi.agustian@unpad.ac.id

***Abstract*—The addition of Covid-19 cases is still uncontrolled, especially in Indonesia. Often the addition of Covid-19 cases in Indonesia always experiences a significant upward trend after a slightly loose government policy. This is because the government does not think there will be a spike in cases after cases go down. This is where the importance of predicting new cases of Covid-19 in Indonesia to be a reference for the government in taking policy. With deep learning, the prediction results will be more accurate. The implementation of vector autoregression (VAR) and long-short term memory (LSTM) methods can reach an accretion rate of up to 98%. With this method, the prediction results can be used for the government in anticipating if there is a surge in new cases per day because it has been predicted from the beginning. In fact, this method can predict new cases for up to a year.**

Keyword—New Cases, Prediction, Artificial Intelligence, Vector Autoregression (VAR), Long-Short Term Memory (LSTM).

I. INTRODUCTION

Already almost 2 years Coronavirus (COVID-19) haunts the entire population of the world, especially Indonesia [1]. Since it was first detected in Wuhan in early December 2019, COVID-19 has killed 4.55 million people worldwide. In Indonesia alone, as many as 141 thousand people have died because of this virus. Indonesia first confirmed the Covid-19 case on Monday, March 2, 2020. At that time, President Joko Widodo (Jokowi) announced that there were two Indonesians who tested positive for the coronavirus, namely a 31-year-old woman and a 64-year-old mother. Based on the Covid distribution map uploaded on the official website of Covid-19 Indonesia, since the first positive case was confirmed in Indonesia, other positive cases increased until the end of May 2021 but at the beginning of June 2021the us cash confirmed positive increasingly experienced an increase and the peak of the increase in cases of Covid-19 occurred on July 15, 2021 where positive confirmed cases reached 56 thousand confirmed cases. The data shows how dangerous and acute the Covid-19 virus is [2].

With the increasing number of cases of Covid-19 in Indonesia that eat a lot of souls, people are made restless, worried, and wondering when the Covid-19 virus ends up hitting Indonesia. While the only info about Covid-19 that people get is only through the official website of Covid-19 Indonesia, although the website has provided information about Covid-19 cases in Indonesia, the People of Indonesia often want to know how the Covid-19 case in the future. For that in This research then we want to help the community and interested parties to predict the trend of Covid in the next year with the use of deep learning.

With this research, it is hoped that this prediction can be used to prepare if there is an increase in new cases per day which can be anticipated by adding health workers or health facilities. In addition, other preparations such as acceleration of vaccination and preparation of self-isolation places so that no hospitals are full or health workers fall due to fatigue.

II. DEEP LEARNING

Deep learning is part of artificial intelligence and machine learning [3], which is the development of multiple layer neural networks to provide precision tasks such as object detection, voice recognition, language translation, and others. Deep learning is based on traditional machine learning techniques [4], because deep learning automatically performs representations of data such as images, videos, or text without introducing code rules or human domain knowledge [5]. Deep learning was developed in 1950. However, only in 1990 can be applied successfully. The learning algorithms used now on complex tasks are almost the same as the learning algorithms used to solve game problems in the 1980s, although the algorithm mode used was changed to a simple training of deep learning architecture. The important thing in the development of the current model is that we can support with the resources needed to be successful. The development of an increasing data set leads to a centralized data set that facilitates in its management [6].

Deep learning has a variety of benefits that can bring benefits to other technologies. The benefits of deep learning

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

technology include maximizing the unstructured performance of data in applications or websites, eliminating the need for technology for feature engineering, providing a much more quality output display, reducing operational development costs, and effective data manipulation techniques [6].

In addition, there are several types of deep learning algorithms. One of them is long-short term memory (LSTM). To predict new cases of Covid-19 in Indonesia, more effectively we combine this type of long-short term memory (LSTM) algorithm with vector autoregression (VAR).

III. VECTOR AUTOREGRESSION AND LONG-SHORT TERM MEMORY FOR PREDICTING NEW CASES OF COVID-19 IN INDONESIA

The structure of the deep learning layer is used to analyze objects and phenomena so that multi-level characteristics are extracted by the learning object. Layer structures allow computers to learn complex concepts with simple structures. Due to its strong data reading capabilities and the great computing power of high-dimensional data, deep learning can achieve strong representational capabilities with different types of data [7]. To predict the addition of new cases of Covid-19 in Indonesia using vector autoregression and long-short term memory methods consisting of three parts namely data set, vector autoregression, and long-short term memory.

A. Data Set

We use data from kawalcovid19 which is a task force to accelerate the handling of Covid-19 in Indonesia. Kawalcovid19 has been collecting data on Covid-19 cases in Indonesia since the beginning of the entry of Covid-19 into Indonesia until now. But all we use is data from the beginning of the entry of Covid-19 into Indonesia on January 8, 2020 to July 9, 2021.

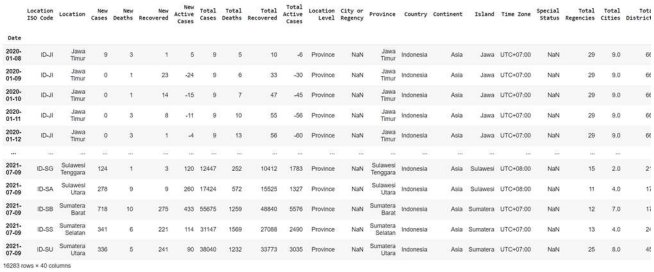


Fig. 1. Data Set Covid-19 in Indonesia.

TABLE I. DATA MODELING AVERAGE FROM ALL PROVINCES IN INDONESIA

Date	New Cases	New Deaths	New Recovered
2020-01-08	9.000000	3.000000	1.000000
2020-01-09	0.000000	1.000000	23.000000
2020-01-10	0.000000	1.000000	14.000000
2020-01-11	0.000000	3.000000	8.000000
2020-01-12	0.000000	3.000000	1.000000
…	…	…	…
2021-07-05	1747.470588	33.529412	842.411765
2021-07-06	1833.558824	42.176471	933.823529
2021-07-07	2019.735294	57.704882	865.470588
2021-07-08	2257.088235	50.352941	1241.205882
2021-07-09	2242.264706	50.735294	1694.382353

The data set we took was shown in Fig. 1. We model new cases, new deaths and new cures per day. Modeling of new cases, new deaths and new cures from the Covid-19 data set in Indonesia is shown in Fig. 2. The modeling represents the average of cases in all provinces in Indonesia. Modeling of data on table I can be described in graph form. For graphic depictions of new cases, new deaths and new recovered are found in Fig. 2.

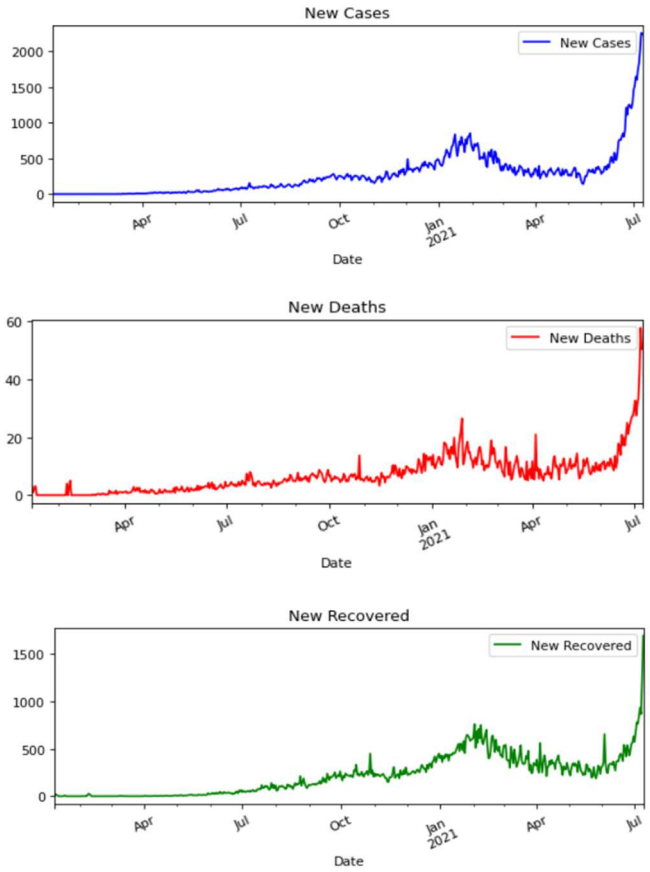


Fig. 2. Graphs of New Cases, New Deaths and New Recovered.

B. Vector Autoregression (VAR)

VAR is a linear statistical model that connects multiple data between several time series. VAR is particularly useful in modelling temporal and spatial correlations between new cases, new deaths and new recovered. Each variable has an equation that describes changes based on the interval of time per day [8].

However, models need to access complex structures that cause risk to parameters when using data sets with small dimensions. By arena, the VAR 1 model as represented in equation 1.

$$z_t = \emptyset_1 z_{t-1} + \varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

According to this, $\boldsymbol{z}_t$ is the vector of observations at t-time and n-th is the location of size $(n \times 1)$. Matrix parameters ϕ_p VAR p-order of size$(n \times n)$. ε_t explains the white noise vector $\varepsilon_t \sim MN(0, \Sigma)$ with size $(n \times 1)$. The identification step for determining the order of VAR model can be seen from the

Multivariate Partial Autocorrelation Function (MPACF) plot. The VAR (1) model has a Partial Autocorrelation Function cut-off pattern, after the 1st lag, while the stationarity could be seen from the characteristic root value of the matrix parameter [9].

VAR can process the correlation between new cases with new deaths, new cases with new cures, and new deaths with new cures. That way, the use of variables to predict new cases of Covid-19 not only uses new case variables but new death and cure trends also affect the trend of new cases that occur in Indonesia. So, the non-factor Technicalities such as vaccination and community compliance can be parameters denatured predicting new cases of Covid-19 in the future. The results of VAR modeling are a matrix of residual correlations between new cases, new deaths and newly recovered as shown in Table II.

TABLE II. CORRELATION MATRIX OF RESIDUALS

	New Cases	New Deaths	New Recovered
New Cases	1.000000	0.189992	0.227248
New Deaths	0.189992	1.000000	0.458126
New Recovered	0.227248	0.458126	1.000000

After that, the matrix is used for training. Training or training results data can be a table or a new case graph. A graph of the new case after training is shown in Fig. 3.

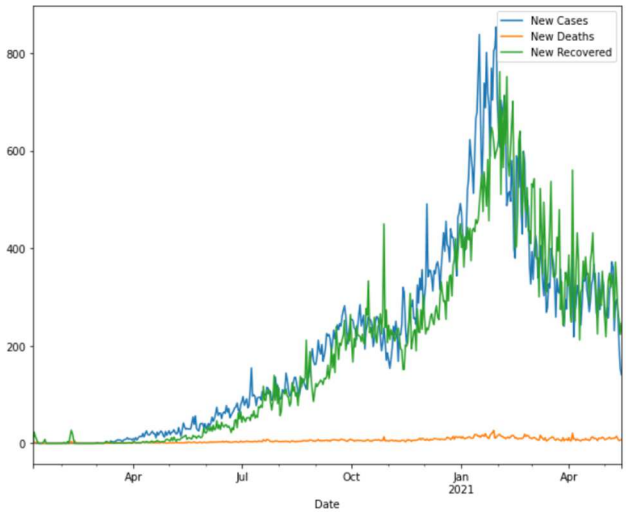


Fig. 3. Graphs of New Cases After Training Using VAR Methods.

C. *Long-Short Term Memory*

Long-short term memory (LSTM) is the most widely used recurrent neural network architecture in everyday life to address gradient loss problems. An LSTM is an interconnected computing unit [10], not only with the sound state h, but the state of cell c also acts as storage [11]. The transition created by constant gain transfer is equal to 1, so the error causes LSTM to perform previous steps (up to 1000 steps in the past) without eliminating gradients.

We use LSTM to conduct data training that has been processed with VAR to determine the prediction of new cases of Covid-19 in Indonesia. Using the VAR method combined with LSTM can improve prediction accuracy by up to 98%. A training of VAR will be incorporated into LSTM modelling so that there are parameters used for training as shown in table III.

TABLE III. SEQUENTIAL MODELING FOR LSTM TRAINING

Layer (type)	Output Shape	Param #
lstm_4 (LSTM)	(None, 14, 64)	17408
lstm_5 (LSTM)	(None, 32)	12416
dropout_2 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 1)	33

The sequential modeling above and training with 100 epochs resulted in training loss and validation loss shown in Fig. 4.

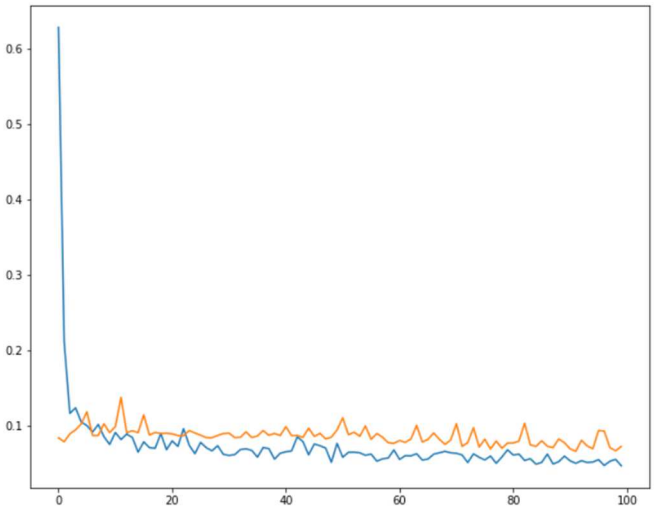


Fig. 4. Training Loss and Validation Loss After LSTM Training.

After that, data from the training can be used to predict new cases of Covid-19 in Indonesia.

IV. RESULT AND ANALYSIS

In predicting Covid-19 cases in the future, we do not know that later cases will rise significantly or have subsided, even there is no such virus. By utilizing artificial intelligence (AI) it is expected that we as humans can predict Covid-19 cases in the future, namely until 2022.

By using vector autoregression (VAR) and long-short term memory (LSTM) methods or algorithms, the addition of Covid-19 cases in the future until 2022 can be predicted. On Fig 5, it appears that at the beginning of 2022 experienced a continuous increase in intensive. And the rise in such cases reached its peak between the months and months. 3 to the 4th month. The increase in cases almost reached the 900 mark. Between the 4th month and the 6th month, the case decreased yang can be said to be quite decent. At the beginning of the 7th month, the case had gone up even slightly. But in the end, the case went down again.

Thus, the case of Covid-19 in Indonesia can be predicted using vector autoregression (VAR) and long-short term memory (LSTM) methods or algorithms. Even the method can be said to be suitable and accurate to predict a case or case.

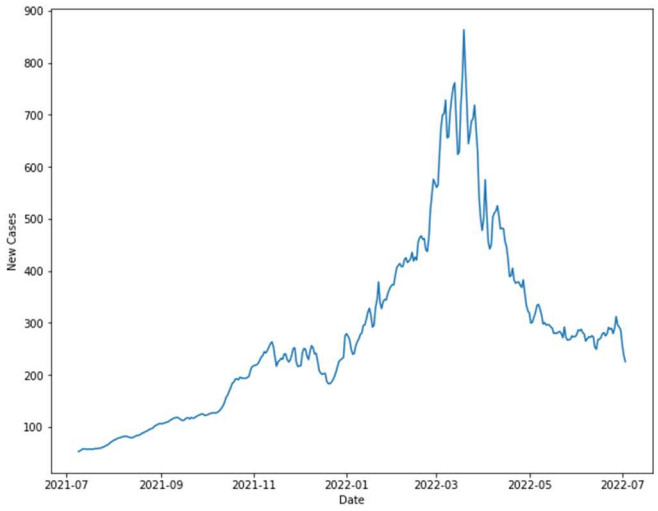


Fig. 5. Prediction Results of New Covid-19 Cases in Indonesia Using the VAR and LSTM Methods.

V. CONCLUSIONS AND SUGGESTIONS

From the above research, conclusions and suggestions are also obtained.

A. Conclusions

Based on the results of processing Covid-19 data sets in Indonesia with VAR and LSTM methods it can be concluded that the prediction of new cases of Covid-19 in Indonesia will experience ups and downs with the highest case peak between the 3rd month of the 4th month which is about 900 new cases of Covid-19. Data processing sets are used by VAR and LSTM methods that improve the accuracy of AI use. VAR and LSTM methods have an accuracy rate of up to 98%. This, exceeds the level of accuracy of the VAR method itself even the LSTM itself. With a high level of accuracy, the results of processing this data set can be used as a reference by the government in taking policy. So, it is expected that there will be no addition of new cases of Covid-19 in Indonesia again.

B. Suggestions

From the above research, VAR and LSTM methods have an accuracy rate of 98%. However, there are still methods that have a higher degree of accuracy than combining such methods. Hopefully, methods that have the level of accuracy of combining VAR and LSTM methods will soon be discovered. Thus, data processing sets using AI can achieve an accuracy rate of up to close to 100%.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[2] G. Eason, B. Noble, and I. N. Sneddon. On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[3] L. Deng,and D. Yu, "Deep learning: methods and applications" Foundations and trends in signal processing, 7(3–4), 197-387, 2014

[4] A. Hassan, A. Mahmood, Convolutional Recurrent Deep Learning Model for Sentence Classification, IEEE Access, 6, pp. 13949-13957, 2018.

[5] K. Yu, L. Jia, Y. Chen, and W. Xu, "Deep learning: yesterday, today, and tomorrow", Journal of computer Research and Development, 50(9), 1799, 2013.

[6] Y. Guo, et al "Deep learning for visual understanding: A review", Neurocomputing, 187, 27-48, 2016.

[7] Liu, T., Wu, T., Wang, M., Fu, M., Kang, J., & Zhang, H. 2018. Recurrent neural networks based on LSTM for predicting geomagnetic field. In 2018 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES) (pp. 1-5). IEEE.

[8] Y. Liu, M. C. Roberts, and R. Sioshansi, "A vector autoregression weather model for electricity supply and demand modeling", Journal of Modern Power Systems and Clean Energy, 6(4), 763-776, 2018

[9] R. E. Caraka, et al,"Hybrid Vector Autoregression Feedforward Neural Network with Genetic Algorithm Model for Forecasting Space-Time Pollution Data", Indonesian Journal of Science and Technology, 6(1), 243-266, 2021

[10]H. Bouhamed, "Covid-19 cases and recovery previsions with deep learning nested sequence prediction models with long short-term memory (LSTM) architecture", Int. J. Sci. Res. in Computer Science and Engineering Vol, 8(2), 2020.

[11]I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

Application of Bidirectional Gated Recurrent Unit (BiGRU) in Sentiment Analysis of Tokopedia Application Users

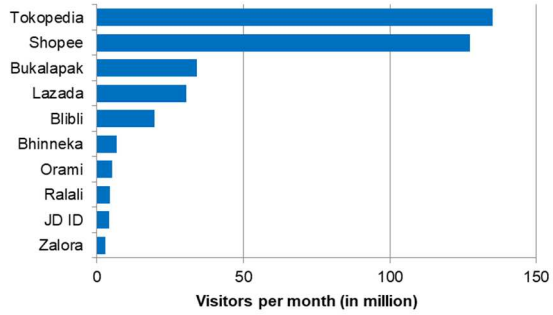
Dimas Ananda[ab1], Teguh Ammar Taqiyyuddin[a2], IIyen Nugraha Faqih[a3], Raihan Badrahadipura[ab4], Anindya Apriliyanti Pravitasari[ab5]
[a]*Department of Statistics, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Bandung, Indonesia
e-mail: [1]dimas18001@mail.unpad.ac.id, [2]teguh18001@mail.unpad.ac.id, [3]iyen18001@mail.unpad.ac.id, [4]raihan180021@mail.unpad.ac.id, [5]anindya.apriliyanti@unpad.ac.id

***Abstract*— Customer satisfaction is an important thing in business. The aim is to develop future business strategies. Tokopedia is one of the largest marketplaces in Indonesia and has become the fastest growing online company since its official launch. This study conducted a sentiment analysis of the Tokopedia application on the Google Play Store by classifying reviews into three groups, namely Positive, Negative and Neutral. The method chosen is Bidirectional gated recurrent unit (BiGRU), because BiGRU has the best accuracy rate compared to other methods. BiGRU is able to predict customer reviews with an accuracy 74.79%, recall 74.80%, and F1-Score 70.88%, which is considered good enough and can be used to help the development of Tokopedia.**

Keywords*— *Bidirectional gated recurrent unit, Tokopedia, Sentiment analysis, Reviews

I. INTRODUCTION

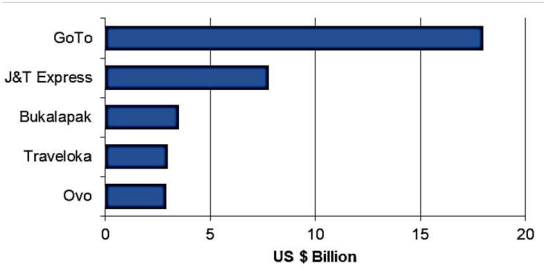
E-commerce is a set of many recent technological advances, specifically in the transformation of business processes and the transaction of goods and services [1]. Tokopedia is one of the most popular e-commerce sites in Indonesia (see Fig. 1.). This can be proven from the highest number of Tokopedia visitors in the first quarter of 2021 [2].



Source: https://databoks.katadata.co.id/

Fig. 1. Number of E-commerce Visitors Q1-2021

As of August 2021, Tokopedia had over 50 million downloads on the Play Store and 4.7 million reviews. Tokopedia is one of Indonesia's major digital-based buying and selling platforms. Tokopedia has grown to become one of Indonesia's fastest-growing online firms since its official launch. Tokopedia lets any person (small company) or well-known brand create and operate an online store by using an online marketplace and mall business model. Tokopedia's value has risen to the top in Indonesia since its merger with GoJek. [3] which can be seen in Fig. 2.



Source: https://databoks.katadata.co.id/

Fig. 2. Unicorn Valuation Value

WOM is a useful marketing resource for both consumers and marketers, as well as a dependable and effective statistic for evaluating customer loyalty with significant consequences for a product's performance. Word-of-mouth marketing is presently carried out both online and in person. According to the Ehrenberg-Bass Institute of Marketing Science, businesses must generate word of mouth outside of their existing group in order to expand. [4].

Customers who have previously purchased items leave online product reviews, which are the primary source of information for consumers and marketers on the quality of the products offered. The tendency of consumers to suggest items to others, known as "reference value," is the most significant indicator of success in business today, according to (Reichheld). [5].

Sentiment analysis is the process of automatically extracting, interpreting, and processing data in the form of unstructured text in order to extract sentiment information from a sentence of opinion or an opinion statement [6]. Sentiment analysis of a text means establishing a measure of how positive, neutral or negative the text is. Sentiment analysis is used in this study to classify the sentiments of users of the Tokopedia application.

Several studies apply algorithms in deep learning for sentiment analysis, including Abid et. A. [7] who employed the Bidirectional Gated Recurrent Unit (BiGRU) for Multi-source social media data sentiment analysis, Liu and Qi [8] use the multilayer BiGRU as a part of ResGCNN for

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

analyzing the text sentiment analysis. There are a lot of algorithms in Natural Language Processing for sentiment analysis, i.e. Long Short Term Memory (LSTM) [9] and Gated Recurrent Unit (GRU) [10]. However, in sentiment analysis, the bidirectional network structures are better in learning the text context information [7]. Considering this information, this study choose the BiGRU as the algorithm for sentiment analysis of Tokopedia application users. In addition it will also compare with LSTM, BiLSTM, and GRU based on several metrics of evaluation.

II. METHODOLOGY

The sentiment analysis for Tokopedia application users review is run under several steps, including the initialization, learning, and evaluation step. The general step is visualized by Fig. 3.

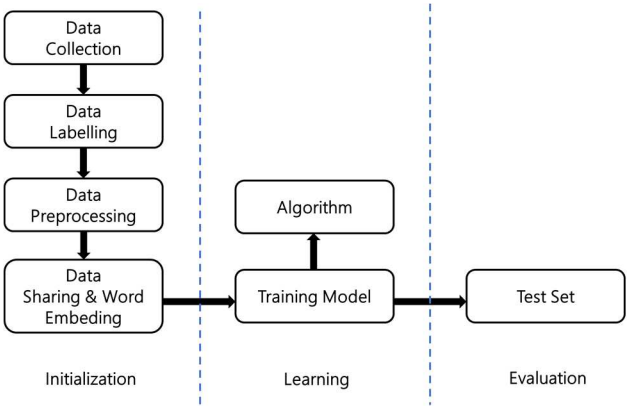


Fig. 3. General step of sentiment analysis

In particularly, the data collection is done by scrapping the reviews from Google Play Store, while the preprocessing is the sequence of Case Folding, Tokenizing, Filtering, Stemming task. The data then divided into two groups for training model and for testing and evaluation. BiGRU as the algorithm and some literatures are explained bellow.

A. Sentiment Analysis

Sentiment analysis (also known as opinion mining or emotional Artificial Intelligent) is the systematic identification, extraction, measurement, and investigation of emotional states and subjective information using natural language processing, text analysis, computational linguistics, and biometrics in the process. For applications ranging from marketing to customer service to clinical medicine, sentiment analysis is often applied in customer vocal materials such as survey evaluations and responses, online and social media, and health materials typically applied by agencies or organizations. [11].

B. Gated Recurrent Unit

Kyunghyun Cho et al. [10] presented the Gated Recurrent Unit (GRU) as a gate mechanism in repeating neural networks in 2014. Similar to long-term memory (LSTMs) with forget gates, the only difference between them and GRUs are that they lack output gates and possess fewer parameters. Certain tasks such as polyphonic music modelling and natural language processing have shown the performance of LSTMs and GRUs to be equal. On some smaller and less frequent data sets, GRUs have been found to perform better [10].

C. Bidirectional Gated Recurrent Unit (BiGRU)

Bidirectional GRU, sometimes known as BiGRU, is a two-GRU sequence processing paradigm. One accepts input in one way, while the other takes input in the other direction. As illustrated in Fig. 4., it is a bidirectional iterative neural network with just input gates and forgetting gates [12].

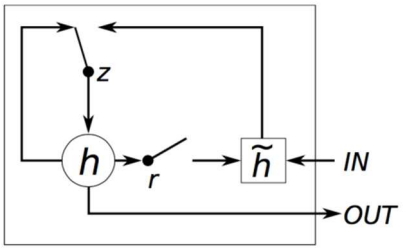


Fig. 4. Bidirectional Gated Recurrent Unit (BiGRU) Diagram

D. Research Sources and Variables

The data used in this research is secondary data. The data is taken from using the Scraping process in the Tokopedia review column on the Google Play Store site. The population in this study is all reviews about Tokopedia on the Google Play Store. The sample used is the data taken for eight months, namely January to August 2021.

The variables used are two attributes and one label. The attributes used include reviews and ratings. The labels used are labeling data from sentiment analysis. The types of label data are categorical, namely positive, neutral, and negative.

III. RESULT AND DISCUSSION

The following are the results of the research following the above methodology, namely sentiment analysis modeling using a Bidirectional Gated Recurrent Unit.

A. Scraping Data

The results of Scraping Tokopedia review data on the Google Play Store site from January to August 2021 obtained Tokopedia reviews of 32997 records.

B. Data Labeling

There are 3 labels used, namely "Negative", "Neutral", and "Positive". A “Negative” label is a review that contains bad words, ridicule, or contraries. The "Neutral" label is a review that has words that mean information, and does not side with the pros or cons. The "Positive" label is review data containing words of kindness, praise, agreement, or support. Labeling was done by the author and then justified by linguists against 32997 review data. The number of labeling for each class can be seen in Table I.

TABLE I. TOTAL SENTIMENT CLASS LABELING

Sentiment	Tokopedia Reviews
Negative	19768
Neutral	9805
Positive	3424

C. Preprocessing Data

The results of preprocessing the Tokopedia review data can be seen in Table II.

TABLE II. REVIEW DATA PREPROCESSING RESULTS

Raw Data and Case Folding Results	
Pengiriman barang di Tokopedia sangat CEPAT sekali, saya jadi senang!	pengiriman barang di tokopedia sangat cepat sekali, saya jadi senang!
Case Folding and Tokenizing Results	
pengiriman barang di tokopedia sangat cepat sekali, saya jadi senang!	pengiriman, barang, di, tokopedia, sangat, cepat, sekali, saya, jadi, senang!
Tokenizing and Filtering Results	
pengiriman, barang, di, tokopedia, sangat, cepat, sekali,, saya, jadi, senang!	pengiriman, barang, tokopedia, sangat, cepat, sekali, saya, senang
Filtering and Stemming Results	
pengiriman, barang, di, tokopedia, sangat, cepat, sekali, saya, jadi, senang	kirim, barang, tokopedia, sangat, cepat, sekali, saya, senang

D. Split Data

The data in this study will be grouped into training data and testing data. Training data is used to create and train the model to be used. Data testing is used to test the model that has been made. The distribution of training data and testing data in this study is 70:30, so 70% of the data from each class is used for training while the remaining 30% is used for testing.

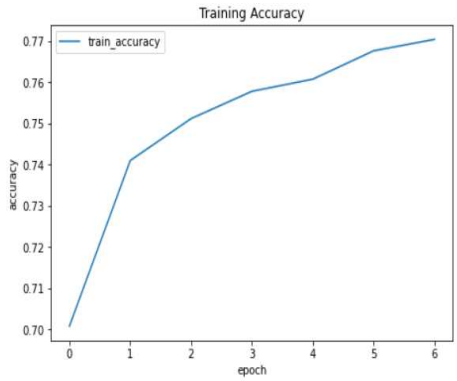
E. Word Embedding

Mapping in a sentence based on its words can be done so that it can be understood by the machine. The mapping can be done by forming a vector containing numeric values. Word Embedding is the name of the method used to carry out the mapping to retrieve learning features in Natural Language Processing (NLP). Each word in the sentence will be replaced by the corresponding numeric value in the corpus and transferred in the vector. Word2Vec in this study uses a library called genism which is available in the Python programming language. This model training process uses Skip-Gram which is an algorithm to predict the context of words by looking at the proximity of each word and its use in sentences.

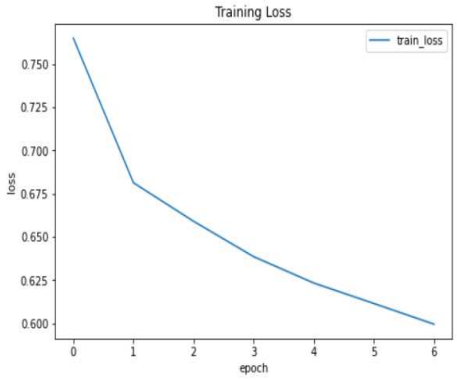
F. BiGRU Modeling

The modeling is done using BiGRU with the last layer in the form of a softmax layer with SoftMax activation function with dense 3 to classify between 3 classes, namely negative, neutral and positive. The model training is carried out with 7 epochs and 32 batch sizes. The loss used is categorical crossentropy, the optimizer uses Adaptive Moment Estimation or Adam, and uses accuracy as the metric.

Furthermore, it is obtained that training accuracy continues to increase according to Fig. 5(a) and Fig. 5(b) shows that the training loss decreases constantly. This shows that the model is optimally trained so that it continues to improve at each epoch even though there are slight fluctuations.



(a)



(b)

Fig. 5. Line chart of Training Accuracy (a) and Training Loss (b)

G. Model Evaluation

The evaluation of the model uses a confusion matrix which is a method for assessing the quality of the model. With the values obtained from the confusion matrix, the values for accuracy, precision, recall, f1-score are obtained. The results of the model predictions can be found in Fig. 6.

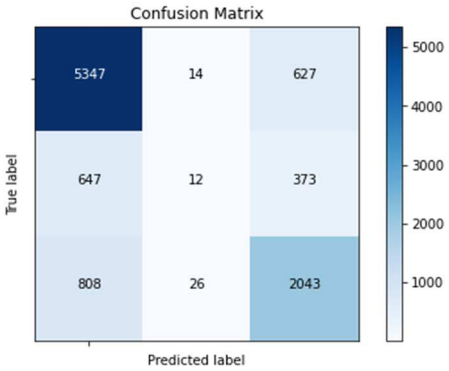


Fig. 6. Confusion Matrix

From the confusion matrix, another evaluation value can be generated. Table III is the metric of evaluation for BiGRU, BiLSTM, LSTM, and GRU. In Table III, we can see that the BiGRU is way better in predicting the sentiment reviews for Tokopedia application.

TABLE III. EVALUATION RESULTS USING BIDIRECTIONAL GATED RECURRENT UNIT (BIGRU)

Results	BiGRU	GRU	LSTM	BiLSTM
Accuracy	0.7479	0.7393	0.7405	0.7477
Precision	0.6948	0.6951	0.6988	0.6671
Recall	0.7480	0.7390	0.7410	0.7477
F1-Score	0.7088	0.7085	0.7053	0.7042

From Tabel III based on the accuracy, precision, recall, and F1-Score, the Bi-GRU has the highest value on accuracy, recall, and F1-Score. This is enough to say that this algorithm is powerfull and can be an alternative to sentiment analysis for Tokopedia.

IV. CONCLUSION

A. Conclusions

The research was conducted to analyze sentiment for review data from users of the Tokopedia application, which amounted to 32,996 with 19,768 negative class data reviews, 9,805 neutral data, and 3,424 positive data. This study performs data preprocessing processes including case folding, noise removal, stopword removal and tokenization. The Bidirectional Gated Recurrent Unit method is implemented to conduct sentiment analysis in the case study of this research, namely the public opinion of the Tokopedia application.

The results of sentiment classification using the Bidirectional Gated Recurrent Unit (BiGRU) 15method with 70:30 data sharing obtained an accuracy level of 75%. In addition, precision is 70%, recall is 75% and F1-Score is 71%. This value is higher than the other algorithm such as LSTM, BiLSTM, and GRU. It shows the predictive ability using the Bidirectional Gated Recurrent Unit (BiGRU) is better and can be used as a reference method for processing text data for Tokopedia review.

B. Recommendations

In future research, it is recommended to use a more balanced dataset between positive, neutral and negative classes so as to produce better accuracy. Also it is recommended to use a different preprocessing method like Glove or FastText.

ACKNOWLEDGMENT

The Authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] E. Turban, J. Lee, D. King, and H. M. Chung, "Electronic commerce: A managerial perspective New Jersey." Prentice Hall, 2000.

[2] D. H. Jayani, "Jumlah Pengunjung Tokopedia Kalahkan Shopee pada Kuartal I-2021," 2021, [Online]. Available: https://databoks.katadata.co.id/datapublish/2021/06/11/jumlah-pengunjung-tokopedia-kalahkan-shopee-pada-kuartal-i-2021.

[3] D. H. Jayani, "GoTo Miliki Valuasi Tertinggi di Indonesia," 2021. https://databoks.katadata.co.id/datapublish/2021/05/27/goto-miliki-valuasi-tertinggi-di-indonesia#.

[4] N. Hu, L. Liu, and J. J. Zhang, "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Inf. Technol. Manag.*, vol. 9, no. 3, pp. 201–214, 2008.

[5] F. F. Reichheld, "The one number you need to grow," *Harv. Bus. Rev.*, vol. 81, no. 12, pp. 46–55, 2003.

[6] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*, vol. 2. CRC Press, 2010.

[7] Abid, F., Li, C., & Alam, M. (2020). Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks. Computer Communications, 157, 102-115.

[8] Liu, C., & Qi, J. (2019, November). Text Sentiment Analysis Based on ResGCNN. In 2019 Chinese Automation Congress (CAC) (pp. 1604-1608). IEEE.

[9] Zhao, J., Zeng, D., Xiao, Y., Che, L., & Wang, M. (2020). User personality prediction based on topic preference and sentiment analysis using LSTM model. Pattern Recognition Letters, 138, 397-402.

[10] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734, Jun. 2014, Accessed: Sep. 09, 2021. [Online]. Available: https://arxiv.org/abs/1406.1078v3.

[11] V. A. Ho *et al.*, "Emotion Recognition for Vietnamese Social Media Text," *Commun. Comput. Inf. Sci.*, vol. 1215 CCIS, pp. 319–333, Nov. 2019, Accessed: Sep. 09, 2021. [Online]. Available: https://arxiv.org/abs/1911.09339v2.

[12] C. Li *et al.*, "Bi-directional gated recurrent unit neural network based nonlinear equalizer for coherent optical communication system," *Opt. Express, Vol. 29, Issue 4, pp. 5923-5933*, vol. 29, no. 4, pp. 5923–5933, Feb. 2021, doi: 10.1364/OE.416672.

Adaptive Neuro-Fuzzy Inference System for Predicting Stock Price of ITMG Issuer

Anne Audistya Fernanda[ab1], Aprischa Nauva Miliantari[ab2], Akik Hidayat [a3], Anindya Apriliyanti Pravitasari[ac4]
[a]*Research Center for Artificial intelligence and Big Data, Universitas Padjadjaran*
[b]*Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Padjadjaran*
Sumedang, Indonesia
[c]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]anne18003@mail.unpad.ac.id, [2]aprischa18001@mail.unpad.ac.id, [3]akik@unpad.ac.id,
[4]anindya.apriliyanti@unpad.ac.id

***Abstract*— ITMG stock price reflects investor perception of its ability to earn and grow its profits in the future. The prediction is beneficial for determining the movement of the stock price of financial exchange, therefore both issuers and investors could take important steps regarding the financial condition. In this study, a stock price prediction was carried out using the Adaptive Neuro Fuzzy Inference System (ANFIS) method with the Fuzzy Inference System (FIS) Grid Partitioning type. The initial step was to determine the parameters that produce gbellmf as a membership function type, 300 epochs, and a learning rate of 0.03. The calculation of accuracy was carried out using the RMSE calculation. The experiments conducted found that the smallest RMSE calculation result was 0.021942 with 80% of training data. Other issuers can use this method because it can predict stock prices quite well.**

Keywords—Stock, Forecasting, ANFIS, Fuzzy, Economy

I. INTRODUCTION

Indonesia is one of the world's developing countries where the finance sectors play a significant role in economic growth. The financial sector in Indonesia is divided into three main sectors, namely the banking sector, capital market, and nonbank financial industry. One sector that has a large portion of economic growth in Indonesia is the capital market sector. The capital market sector is currently in the spotlight because it improves the national economy and is globally competitive due to the availability of Indonesian capital market facilities and instruments that can compete with other countries' capital market instruments [1].

One of Indonesia's main capital market sectors is the mining sector, which operates as a producer of raw materials. The mining sector supports the national economy and national energy security, both in employment and foreign exchange earnings through exports. The mining sector is divided into five sub-sectors, one of which is the coal mining sub-sector. The coal mining sub-sector has contributed 75 to 80 percent of the mineral and coal sector's total Non-Tax State Revenue (PNBP) [2].

Therefore, to maintain and even improve it, it is necessary to predict the stock price of coal issuers. Inappropriate capital structure policies will later signal investors about the company's performance and prospects, which will affect the company's stock price [3]. Stock price prediction is an analytical technique to find out stock prices in the future by using the history of stock prices in the past. This technique avoids losses due to the nature of stock prices that fluctuate and tend to be dynamic every time, so stock price predictions are needed to maximize profits for investors. Understanding trends supported by prediction software for forecasting is critical for decision-making.

The Adaptive Neuro Fuzzy Inference System (ANFIS) method is one of the methods that has been implemented in several prediction studies, the first of which is by Mohammad Ridwan [4] regarding Prediction of Lamp Price Using Adaptive Neuro Fuzzy Inference System. This study used lamp price data every four months within six years. The result in this study is an RMSE value of 0.05. Then research by Raharja et. Al [5] which analyzes the fuzzy membership function (MF) which can give the best results in the application of the Adaptive Neuro-fuzzy inference system (ANFIS) method to predict inflation growth in Bali Province. Based on the research results, the fuzzy membership function analysis in the inflation prediction system produces the best error 1.35 x 10^{-7} with the type of triangular membership function (MF). In addition, a study was conducted by Abid et al [6] titled Car Sales Prediction System Based on Fuzzy Time Series and Adaptive Neuro-Fuzzy Inference System. This paper found that the combination of ANFIS and the Average Forecasting Error Rate (AFER) obtained a 15% better value than the Fuzzy Time Series, and the MSE value obtained by ANFIS is far below the Fuzzy Time Series. Several studies above inspire to predict stock prices using the ANFIS time series method with much larger data to prove that this method can produce an RMSE evaluation value that is smaller and far below 1.

Soft computing is currently one of the most popular solutions that can assist humans in solving problems. The methods in soft computing will complement each other and work together to form a stronger method. One of these methods is combining a fuzzy system with an artificial neural network called the ANFIS method. The ANFIS model has the advantage of having numerical and linguistic knowledge to classify data and identify patterns. The ANFIS model is more transparent to users than Artificial Neural Network and causes fewer memorization errors. Some of the advantages of ANFIS are its adaptability, nonlinearity, and fast learning capacity [7]. Therefore, ANFIS is trusted to predict stock prices.

II. RESEARCH METHODS

The data used to be tested in this study is obtained from Investing.com (web of world stock price movements)

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

https://www.investing.com/equities/indo-tambangra-historical-data. The data is stock price data with time intervals every day from PT Indo Tambangraya Megah Tbk from May 20, 2018 to May 20, 2021. The total data obtained is 1097 data.

The data analysis stage is conducted using Matlab R2018a software, with the intel core i5 processor and 8 GB DDR 4 Ram following computer specifications. The data that has been collected will be normalized first and then divided into two data, namely training data and testing data. Training data is used to train ANFIS work. Then, the testing data is used to measure the accuracy of the predictions produced by ANFIS. As for measuring the level of system accuracy, the RMSE method can be applied. The stages of data processing can be seen in Fig. 1.

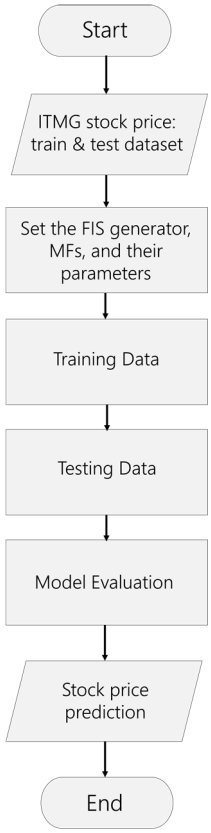


Fig. 1. Stock Analysis Stages with ANFIS

A. Determining Initial Parameters.

Before training and testing using ANFIS, it takes some of the best parameters so that the ANFIS model can produce the smallest RMSE value. The initial step to determine the appropriate initial ANFIS parameters is determine the type of membership function, the number of iterations (epochs), and the learning rate. The membership functions tested are trapmf, trimf, gaussmf, and gbellmf; then for the number of epochs used 100, 200, and 300; and finally, the learning rate test was tested from 0.01 to 0.09. After obtaining the correct initial input parameters values, then the ANFIS model can be run.

B. Adaptive Neuro-Fuzzy Inference System

The ANFIS has two methods namely the fuzzy system method and the artificial neural network (ANN). Fuzzy logic can design a qualitative aspect model based on the knowledge possessed by humans and the decision-making process by making and applying rules [10]. In comparation the Artificial Neural Network can recognize patterns, do learning to solve problems without mathematical modeling, and work based on previous data [8] that has been entered so that it can predict something that will come based on that data. The architecture of ANFIS has shown in Fig. 2.

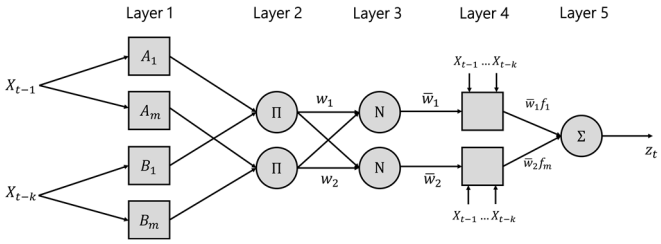


Fig. 2. ANFIS Architecture

1) Layer 1

Each neuron in the first layer is a neuron that is adaptive to the activation function. The output is a membership function of the input.

$$\mu A_1(X_{t-1}),\mu A_m(X_{t-1}), \dots, \mu B_1(X_{t-k}), \mu B_m(X_{t-k}) \quad (1)$$

2) Layer 2

Each neuron in this layer is a fixed (non-adaptive) neuron which is labeled with π. The output is a firing strength of each rule symbolized by w_j from the j^{th}-rule [9].

$$w_j = \prod_{k=1}^{p} \mu_{A_{kj}}(X_{t-k}), j = 1,2, \dots, m \quad (2)$$

3) Layer 3

Each neuron in this layer is a fixed (non-adaptive) neuron labeled N. This calculates the j^{th}'s firing strength divided by the sum of all firing strength. The output is a normalized degree of activation as follows.

$$\bar{w}_j = \frac{w_j}{\sum_{j=1}^{m} w_j} \quad (3)$$

4) Layer 4

Each neuron in the fourth layer, also known as the defuzzification layer, is an adaptive neuron. This layer has a node function:

$$\bar{w}_j f_j = \bar{w}_j(p_ı x + q_ı y + r_ı) \quad (4)$$

where $p_{\text{i}}, q_{\text{i}}, r_{\text{i}}$ are the multiple linear equation parameters which contained in this layer and it called consequent parameter.

5) Layer 5

There is one fixed (non-adaptive) neuron in this layer which is the sum of all input signals of the fourth layer. The output is as follows:

$$\hat{Z}_t = \sum_{j=1}^{m} \bar{w}_j f_j \quad (5)$$

C. Metric Of Evaluation

Tests based on absolute error are known as scale-dependent measures because their scale depends on the scale

of the data [11]. They are useful in comparing forecasting methods on the same data set. However, they should not be used across data sets that are at different scales. One of the most used is Mean Squared Error (MSE) and Root Mean Square Error (RMSE). MSE and RMSE can be calculated by the equation:

$$MSE = \sum_{i=1}^{n} \frac{(Z_i - \hat{Z}_i)^2}{n} \quad (6)$$

$$\text{RMSE} = \sqrt{MSE} \quad (7)$$

Where Z_i is the actual data, $\hat{Z}_i$ is the result of prediction/forecasting, and n is the amount of data.

III. Result and Discussion

A. Determining Initial Parameters

1) Membership Function (MF) Test Result

The first test in determining the initial parameters of ANFIS in this study was conducted by comparing four types of Membership Functions, namely trapmf, trimf, gaussmf, and gbellmf.

The test results in Table I shows that the gbellmf type has the smallest RMSE value so that it can be used for more optimal predictions.

TABLE I. *RESULT OF COMPARISON OF MF TYPE*

MFs Type	RMSE
trapmf	0.008786
trimf	0.00620
gaussmf	0.00458
gbellmf	**0.00444**

2) Epoch Test Result

The next test to determine the initial parameters of ANFIS is the Epoch ratio of 100, 200, and 300.

TABLE II. *RESULT OF COMPARISON OF THE NUMBER OF EPOCHS*

Epochs	RMSE
100	0.00444
200	0.003775
300	**0.00373**

From the Table II above, we get the smallest RMSE with 300 epochs.

3) Learning Rate Test Result

The Table III is the results of the learning rate test that has been done using the learning rate value tested between 0.01 to 0.09.

TABLE III. RESULT OF COMPARISON OF LEARNING RATE

Learning Rate	RMSE
0,01	0,004591
0,02	0,004060
0,03	**0,003588**
0,04	0,003730
0,05	0,003779
0,06	0,003701
0,07	0,003772
0,08	0,003637
0,09	0,003597

Based on Table III, the average RMSE value with the learning rate tested is good because the RMSE value is below 1. The smallest RMSE is found in the learning rate with a value of 0.03. The smaller the RMSE value, the more optimal the performance in predicting variables will be. So, it can be concluded that the optimal RMSE is obtained by using a learning rate of 0.03. This learning rate will be used for RMSE testing on training data and test data.

B. Testing Using Adaptive Neuro-Fuzzy Inference System (ANFIS) Method

In the process of testing the data, the author uses the Adaptive Neuro-Fuzzy Inference System (ANFIS) method and the Fuzzy Inference System (FIS) type to display the diagram. In addition, the author uses Grid Partitioning to perform time series prediction. In this testing process, the number of three MF's is used based on the fuzzy membership of the stock price, namely low, average, and high as well as the initial step size (learning rate) as in table 1. From the testing process based on membership function, epoch testing, and testing percent of training data and certain test data that have been done, the results obtained are as follows:

1) Training data is 50% and testing data is 50%

Based on the test results, the resulting output when we use 50% of data train, shows that the RMSE value obtained is 0.22432. Then we must try a larger data train in order to find better accuracy. The results of testing dan training in the ratio of 50:50 is shown in Fig. 3.

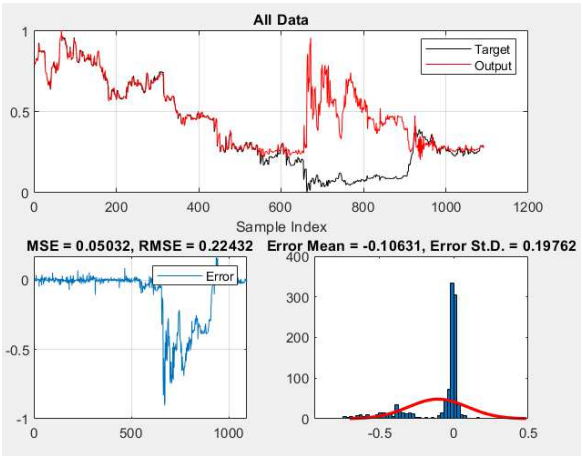


Fig. 3. Result of testing 50% of training data and 50% of testing data

2) Training data is 60% and testing data is 40%

Based on the test results, it can be seen that the output produced is much smaller than 50% of data train. It shows that the RMSE value obtained is 0.022839. The results of testing dan training in the ratio of 60:40 is shown in Fig. 4.

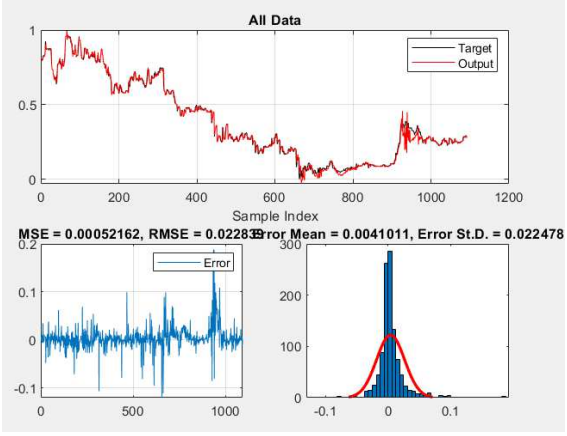


Fig. 4. Result of testing 60% of training data and 40% of testing data

3) Training data is 70% and testing data is 30%

Based on the test results, it can be seen that the resulting output is getting closer to target (which is zero). It shows that the RMSE value obtained is much better, which is 0.021983. The results of testing dan training in the ratio of 70:30 is shown in Fig. 5.

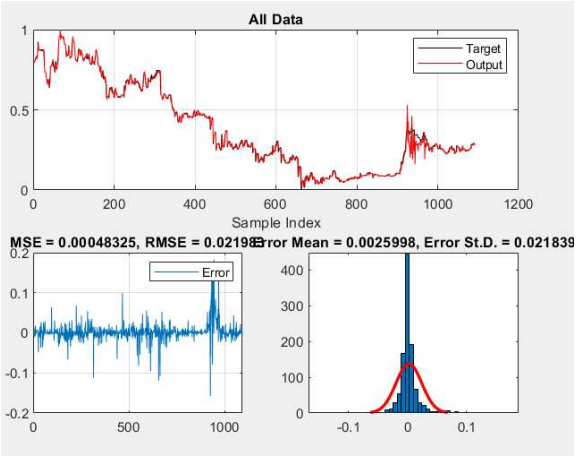


Fig. 5. Result of testing 70% of training data and 30% of testing data

4) Training data is 80% and testing data is 20%

Based on the test results, the resulting output is the closest to the target. It shows that the RMSE value obtained is 0.021942. The results of testing dan training in the ratio of 80:20 is shown in Fig. 6.

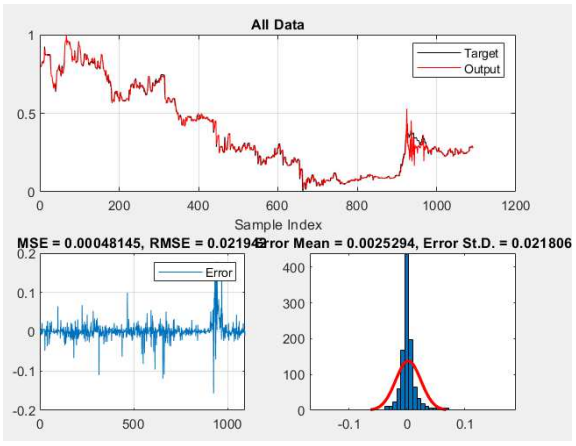


Fig. 6. Result of testing 80% of training data and 20% of testing data

C. Test Result Using Adaptive Neuro-Fuzzy Inference System (ANFIS) Methods

From the RMSE testing process conducted by training data of 50%, 60%, 70%, and 80% of the total data and test data that will be conducted for testing are 50%, 40%, 30%, and 20% of the total data or the rest of the training data, the results of the RMSE values are obtained in Table IV.

Based on the results of the RMSE values contained in Table IV, the smallest RMSE value is found in the combination of 80% training data and 20% test data. The smaller the RMSE, the higher the accuracy. The more training data used for the learning process; the fewer error value achieved.

TABLE IV. RESULT OF RMSE TEST ON EACH COMBINATIONS DATA

Training Data Percentage	Testing Data Percentage			
	20%	30%	40%	50%
	RMSE	RMSE	RMSE	RMSE
50%	0.22432	0.22432	0.22432	0.22432
60%	0.022839	0.022839	0.022839	-
70%	0.021983	0.021983	-	-
80%	0.021942	-	-	-

Based on the test results, it can be concluded that the Adaptive Neuro-Fuzzy Inference System (ANFIS) method has a good ability to predict the stock price of ITMG issuers. This result can be seen from the average RMSE value, which is close to zero, so it has a fairly good error.

IV. CONCLUSION

Based on the test results on the stock price prediction model of ITMG issuers that have been conducted, the following conclusions can be drawn:

1. The Adaptive Neuro-Fuzzy Inference System (ANFIS) method has a good ability to predict the stock price of ITMG issuer.
2. The prediction model using the Adaptive Neuro-Fuzzy Inference System (ANFIS) method has the smallest error value of 0.021942.

Based on the tests conducted, there are several things that need to be considered and the authors suggest the following:

1. It is necessary to test various methods to get a better comparison of accuracy's results.
2. Perform error or accuracy calculations using methods other than RMSE to improve the results of a better prediction model to compare each method's results.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] G. M. Hutapea, A. F. Fathoni and E. Yulia, "Investigation of Capital Market Efficiency in Indonesia," *AFEBI Management and Business Review (AMBR),* vol. 04, no. 02, 2019.

[2] EITI Indonesia, "Report Of EITI Indonesia 2018 EXECUTIVE SUMMARY (Flexible Report)," Ministry of Energy and Mineral Resouces Republic Of Indonesia, 2018.

[3] R. S. Indahwati, "Consistency Of Pecking Order Theory: Evidence From Indonesia Capital Market," 2021.

[4] M. Ridwan, "Prediction Of Lamp Price Using Adaptive Neuro Fuzzy," in *ICCSET*, Kudus, 2018.

[5] M. A. Raharja, I. D. M. B. A. Darmawan, D. P. E Nilakusumawati, D. P. E., and I. W Supriana, "Analysis of membership function in implementation of adaptive neuro fuzzy inference system (ANFIS) method for inflation prediction", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012005), 2021.

[6] A. Falih Zuhdi, A. R. Taufani, A. Firmansah and G. Jiun Horng, "Car Sales Prediction System Based on Fuzzy Time Series and Adaptive

Neuro Fuzzy Inference System," in *International Computer Symposium (ICS)*, 2020.

[7] M. Sahin and R. Erol, "A Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games," *Mathematical and Computational Applications,* vol. 22, no. 4, 2017.

[8] N. F. Mohammed and N. Omar, "Arabic Named Entity Recognition Using Artificial Neural Network," *Journal of Computer Science 8,* 2012.

[9] A. Rusgiyono , "Adaptive Neuro Fuzzy Inference System (ANFIS) approach for modeling paddy production data in Central Java," *Journal of Physics: Conf. Series 1217,* 2019.

[10] I. N. Yulita, A. Hidayat, A.S. Abdullah, and E. Paulus, "Combining Fuzzy Clustering and Hidden Markov Models for Sundanese Speech Recognition", Journal of Physics: Conference Series (Vol. 1028, No. 1, p. 012239), 2018.

[11] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

A Comparison of Support Vector Machine and Naïve Bayes Classifier in Binary Sentiment Reviews for PeduliLindungi Application

Isal Firmansyah[1], Mohammad Hamid Asnawi[2], Syifa Auliyah Hasanah[3], Rafly Novian[4], Anindya Apriliyanti Pravitasari[5]
Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran
Department of Statistics, Universitas Padjadjaran
Bandung, Indonesia
e-mail: [1]isal19001@mail.unpad.ac.id, [2]mohammad19011@mail.unpad.ac.id, [3]syifa19010@mail.unpad.ac.id,
[4]rafly19001@mail.unpad.ac.id, [5]anindya.apriliyanti@unpad.ac.id

***Abstract*— COVID-19 statistics in Indonesia show more than 4.2 million active confirmed cases with more than 140 thousand deaths. The Indonesian government has made several policies to reduce the number of COVID-19 cases, one of them is by implementing the PeduliLindungi application. The government has socialized and recommended this application as an effort to fulfill the tracking, tracing, and fencing program. Various kinds of responses appear in the community to this application, therefore sentiment analysis is needed to find out public trends so that the government can evaluate the policies that have been made. This study aims to determine the best model from the comparison of the Naïve Bayes algorithm and the Support Vector Machine, besides that this study will also see whether a simpler model such as Naive Bayes is still good in handling binary sentiment for PeduliLindungi data reviews. The data was obtained by web scraping from the PeduliLindungi application review on the Google Play Store. The Naïve Bayes accuracy value is 81%, smaller than the Support Vector Machine which has an accuracy of 84%, although the Support Vector Machine is the best model we have, Naive Bayes itself can still be used to handle binary sentiment data because the difference in accuracy values is not too far.**

Keywords*— *Sentiment Analysis, Naïve bayes, Support Vector Machine, PeduliLindungi, COVID-19

I. INTRODUCTION

The outbreak of the COVID-19 virus has become a scourge of problems throughout the world and to be the main focus for world leaders in finding ways to deal with this problem. The global pandemic was caused by a virus that first appeared in Wuhan, China. It has had a great impact and influence on various sectors of life in Indonesia, especially in the socio-economic. Indonesia's economic condition is greatly affected by the COVID-19 pandemic, in response to these conditions, in 2020 the government allocated Rp695.2 trillion for the national economic recovery (PEN) program. In February 2021, the government announced a budget of Rp699.43 trillion for the PEN program due to the unfinished condition of COVID-19 in Indonesia [1]. Not only economically, the impact of COVID-19 is also felt in various other fields, such as education, health, and even social.

Various policies have been done by the government to be able to stop the COVID-19 pandemic in Indonesia, one of which is through the Regulation of the Minister of Health of the Republic of Indonesia No. 18 of 2021 concerning the Implementation of Vaccination in the Context of Coping with COVID-19 helps reduce the number of positive cases of COVID-19 in Indonesia. It is necessary to continue to hold efforts to control COVID-19 cases in Indonesia through 3T (Testing, Tracing, Treatment) by utilizing current technological advances. The decree of the Minister of Communication and Informatics No. 171 of 2020 stipulates that the basis for carrying out tracing, tracking, and fencing activities that utilize telecommunication systems and applications to support health surveillance is the PeduliLindungi application. In helping the government to carry out tracking to suppress COVID-19 cases in Indonesia, PeduliLindungi is expected to be the right solution to deal with these cases, every information about COVID-19 is available including vaccines, etc. The users will also get notification if they are in a crowd or in a red zone. Health Minister Budi Gunadi Sadikin confirmed that the PeduliLindungi application will be included in some community activities. The application has started to be implemented in the trading sector, both in traditional markets or modern markets such as supermarkets, malls, etc. This application will also be applied in various sectors such as transportation, tourism (hotels, restaurants, performances), work, religious (mosques, churches, temples, monasteries, religious activities), and education sectors [2].

Various responses from the public to the PeduliLindungi application emerged along with the government's recommendation to use this application, this can be analyzed using sentiment analysis based on user reviews on the Google Play Store application. Sentiment analysis is a combination of text mining with natural language processing which aims to find opinions, identify what sentiments they express, and then classify them based on the values contained in them [3]. The benefit of looking for public sentiment is to know the position of this application in the community, so that the government can take another approach in recommending and socializing this application.

This paper is inspired by former paper research entitled Analysis of User Reviews for the PeduliLindungi Application on Google Play Store Using the Support Vector Machine and Naive Bayes Algorithm Based on Particle Swarm Optimization by A. Mustopa, et al [4], the conclusion from this paper can be different due to changes in situations and conditions that Indonesia faced such as the PPKM regulation, this caused different patterns of public reviews toward PeduliLindungi application, this regulation also makes the PeduliLindungi application users number increased starting August 2021, this increase was also triggered by the opening of several public services that require visitors to have PeduliLindungi application. The previous study that conducted sentiment analysis was Kristiyanti, et al [5] who

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

applied Support Vector Machine (SVM) and Naïve Bayes Classifier for sentiment analysis on the West Java Governor Candidate, then there was Rana and Singh [6] who compared Naïve Bayes and SVM for user reviews about movies. Comparison between other methods is given by Poornima and Priya [7] which results that Naïve Bayes being more suitable for simpler data. Inspired from previous research, it appears that SVM and Naive Bayes are often used in text analysis. The two methods have different algorithms, SVM is more complex than simple Naive Bayes. This study uses SVM and naive Bayes to predict public sentiment towards PeduliLindungi with only two categories, i.e., Positive and Negative (Binary Sentiment Analysis). This study aims to determine the best model in predicting public sentiment of PeduliLindungi application based on performance of these two algorithms. Besides it, this study also wants to see if the simpler Naïve Bayes algorithm is able to handle binary sentiment of PeduliLindungi application review after some new provision applied by Indonesia's government, or is it necessary to use a more complex algorithm such as SVM to predict sentiment.

II. METHODOLOGY

Sentiment analysis of the PeduliLindungi review begins with scraping data from Google Playstore (the data in Bahasa Indonesia) with WebHarvy followed by pre-processing such as case folding, data cleaning, tokenization, stopword removal, normalization, and stemming. The next step after pre-processing is weighting with the TF-IDF algorithm which will then be followed by Naïve Bayes and Support Vector Machine Classifier, in this research hold out validation is used to evaluate both algorithms. The flow chart of sentiment analysis can be seen in Fig 1.

A. Pre-Processing

Pre-processing is one of the most important steps in analysis, pre-processing ensures that the output of a data set is ready for analysis. Preprocessing identifies and fixes problems in the raw data in-to cleaner information that can be used for further processing. Here is the explanation per step.

1. Case Folding and Data Cleaning

The case folding stage basically unifies all text into lowercase, at this step the capital letters are changed to lowercase using the "RegEx" module in python. The purpose of case folding is to eliminate redundancy, which is the repetition of the same data in a database which results in wastage of storage. If case folding is not done, the same word may be counted or defined into two entities due to differences in the writing system that have not been removed.

The data cleaning step also utilizes the same module as the case folding, namely "RegEx", at this step we clean elements in the text that have no meaning at all on the results of sentiment analysis, therefore at this stage the author performs several element deletions such as removing punctuation, eliminate numbers, and so on.

2. Tokenizing

The tokenizing stage is the step of cutting sentences into a list of words that make up the sentence separated by commas and spaces so that the results are single words that are collected in the array data which will later be used in the weighting process, examples of tokenized sentences: “halo warga indonesia” to “halo, warga, indonesia”.

3. Stopword Removal

Stopword removal is the process of removing words that are included in the stopword category, stopwords are words that often appear but have no meaning on the analysis. At this step we use the nltk package in python. The examples of words that are stopwords are conjunctions/prepositions and slang words that are inserted at the end of sentences that have a function only to make an informal impression, such as the word "dong".

Stopwords are removed because they are considered unable to represent the contents of the review sentences that we have. The stopword removal process is carried out by making a stopword database. The database created will be compared with the PeduliLindungi application review data so that the result of this process is the elimination of words that exist in the database we made previously.

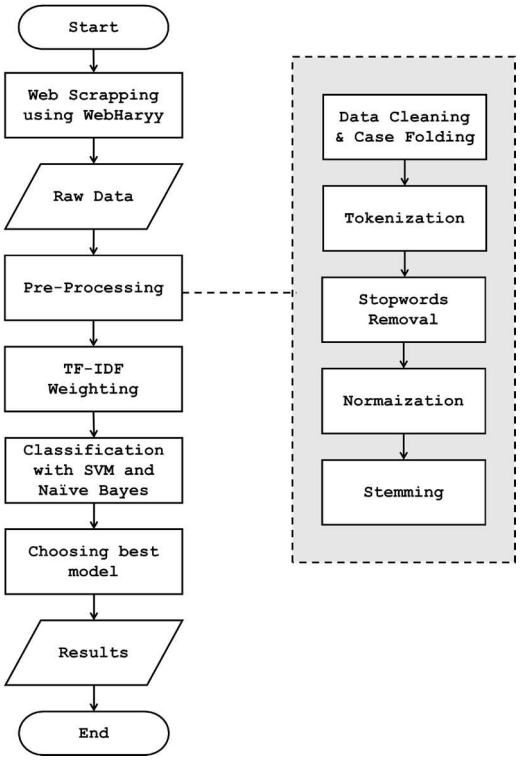


Fig. 1. Flow Chart of Sentiment Analysis for this study

4. Stemming

The stemming process is the process of changing and reducing words into their basic form [8]. The stemming process removes suffixes, confixes, and prefixes in the existing text. At this stage the author uses the pipeline and StemmerFactory modules in python.

B. TF-IDF Weighting

In the weighting process, each word will be weighted with certain rules to see the tendency of the response of the text it has. In this study, weighting was used with the TF-IDF technique. The TF-IDF method combines the weighting process of TF and IDF, calculating the frequency of occurrence of a word in a particular document and reducing the weight of a word if the word appears a lot in a document, then both are multiplied [9].

C. Sentiment Analysis with Naïve Bayes Classifier

In this study, two algorithms are compared, one of them is Naive Bayes. Naïve Bayes Classifier is a classification method based on the bayes theorem, Naïve Bayes Classifier It is popular for its ease and simplicity, although this classification provides classification results equivalent to decision tree and neutral network, moreover Naïve Bayes Classifier also provides speed in processing data in large quantities [10].

Naïve Bayes Classifier assumes that the presence or absence of a feature in a class is independent, that means a feature in a class has no connection to the existence of other features of the same data.

The Naïve Bayes Classifier process can generally be written in the following equations:

$$P(w_i) = \frac{P(c_j)xP(w_i|C_j)}{P(w_i)}, \quad (1)$$

where:

$P(c_j|w_i)$: Probability of hypothesis based on conditions $C_j w_i$(Posteriori probability)
$P(w_i|C_j)$: Probability based on conditions on w_i hypothesis C_j
$P(c_j)$: Probability hypothesis C_j (prior probability), $P(c_j) = \frac{N_{c_j}}{N}$
$P(w_i)$: Probability w_i
w_i : Unknown class
c_j : A data hypothesis that is a class C_j Specific
N_{c_j} : Documents that fall into categories c_j
N : Number of all training documents used

In the Naïve Bayes Classifier, each review is represented in attribute pairs $(k_1, k_2, \ldots, k_n)$ where k_1 is the first word, k_2 is the second word, and so on until the n-th word. In the classification process, Bayes' approach selects the category that has the highest probability (V_{MAP}) formulated in the equation as follows:

$$V_{MAP} = argmax \frac{P(c_j)xP(w_i|c_j)}{P(w_i)} \quad (2)$$

Since the value $P(W_i)$ is constant for all c_j, the value of $P(w_i)$ can be ignored so that the above equations can be written as:

$$V_{MAP} = argmax\ P(c)P(w_i|c_j) \quad (3)$$

Value $P(c_j)$ and $P(c_j)$ are calculated during the training process where the equation of the two is as follows:

$$P(c_j) = \frac{|docs_j|}{training}$$
$$P(w_i|c_j) = \frac{n_{i+1}}{vocab+n} \quad (4)$$

where:

$P(c_j)$: The probability of the word w_i in the category c_j
$|docs_j|$: Number of documents in category j
$training$: The total number of samples used in training process
n_i : The frequency of occurrence of the word w_i in the category c_j
$vocab$: Number of unique words in all training data

D. Sentiment Analysis with Support Vector Machine

Besides Naïve Bayes, another algorithm to be compared is Support Vector Machine. Support Vector Machine (SVM) is one of the algorithms used to predict regression and classification. SVM is a supervised machine learning algorithm that uses kernel functions to map data point spaces. Linearly the data cannot be separated into a new space in which it is [11]. The SVM algorithm looks for the largest hyperplane value, the classification of hyperplanes is notated as follows:

$$f(x) = w^T x + b \quad (5)$$

Similarities formed according to Vapnik and Cortes [12]:

$$[(w^T x_i) + b] \geq 1 \text{ for } y_i = +1$$
$$[(w^T x_i) + b] \leq -1 \text{ for } y_i = -1$$

where:
x_i = training data set, $i = 1,2,\ldots,n$
y_i = class label of x_i

Support Vector Machine looking for the best hyperplane located in the middle of a class divider, and by maximizing the margin or distance between two sets of objects of different classes.

With Quadratic Programming (QP) Problem Method, The Lagrange function used to optimize found by Vapnik is as follows:

$$L(w, b, a) = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i \{y_i[(w^T x_i) + b] - 1 \quad (6)$$

where $\alpha_i =$ the lagrange and $i = 1,2,\ldots,n$ is a function multiplier.

Kernel functions in SVM algorithm can be used to lower dimensions map into higher dimensions on non-linear data. Some kernel functions include:

1. Kernel Gaussian Radial Basis Function (RBF)

$$K(x_i, x_j) = exp\left(-\frac{||x_i, x_j||^2}{2\sigma^2}\right) \quad (7)$$

2. Kernel Polynomial

$$K(x_i, x_j) = \left((x_i, x_j) + c\right)^d \quad (8)$$

E. Metric of Evaluation

In classification analysis, the model will be evaluated using the metric of evaluation. Confusion matrix or error matrix is a matrix table that displays a description of the performance of the classification model on a test data set (testing) whose actual value has been known, confusion matrix provides information comparing the results of the system classification with the actual results. Confusion matrix is used to determine accuracy, precision, recall, and error rate.

TABLE I. CONFUSION MATRIX

	Positive	**Negative**
Positive	*True Positive* (TP)	*False Positive* (FP)
Negative	*False Negative* (FN)	*True Negative* (TN)

Four important terms that represent the results of the classification process in the matrix confusion are: True Positive, True Negative, False Positive, and False Negative.

1. Accuracy

Accuracy is the correct prediction ratio of the number of diagonal elements to the sum of the total matrix elements, mathematically accuracy can be formulated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (9)$$

The result of the calculation above is the percentage of the predicted amount of data that is correctly valued against the overall amount of data. Accuracy is only suitable if used at the time of comparison of the number of actual data labels relatively the same.

2. Precision

Precision can be defined as the degree of reliability of a model when the model produces a positive prediction. In calculating precision, only the first or second line of the confusion matrix is required. Precision is the proportion of positively correct predictions against the overall positive prediction. Mathematically precision can be formulated as follows:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

3. Recall

Recall or sensitivity is a method used to measure how well a test can identify a true positive, a recall describes the success of a model in rediscovering information. Recall is a comparison of positive correct predictions with overall positive correct data. Mathematically the recall value can be formulated as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

4. *F*1-*Score*

Calculations that summarize precision and sensitivity/recall by taking harmonic average calculations of both. Mathematically the value of F1-Score can be made the following formula:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (12)$$

III. Result and Discussion

The data used in this study is from the review of PeduliLindungi application on Play Store. This study mined the documents of about 2,840 comments or reviews. Table II shows the example of raw data we used.

TABLE II. EXAMPLE OF POSITIVE AND NEGATIVE SENTIMENT

Review	Sentiment
Aplikasi nya bagus banget, sangat efektif & bermanfaat sekali ? Mungkin yang belum bisa login, kurang teliti besar / kecil huruf di email. atau juga bisa login di website dulu lalu lihat akun kita agar bisa login di aplikasi (pengalaman saya) Budayakan cari info dulu / hubungi CP jika ada kendala, sebelum menghujat aplikasi & sistem ini ?? Be smart guys! Stay Safe & semangat semuanya ???	Positive
Aplikasi banyak bug nya. Input tanggal lahir jangan dibikin pake tampilan calender kalau nyatanya gak berfungsi. Bikin input manual aja pake keyboard. Setelah sekian cara saya coba, akhirnya bisa juga input tanggal lahir. Saya harus ketik yyyy-mm-dd di kolom NIK trus dicopy ke kolom tanggal lahir. Kalau bikin aplikasi jangan setengah, kalau aplikasi masih belum siap jangan paksa masyarakat gunakan.	Negative
Sejauh ini sangat membantu aplikasinya. Dari awal buka aplikasi, registrasi, hingga download sertifikat vaksinnya cuma butuh 10-15 menit saja. Saya baca ada beberapa keluhan dari pengguna lain terkait input tanggal lahir dan sertifikat vaksin belum tersedia. Waktu input tanggal lahir, bisa di klik bagian tahun jadi bisa milih tahunnya dulu, baru bulan dan tanggal. Jadi tidak perlu klik geser bulan satu per satu. Saya vaksin 18 Agt, lalu saya register disini 20 Agt alhamdulillah sudah ada sertif.	Positive
Aplikasi buruk. Ini gimana? Udh download dan login. Tapi baru keluar buka aplikasi lain sebentar udh gk bisa di pake lagi . Padahal gk log out . Pas masuk aplikasi lagi malah minta akses GPS terus . Padahal GPS aktif. Sampai restart juga tidak bisa digunakan .	Negative

Table III shows the comparison of one example of a comment that has been done through the text pre-processing stage.

TABLE III. PREPROCESSING RESULT TABLE

Raw Data	Data after Preprocessing
Saya sudah vaksin 2x, tapi knapa hanya di beritahukan vaksin pertama saja, yg kedua belum vaksin „padahal sudah vaksin 2x, parahnya sertifkat vaksin pertama juga blum ada katanya, padahal sudah vaksin pertama bulan april, Tolong perjelas, jangan buat aplikasi abal2 dan menipu publik!!!!!!!!!!!!!!!!	vaksin kenapa beritahukan vaksin parah sertifikat vaksin belum vaksin tolong jelas aplikasi abal tipu publik
⋮	⋮
Aplikasi apa ini!!!! Masa terdeteksi make fake GPS, padahal gk make tuh. Ayolah Menkominfo, saya cuman mau *download* sertifikat vaksinasi buat ke sekolah, klo masih kaya gini, mending hapus aja dari *playstore*. Daripada cuman jadi beban.	aplikasi deteksi pakai palsu gps pakai ayo menkominfo hanya unduh sertifikat vaksinasi sekolah seperti begini mending hapus hanya beban

The comments have undergone changes in the text pre-processing. At this phase, the process of case folding and data cleaning, tokenization, stopword removal, normalization, and stemming is carried out. The results of the text pre-processing phase produce text data that are ready to enter the next stage, namely the weighting stage. The following is a data visualization of the results of text pre-processing.

Fig. 2 shows a visualization to represent the number of words that appear in the PeduliLindungi application review. It can be seen that there are 10 words that users often use to express their opinion about the application. It can be seen that

the word 'aplikasi' is the word with the highest number of uses, namely 5,041 words, followed by the word 'vaksin' as many as 4,118, the word 'sertifikat' as many as 2,698, the word 'masuk' as many as 1,491, the word 'pakai' as many as 1278, the word 'tolong' as many as 1,136, the word 'lahir' as many as 994, the word 'tanggal' as many as 994, the word 'nik' as many as 781, and the word 'unduh' as many as 710 words.

At the stage of determining the TF-IDF matrix normalization is carried out on the TF and IDF to obtain the TF-IDF values and the ranking are obtained based on the weighting above as in Table IV.

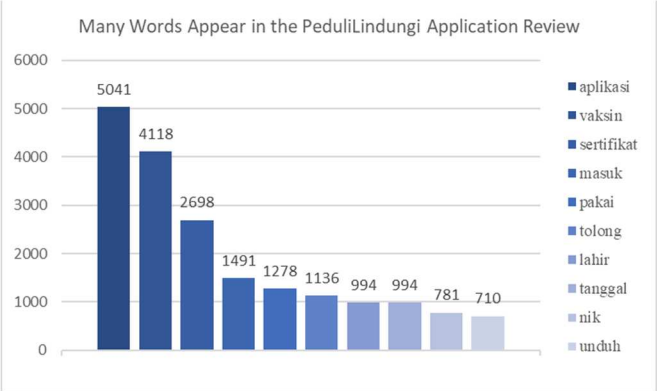


Fig. 2. Bar chart of the number of words that appear in PeduliLindungi app reviews

The results obtained from the ranking of the top three words, namely the word 'aplikasi' which got the first rank with a weighting value of 381.0746. The word 'vaksin' got the second rank with a weighting value of 344.266 and for the third rank the word 'sertifikat' was obtained with a weighting value of 329.8253.

TABLE IV. WEIGHTING RESULT TABLE

Word	Weighting
aplikasi	381.0746
vaksin	344.266
sertifikat	329.8253
tanggal	258.5432
tolong	244.0897
pakai	230.3026
masuk	228.5307
unduh	210.3529
nik	187.1637
lahir	178.803

Sentiment analysis was performed using the SVM and Naïve Bayes algorithm on the PeduliLindungi application review data set, this research using hold out validation, in this case 80% split was applied to training data from all data (2,271 data) and 20% split to test data from all data (568 data). The distribution of data is given in Fig. 3.

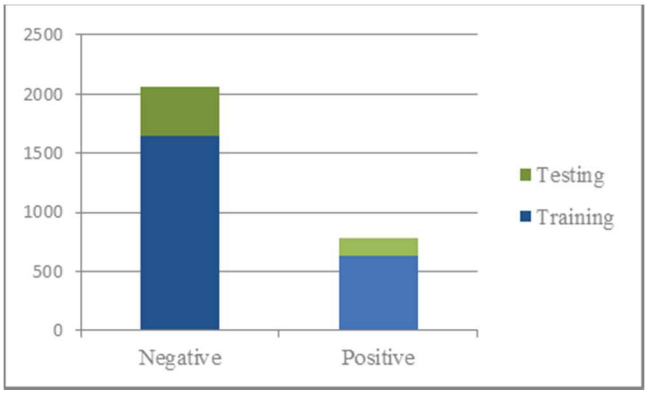


Fig. 3. Distribution of the document labels.

Fig. 3 shows the distribution sentiment label for the data. It appears that mostly the reviews or public sentiments towards this application are negative. This information can be a concern for the government to carry out a strategy to better introduce and socialize this application to the community.

The confusion matrix of Naïve Bayes and SVM algorithm can be seen in Table V and Table VI.

TABLE V. CONFUSION MATRIX NAÏVE BAYES CLASSIFIER

	Positive	Negative
Positive	367	48
Negative	59	*94*

TABLE VI. CONFUSION MATRIX SUPPORT VECTOR MACHINE

	Positive	Negative
Positive	381	34
Negative	55	*98*

The comparison of the algorithm through the accuracy, precision, recall, and F1-score from testing data is given in Table VII. From Table VII.

TABLE VII. MODEL EVALUATION

Performance	Naïve Bayes	SVM
Accuracy	81.16%	84.33%
Precision	86.15%	87.38%
Recall	88.43%	91.80%
F1 Score	87.27%	89.50%

In a previous study conducted by Mustopa et al [4], the simpler Naïve Bayes algorithm could not classify the sentiment review of the PeduliLindungi application properly because the accuracy of 69% was quite small compared to the Support Vector Machine algorithm which has 93% of accuracy. There is a different result in this research, the Naïve Bayes accuracy is not much different from the Support Vector Machine, this is because the data used is new data after the 2021 PPKM policy.

This research can classify public sentiment towards PeduliLindungi application reviews so that the tendency of public sentiment can be known. The results of this public sentiment classification can be used by the government as an evaluation to develop better PeduliLindungi applications. From this research, we got a list of words that have the biggest weight, the list is the words that are most often written in the PeduliLindungi application reviews on Google Playstore. From the list, the government can make improvements that are more specific to several sectors. One of the words on the list is 'sertifikat' or certificate, after we searched it turned out that many reviews complained that the issuance of vaccine certificates was taking too long to process.

IV. CONCLUSION

In this study, the Naïve Bayes algorithm and SVM are compared to see which algorithm is the best in classifying the sentiment review of the PeduliLindungi application, the simpler Naive Bayes algorithm can still be used properly, the simpler Naïve Bayes algorithm can still be used properly to handle binary sentiment analysis of PeduliLindungi application because the accuracy value is not too far from the SVM algorithm. but the SVM is considered better in predicting binary sentiment review for PeduliLindungi application. However, the difference in the accuracy is about 3%, which means the valid prediction difference is about 18 documents. If we look at the confusion matrix, the number of miss-prediction comments or reviews is quite large, this could happen due to the possibility of another type of sentiment that is not measured in this study, namely neutral sentiment.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] I. N. Yulita, et al, "Comparison multi-layer perceptron and linear regression for time series prediction of novel coronavirus covid-19 data in West Java", Journal of Physics: Conference Series (Vol. 1722, No. 1, p. 012021), 2021.

[2] P. Violleta, U. Liman and S. Haryati, "PeduliLindungi to be used in people's daily activities: Minister," ANTARA News, 7 October 2021. [Online]. Available: https://en.antaranews.com/news/193073/pedulilindungi-to-be-used-in-peoples-daily-activities-minister. [Accessed Oktober 2021].

[3] B. Liu, "Sentiment Analysis and Opinion Mining," *Morgan & Claypool Publishers. ,* 2012.

[4] A. Mustopa, H. A. E. P. Pratama, A. Hendini and D. Risdiansyah, "Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naive Bayes Algorithm Based on Particle Swarm Optimization," 2021.

[5] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin and L. Marlinda, "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," *The 6th International Conference on Cyber and IT Service Management (CITSM 2018),* 2018.

[6] S. Rana and A. Singh, "Comparative Analysis of Sentiment Orientation Using SVM and Naive Bayes Techinques," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016),* 2016.

[7] P. A and K. S. Priya, "A Comparative Sentiment Analysis of Sentence Embedding Using Machine learning Techniques," *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS),* 2020.

[8] C. Gallagher, E. Furey and K. Curran, "The Application of Sentiment Analysis and Text Analytics to Costumer Experience Reviews to understand What Costumers Are Really saying," *International Journal of Data Warehousing and Mining,* vol. 15, no. 4, pp. 21-47, 2019.

[9] B. G. Gebre, M. Zampieri, P. Wittenburg and T. Heskes, "Improving Native Language Identification with TF-IDF Weighting," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, 2013.

[10] C.Aggarwal, Data Classification: Algorithms and Applications, Minneapolis, Minnesota, USA: Chapman & Hall/CRC , 2015.

[11] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, 2004.

[12] C. Cortes and V. Vapnik, "Support-Vector Networks," in *Machine Learning*, Boston, Kluwer Academic Publishers, Boston, 1995, pp. 273-297.

[13] V. K. Chauhan, K. Dahiya and A. Sharma, "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review,* vol. 52, p. 803–855, 2018.

Utilization of Data Warehouse in Business Intelligence with Kimball Method at Company XYZ

Muhammad Himami[a1], Atje Setiawan Abdullah[a2], Intan Nurma Yulita[ab3], Mira Suryani[a4]
[a]*Department of Computer Science, Universitas Padjadjaran*
Sumedang, Indonesia
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Bandung, Indonesia
e-mail: [1]maqwamhimami@gmail.com, [2]atjesetaiwan@gmail.com, [3]intan.nurma@unpad.ac.id, [4]mira.suryani@unpad.ac.id

***Abstract*— For a large company the speed of processing and distribution of data is very important. However, at Company XYZ, this has not yet been implemented in its entirety. This hinders the progress of the company. To be able to solve these problems, a storage system is needed that can access data more quickly and effectively by using a data warehouse system. This study discusses the design and construction of a data warehouse using the Kimball method. The data warehouse schema is created using SQL Server Management Studio and Microsoft SQL Server Integration Service. The analysis results from the data warehouse are displayed in the form of a dashboard using Microsoft PowerBI and a web-based application. Judging from the dashboard display that has been made, the production environment at PT Bio Farma is stable so it does not interfere with production quality. The data is entered into a web-based application, with the percentage of application feasibility of 84% after testing.**

Keywords—Data warehouse, Kimball Method, Production Environment.

I. INTRODUCTION

XYZ is a company with a business focus on the production of vaccines and antisera that have reached international standards. With its various achievements at the international level, It needs to continue to improve its work performance. One of the things that need to be considered to improve work performance is about easy and effective data processing. There are still data processing complexities in some parts, one example is in the Production Department. This section still relies on manual data processing and retrieval from an application called LIMS (Laboratory Information Management System) and needs to be combined with the database on the Information Technology Division server for monitoring the production environment. This is considered less effective considering the monitoring process will often be carried out monthly or every few months. Therefore, we need a data storage system that can simplify data processing and can integrate large amounts of data contained in different sources.

The solution that can be applied to solve the above problems is to implement a data warehouse system. With the implementation of a data warehouse, data from several different sources can be put together in one place so that the processing will be simpler. The data warehouse itself has four main characteristics, namely subject oriented, integrated, time variant and non-volatile, also useful for determining various decisions that are crucial in business [1]. In building a data warehouse, a method that fits the business needs of one division in the company is needed, the Kimball method was chosen because this method focuses on certain business processes and creates a user-friendly data system [2]. According to Kimball, the data warehouse/business intelligence has become like a combination, because the data warehouse and business intelligence are interconnected, and according to him the data warehouse is intended for business intelligence, and the data warehouse is a platform for all forms of business intelligence [2].

While business intelligence is a collection of techniques and tools for transforming raw data into useful and meaningful information for business analysis purposes. BI technology can handle huge amounts of unstructured data to help identify, develop, and otherwise create new business strategic opportunities [6]. Some of the benefits that can be obtained when an organization implements BI (Business intelligence) such as improving organizational data and information by integrating all data so as to produce complete decision making. Then the resulting data and information becomes more accessible and easier to understand, thus facilitating the monitoring of organizational performance. Then increase the value of existing information technology investments without having to change or replace information systems that have been used previously. Then create employees who have good access to information. BI (Business intelligence) can increase efficiency because it makes it easier for someone to do work, saves time, and is easy to use [7].

Hopefully, the development of a data warehouse as support for BI at XYZ can expedite and make it easier to determine company development decisions.

II. LITERATURE STUDY

A. Data Warehouse Architecture

Data warehouse is a unit of parts that are interconnected with one another. These sections are divided into four main parts, namely data source, staging area, warehouse and presentation area [4]. The architecture is shown in Fig 1.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

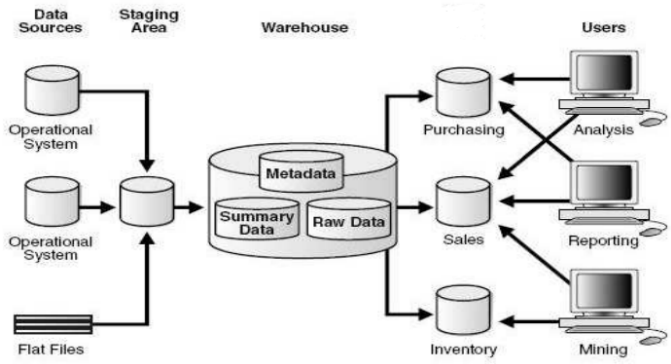


Fig. 1. Data warehouse architecture[4].

B. Kimball Lifecycle

Kimball lifecycle is a data warehouse system development method. This concept consists of several processes that are parallel/sequential, in which the concept has three other important concepts in it [2], namely:

- Focus on adding business value.
- Dimensional structure.
- Better to manage with a manageable cycle than to send it directly in bulk.

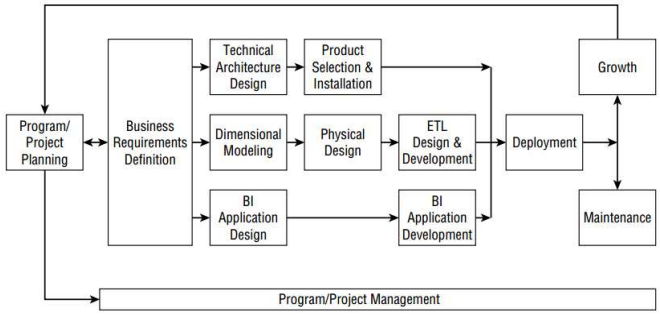


Fig. 2. Kimball Lifecycle [2].

Fig. 2 shows the Kimbal lifecycle process. There are 9 important stages in the Kimbal Lifecycle, which are as follows:

- Define process;
- Determine the grain;
- Dimensional identification and confirmation;
- Determining the facts;
- Storing calculations in a fact table;
- Complete dimension table;
- Determine the duration of the database to be created;
- Seeking a slowly changing dimension;
- Determine the physical design.

C. Alpha and Beta Testing

Alpha and beta testing are two stages of testing an application. Alpha testing is carried out before the launch of the application with the aim of ensuring the application can run well on the operating system while beta testing is testing at the deployment stage and is carried out directly to the user [9][10].

In alpha testing, it is done by black box testing technique. Black box testing is used to test the specific functions of the designed application [11]. At this testing stage, the testers involved are people who know about the workings of the application who are fellow application developers. The main point that is considered in this test is to re-check every feature contained in the application being tested, whether these features can run according to their function and there are no bugs [11].

After alpha testing is complete, beta testing is carried out which is a test carried out in the actual environment, namely to application users without the presence of the application developer. The results of the assessment of this test can be in the form of a questionnaire to application users regarding the application used. The main point that is considered in this test is the readiness of the application, whether the application is in accordance with the needs of the user or not [9]. Then the results of the assessment will be calculated using a Likert scale.

III. Development Methodology

A. Systematics of Data Warehouse Planning and Development Project Planning

This first stage is the planning stage related to the needs of the data warehouse development. This planning process is carried out through interviews with relevant stakeholders. Details of the results of the requirements that show the planning process for developing a data warehouse can be seen in Table I.

TABLE I. Analysis Results of Data Warehouse Development At PT Bio Farma

No	Aspects	Goal
1	*Trend* Data EM Partikel	Supervising Trend Environment Monitoring (TEM) at each sampling point at a certain time.
		Analyze the performance of particle monitoring environment results on the value of alert limit, action limit and specifications on each parameter 0.5µm and 5.0µm
2	*Trend* Data EM Mikrobiologi	Monitoring Trend Environment Monitoring at each sampling point at a certain time.
		Analyzing the performance of environmental monitoring of particles on the value of alert limit, action limit and specifications for each parameter of contact agar (CA), settling plate (SP), air sampling (AS), finger dab (FD).
3	*Key Performance Indicator*	Monitor the number of processes and product batches that have been approved and rejected and whose quality status has been determined.
		Monitor the Environment Monitoring process whose status is registered, received, prepared, tested and validated.

A. Collection and Analysis of Data Sources

At this stage of data collection, data is taken from various different sources and collected. The sources of these data can come from interviews, questionnaires, direct data collection, transaction data, and so on. In this case the data is sourced from data that already exists on the company's side, namely the Information Technology Section sourced from the Production Section taken from the LIMS application at PT Bio Farma.

After the data collection stage has been completed, before the data is entered into the data warehouse system, further analysis will be carried out on the data so that the data that has been obtained can become data that can be used in

business intelligence with good and quality. Table II shows the results of the analysis of data sources at PT Bio Farma.

TABLE II. DATA SOURCE ANALYSIS

No	Task	Data Requirement
1	Trend Data EM Partikel	Register Date; Register Time; Limit Text; Number of Particles; Parameter
2	Trend Data EM Mikrobiologi	Register Date; Register Time; Limit Text; Number of Particles; Parameter
3	Key Performance Indicator	Sampel ID; Product Status; status EM; final review date; approve date;

B. *Technology Track*

This stage is the stage to determine what kind of technology will support the implementation of research and development of data warehouse systems [8]. In addition, at this stage it will be explained how the technical architecture is designed, followed by selecting and using products that are in accordance with system development.

Looking at the various information needs and objectives of the research, the overall performance framework has been determined. In Fig. 5 described the technical architecture design used in the development of business intelligence, starting from the data source, entered into the database, processed and displayed on the dashboard.

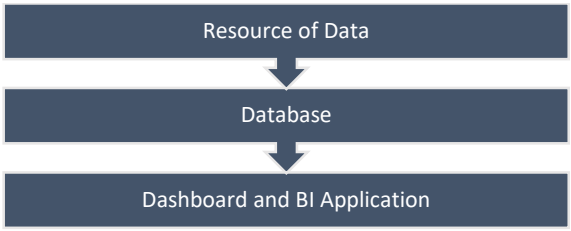


Fig. 3. Technical architecture design.

C. *Data Track*

At this stage, several activities are carried out, starting from dimensional modelling, physical design, ETL design and development. The four activities must be carried out sequentially and carefully so that the results of the data will then be processed.

1) Dimensional Modelling.

- Determining Grain The results of determining the grain containing the dimensional and dimension model schemes can be seen in Table III.

TABLE III. DETERMINING GRAIN

No	Dimensional Model Schematic	Dimension
1	*Trend* Data EM Partikel	Sample; Date; Time; Group; Limit; Parameter;
2	*Trend* Data EM Mikrobiologi	Sample; Date; Time; Group; Limit; Parameter;
3	*Key Performance Indicator*	Sample; Date; Group; Workflow; Produk;

- Identify and Confirm Dimensions. The results of the identification and confirmation of grain dimensions can be seen in Table IV.

TABLE IV. DETERMINING DIMENTION OF GRAIN

No	Dimension	Attribute
1	Date	Date, years, day, month
2	Time	Time30, Hour30, Minute Number,
3	Parameter	ParameterName, ParameterDescription
4	Division	DivisionID, Division Name
5	Group	SamplingPoint, Process, GroupID
6	Limit	LimitText, LimitType, LimitTypeName
7	Product	ProductBets, Class, ProductID
8	QualityStatus	Status, StatusType

Table III and Table IV are an explanation table about what attributes are in each dimension table and an explanation of the value or content of each of these attributes. These tables are master tables that will be used later in the data warehouse and are tables that will be related to the fact table.

- Determining the Facts. The list of facts can be seen in Table V.

TABLE V. DETERMINING THE FACTS TABLE

No	Nama tabel	Keterangan
1	Trend_EM_Partikel	See how the EM particle trend in the Production section of PT Bio Farma.
2	Trend_EM_ *Mikrobiologi*	See how the EM particle trend in the Production section of PT Bio Farma
3	Key_Performance_ Indicatior	See how the work performance has been done for a certain time at PT Bio Farma

- Saving Calculations in a Fact Table. In the fact table that has been determined previously, there are several things that need to be calculated, namely the fact table of Particle EM trends, microbiological EM trends and key performance indicators.
- Determine the Duration of the Database. The data that will be carried out in the data analysis process is sampling point monitoring data in production buildings at XYZ contained in the LIMS application starting from 2019.

2) Physical Design

In the previous stage, dimensional modeling has been formed, the data is then formed into a physical design and entered into the data warehouse database. The following are some examples of physical data warehouse designs:

TABLE VI. WORKFLOW FACT TABLE STRUCTURE

No	Attribute Name	Data Type
1	SK_Key_Perform_Indicator	int
2	IDSample	varchar(50)
3	FK_RegDate	int
4	FK_EnterDate	int
5	FK_Group	int
6	FK_Product	int
7	FK_Workflow	int

TABLE VII. FACT TABLE STRUCTURE OF MICROBIOLOGICAL TRENDS

No	Attribute Name	Data Type
1	SK_TrendMikrobiologi	int
2	IDSample	varchar(50)

No	Attribute Name	Data Type
3	FK_RegisterDate	int
4	FK_RegisterTime	int
5	FK_EnterDate	int
6	FK_EnterTime	int
7	FK_Group	int
8	FK_Parameter	int
9	FK_Limit	int
10	LimitValue	int
11	TestResults	int

TABLE VIII. EM TREND FACT TABLE STRUCTURE.

No	Attribute Name	Data Type
1	SK_TrendMikrobiologi	int
2	IDSample	varchar(50)
3	FK_RegisterDate	int
4	FK_RegisterTime	int
5	FK_EnterDate	int
6	FK_EnterTime	int
7	FK_Group	int
8	FK_Parameter	int
9	FK_Limit	int
10	LimitValue	int
11	TestResults	int

3) ETL Design and Development

After getting data from various sources, then the existing data is collected and carried out in three important stages, namely extraction, transformation and loading. In this study, ETL was performed using Microsoft SSIS.

- *Extraction.* The extraction process is the process of moving and collecting raw data from different data sources into the data warehouse database. In this process, data selection is carried out, which data will be included or not. The first stage is the selection of which columns will be inputted into the data warehouse. Starting from the 93 columns from the source, the required 28 columns have been taken. At the extraction stage, adjustments are also made to the data type from the source table to the destination table.
- Transformation. At this stage, the process of converting raw data from the data source into a form that suits the needs of the data warehouse is carried out. The first process is sorting or sorting existing data as well as deleting duplicate data records. The second is merging. This process is the process of concatenating the same data records but having different values of "limitAlert" and "limiSpecification". Next is cleansing, which is data cleansing. Cleaned data is data that has a value of 'null' will be changed to 0. The last is the lookup, which is to match the foreign key in each column in the fact table that has been related to the dimension tables so that what is obtained in the fact table is the primary key of each dimension table.
- Loading. Send and enter data that has gone through the transformation process to the tables that already exist in the database warehouse. The loading process carried out here is full loading, which is the transfer of all data from the beginning to the latest data. This is done because the implementation of the data warehouse in the production department at PT Bio Farma has never been done before, although it takes a long time, it can facilitate the next ETL process. The full loading process will also be simpler and can minimize errors in data transfer.

D. Dashboard Implementation in Power BI

After carrying out the process of designing and building a data warehouse, the results are in the form of a dashboard display containing graphs to show the results of data processing in the production section at PT Bio Farm.

- Dashboard Trend Environment Monitoring. In Fig. 4 there are results from graphing the trend of EM particle while in Fig. 5 graphs on EM microbiology trends found in the Production Department at PT Bio Farma. The graph shows that the Production Department at PT Bio Farma has a stable production environment so that the graph is in the form of a straight line. The straight line shows that the production process can run as it should. The results of the graphical display are obtained from the data in the data warehouse which is then carried out several OLAP (online analytical processing) operations.

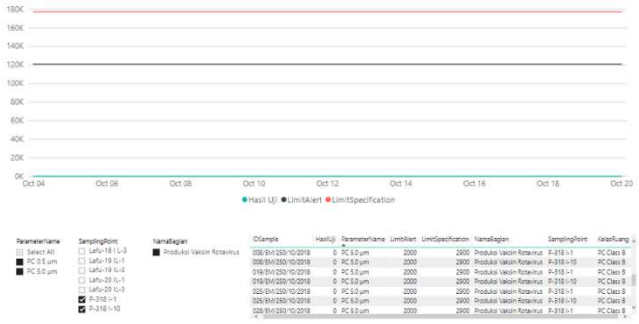


Fig. 4. Power BI particles trend dashboard display.

- OLAP Dicing Operation. The first OLAP operation is dicing, which is taking only a few parts of the entire data. In accordance with the needs of PT Bio Farma itself, the data displayed here is particle test data based on the class of production room, the parameters used, and the part of the production. Illustration of dicing Trend Environment Monitoring (TEM) count can be seen in Fig 6. While the Example of dicing Trend Environment Monitoring particle count can be seen in Fig. 7.

Fig. 5. Power BI microbiology trends dashboard display.

Fig. 6 is an illustration of the dicing process that occurs in the previous two graphs (Fig. 4 and Fig. 5). In the picture it can be seen that the cuts made are on the type of parameters, sampling point and class of space from the existing data. This dicing process can be done in Microsoft PowerBI using the filtering feature to determine the condition of the data to be displayed, such as selecting the sampling point, parameter or

class of space to be displayed. However, judging from the completeness of the existing data, the class A room still cannot be displayed.

- OLAP Slicing Operation. Next is the slicing operation. The operation was carried out to retrieve data regarding the number of particles in the name of a certain section in the Production Division of XYZ. Fig. 7 is an example of a slicing operation that has been performed and is shown in Fig. 8. From the graph in Fig. 7 it can be seen that the section that displays the data is the Rotavirus Vaccine Production Section.

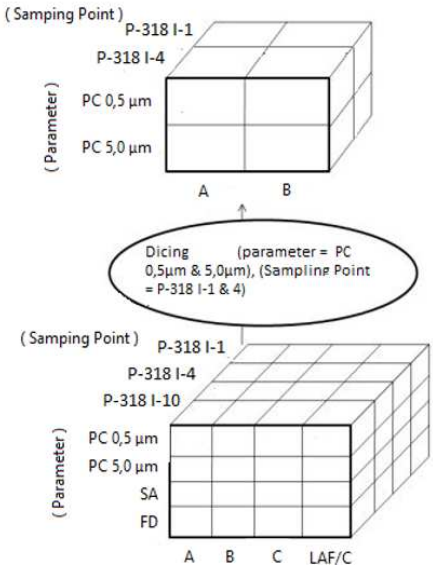


Fig. 6. Illustration of Dicing Trend Environment Monitoring Count

- OLAP Drill-down Operation. After the slicing operation, next is the drill-down operation on the 'date', this operation is carried out because the user needs at XYZ is to see the data per day.

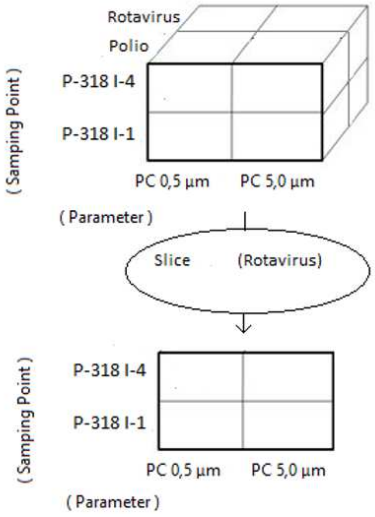


Fig. 7. Illustration of slicing Trend Environment Monitoring particle.

E. *Interface of Power BI Dashboard*

For example, Fig. 8 is a dashboard regarding key performance indicators. This dashboard displays data from the beginning of 2019, there were 426 samples and product samples that had been approved reached 100%. There is also a range of days between the date of preparation and the date of product approval, and the average is 61 days. There is also a calculation of how many product samples were registered, received, prepared, tested and validated and it reached 100% of all. The operation used here is only the calculation of the amount of data and the calculation of the difference between 2 dates.

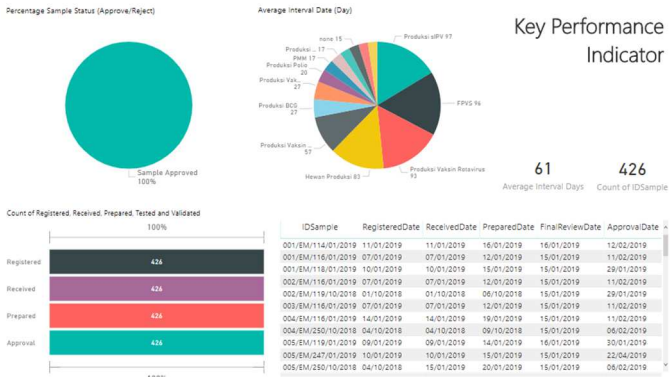


Fig. 8. Key Performance Indicator dashboard interface.

F. *System Testing and Evaluation*

The test is carried out in stages, namely Alpha and Beta testing. Detailed scenarios in alfa testing can be seen in Table 9. While the results of Alpha testing can be seen in Table X.

TABLE IX. ALPHA TESTING SCENARIO

Feature	Testing Pattern	Testing Activity
Login	Fill in the email and password input box	Normal scenario, with correct email and password
		Alternative scenario, with incorrect email and password
Index	View the initial display of the application	Normal scenario, the system displays the start page of the application
Trend EM Particle	See the EM particle trend dashboard display	Normal scenario, the system displays the EM particle trend dashboard display
Trend EM Mikrobiologi	See the EM microiology trend dashboard display	Normal scenario, the system displays the EM microiology trend dashboard
Key Performance Indicator (KPI)	View the KPI dashboard view	Normal scenario, the system displays the EM microiology trend dashboard

TABLE X. ALPHA TESTING RESULTS

Feature	Test Pattern	Testing Activity	Result
Login	Fill in the email and password input box	Normal scenario, with correct email and password	appropriate
		Alternative scenario, with incorrect email and password	
Index	View the initial display of the application	Normal scenario, the system displays the start page of the application	appropriate
Trend EM Particle	See the EM particle trend dashboard display	Normal scenario, the system displays the EM particle trend dashboard display	appropriate
Trend EM	See the EM microiology	Normal scenario, the system displays the EM	appropriate

Feature	Test Pattern	Testing Activity	Result
Mikrobiologi	trend dashboard display	microiology trend dashboard	
KPI	View the KPI dashboard view	Normal scenario, system displays Key Performance Indicator	appropriate

Using the same scenario as the alpha testing, the beta testing process was carried out to respondents outside the team, namely stakeholders from PT Bio Farma. To get the quantification value, the results of the user's assessment of the application built using a number scale from 1 to 10. Then to see the weight and percentage of results in this beta test, it can be calculated using a Likert scale with the following formula:

$$Result = \frac{\text{total score obtained}}{\max score} \times 100\% \quad (1)$$

The results of calculations using these formulas will be seen as material for assessing the feasibility of the application. The eligibility category of this application can be seen based on the following criteria:

TABLE XI. APPLICATION FEASIBILITY CRITERIA

No	Score in percent (%)	Eligibility Category
1	< 21 %	Very Not worth it
2	21 – 40 %	Not feasible
3	41 – 60 %	Fairly
4	61 – 84 %	Worthy
5	85 – 100 %	Very Worthy

From the previous formula, the percentage value of the results in this beta test is 84%. This shows that the business intelligence application that has been built is feasible to use.

IV. CONCLUSION

The development of a data warehouse to support business intelligence at PT Bio Farma begins with the analysis phase in the Production Section regarding monitoring of sampling points at PT. Bio Farma uses the Kimball method. The design of a data warehouse system using SQL Server tools and Microsoft Visual Studio SQL Server Integrated Service can be done using the results of the needs analysis. The analysis is taken based on user requirements and the results of discussions with users, namely PT Bio Farma in the production section, regarding what needs need to be implemented in the data warehouse system to be built. The result of the design is a data warehouse schema with 3 fact tables and 8 related dimension tables and an ETL process schema. The results of the design can be implemented by displaying information in the form of a dashboard related to environmental monitoring trends and key performance indicators in the production department at PT Bio Farma.

ACKNOWLEDGMENT

The author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service Contract, Universitas Padjadjaran No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] S. Saxena and M. S. Mathur, A Lifecycle based Testing of Data Warehouse. International Journal of Advanced Research in Computer Science and Software Engineering, 4(12), 2277–128, 2014

[2] R. Kimball, M. Ross, B. Becker, J., Mundy, and W. Thornthwaite, The Kimball Group Reader, 2010

[3] W. H. Inmon, Building The Data Warehouse. Indianapolish: Wiley Publishing, 2005

[4] I. Singh, Study of Data Warehouse Architecture, 2(7), 2013

[5] M. Suknovic, M. M. M. Cupic, Data warehousing and data mining: A case study. (November 2014), 2005

[6] Negash, S, Business Intelligence, July 2015

[7] E. Z. Ed, Business Intelligence and Big Data, Vol. 324, 2018

[8] R. Kimball and M. Ross, "Relentlessly Practical Tools for Data Warehousing and Business Intelligence," p. 915, 2016

[9] M. A. Bujang, E. D. Omar, and N. A. Baharum, "A review on sample size determination for Cronbach's alpha test: a simple guide for researchers", The Malaysian journal of medical sciences: MJMS, 25(6), 85, 2018.

[10] C. K. N. C. K. Mohd, and F.Shahbodin, "Personalized learning environment: alpha testing, beta testing & user acceptance test" Procedia-Social and Behavioral Sciences, 195, 837-843, 2015.

[11] J. N.Fishbein, et al, "Mobile application to promote adherence to oral chemotherapy and symptom management: a protocol for design and development", JMIR Research protocols, 6(4), e6198, 2017.

Dynamic Channel Allocation Technique for Cognitive Radio based UAV Networks

Md Sabbir Hosen
School of Computer & Communication Engineering
University of Science & Technology Beijing
Beijing, China
Email: s20191488@xs.ustb.edu.cn

Yunfeng Peng
School of Computer & Communication Engineering
University of Science & Technology Beijing
Beijing, China

***Abstract*— Unmanned Aerial Vehicles (UAVs) are the modern technology for Public Safety Communication Systems. With the integration with wireless & cellular networks, UAVs can offer dynamic solutions to resolve various communication challenges. Swarm of UAVs collaborates to form a wireless ad hoc network with applications in different areas of life. The line-of-sight connectivity provided by unmanned aerial vehicles (UAVs) is their most unique feature. Within the shortest time, UAVs can provide wireless access in facing sudden congestion situations. The mobility and connectivity of UAVs are dynamic. Drones can be used to transmit & receive data from a variety of flying and ground devices (connecting with ground controllers & unmanned aerial vehicles). For reliable and secure effective communication, UAVs require massive Bandwidth. However, Bandwidth is a limited resource. So, Cognitive Radio Technology is a prospective technique to overcome from this critical moment. Cognitive Radio technology can detect available free channels, can utilize available Bandwidth, and operate spectrum management. Cognitive Radio-based UAV Network is an intelligent device-to-device communication system where a hundred devices are connected through a wireless communication model. This research article suggests an approach for prioritizing channel allocation in a Cognitive Radio-based UAV Network using the Fuzzy Logic Algorithm.**

Keywords— Cognitive Radio Networks, UAV Networks, Fuzzy Logic Algorithm.

I. INTRODUCTION

Unmanned Aerial Vehicle (UAV), generally known as a drone, has grown increasing popularity in recent years and has been successfully implemented in several public, commercial & industrial applications such as Surveillance, Surveying & Mapping, Search & Rescue, Security & Emergency Response, Disaster relief & management, Weather forecast, Inspections, Precision Agriculture, etc. The integration with wireless and cellular networks, swarms of unmanned aerial vehicles (Drones) can offer innovative solutions and high spectral efficiency. Because of the mobility, adaptability, and dynamic connectivity, autonomous & remote-controlled UAVs can perform a variety of activities without the need of direct human interruption. UAVs can be integrated with communication systems that allow them to communicate with air & ground nodes. Because of the adaptive 3D positioning, UAVs have the capability to operate stable transmissions with shortened path loss effect. UAV network is the two-way communication system where a Swarm of UAVs collaborate and share the data with each other. So, The UAVs need massive volume spectrum space. However, Bandwidth is a minimal resource. Utilizing limited Bandwidth, spectrum, and channel allocation in UAV-based communication is the most critical challenge during the emergency duration. Usually, UAVs operate in Unlicensed bands & ISM bands. The unlicensed spectrum is also used by a rising number of wireless devices (smartphones, tablets & sensors). As a result, the unlicensed spectrum is getting congested, and unmanned aerial vehicles (UAVs) are experiencing interference & spectrum scarcity. So, Cognitive Radio is the best solution to utilize the beneficial use of unused spectrum. Cognitive Radio Technology has been launched for utilizing the radio spectrum resources more systematically; this point of view, Cognitive Radio-based UAVs will be considered a possible explanation for executing an essential Unmanned Aerial Vehicles Network.

A. Related Work

Various works have been published that are belonging to performance improvements in cognitive radio-based UAV networks. Many of these works are intended for improving communication in UAVs without evaluation of Channel Allocation, investigating channel allocation in UAVs even in prompt stages. Very few latest research denoted the problems are discussed here. In[1], authors have suggested Energy Efficient power allocation for UAV Cognitive Radio Systems. A hierarchical game model is designed for Dynamic multi-UAV Device to Device Communication Networks in [2]. Authors in [3] have proposed a method to improve performance to access idle spectrum. Quadratic transformation through Convex Functions (DC) optimization approach has been applied in [4] for UAV placement and Bandwidth Allocation. Authors in [5] have discussed about Spatial Spectrum Sensing in UAV Communication. Conventional Inter-Cell Interference Coordination (ICIC) techniques have been discussed for UAV sensing-assisted cellular interference [6]. The time schedule Optimization model for UAV Altitude has been proposed in [7]. Authors in [8], Optimal beam width Allocation method has proposed for Spectrum Sharing based Cognitive UAV Network. So far, it is visible from these literatures that existing methods have some limitations and channel allocation is negligible; our paper aims to develop an approach or proposal that Fuzzy Logic priority-based channel allocation technique as well as improve performance in Cognitive UAV Networks.

B. Paper Outline

This article is presented as follows. Section II describes the System Model. Section III discusses the formulation of the problem & the proposed algorithm. Section IV presents the simulation results. Section V ends with a conclusion.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

II. SYSTEM MODEL

UAVs, as before we mentioned, are expected to form a standard feature of upcoming aviation traffic. The performance of application features and functionality is dependent on the transmission of information between unmanned aerial vehicles (UAVs) and between unmanned aerial vehicles and ground control stations, which is further dependent on the use of aeronautical channels. When unmanned aerial vehicles (UAVs) are integrated with ground control stations and data links, a UAS (Unmanned Aerial System) is established. The unmanned aerial system (UAS) must be considered as a component of a massive system that includes the command, control, and communication system. Integrating unmanned aerial vehicles (UAVs) and other aerial devices for relaying connectivity is a promising phenomenon; the channel must be specified once communication links can be formed. Three different types of communication links are applied in the UAV Communication System: A2A (air-to-air), A2G (air-to-ground), and G2G (ground-to-ground).

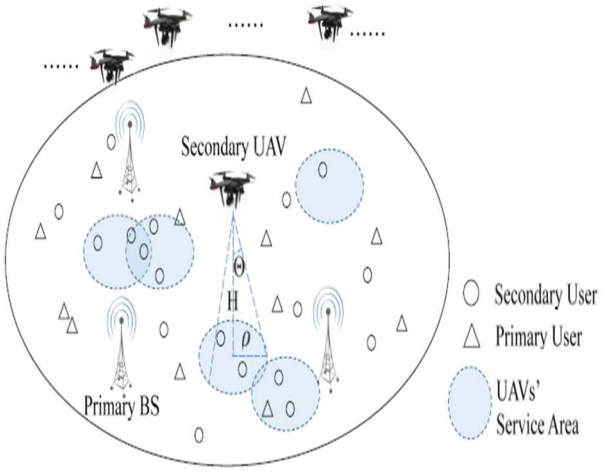


Fig. 1. Cognitive Radio based UAV Networks

Cognitive radio is described as an intelligent & adaptive wireless communication system that is aware of its surroundings and uses that awareness to learn from it and adapt to numerical changes in the input incitation. It enables the resolution of the dynamic spectrum allocation problem and the implementation of the flexible transition strategy for managing the air-ground radio system. The key goal of cognitive radio is to optimize the amount of available spectrum by utilizing cognitive capabilities and reconfigurability. After determining the operational spectrum band, communication may begin. However, since the radio environment is dynamic in terms of time, location, and frequency, cognitive radio devices maintain track of the radio environment's changes. If the currently used spectrum band becomes unavailable, the spectrum mobility feature is activated to ensure uninterrupted transmission. This modification may be triggered by any change in the environment that occurs during the communication, such as the primary user's appearance, mobility, or traffic fluctuation. In this respect, UAVs will significantly benefit from integration with Cognitive Radio Technology (CRT), owing to this technology's benefits, including dynamic spectrum access, decreased energy consumption & delay, and opportunistic spectrum usage depending on application needs. Cognitive radio-based UAVs use centralized and distributed methods to identify the optimum solution for spectrum sharing and position optimization in high-priority missions. While on a mission, all unmanned aerial vehicles (UAVs) are linked to the central controller, both directly and via a relay. The central controller allocates the channel and assigns the UAVs to the priority zones. For optimal positioning inside the operating zone, UAVs use real-time network observations.

III. PROBLEM FORMULATION & PROPOSED ALGORITHM

Bandwidth is a scarce resource in wireless communication system. Effective use of limited bandwidth is a critical technical challenge. In Cognitive Radio System, when a Secondary User is using a channel for transmission, and a Primary User appears, it is the SU that needs to stop transmission and return the channel to PU. SUs are dropped from the queue to access channels on the return of a PU. Every SU will have sensors using which they determine the existence of other available channels. These sensors look for what are the available channels, the channel quality, and the duration for which the channel is available. However, in the emergency situation, if there is no available free channel for SU, then the priority-based channel allocation technique is one of the possible solutions for that kind of emergency issues. Using the priority-based channel allocation technique in Cognitive Radio, a well-organized & synchronized communication system can be initiated for UAV Networks.

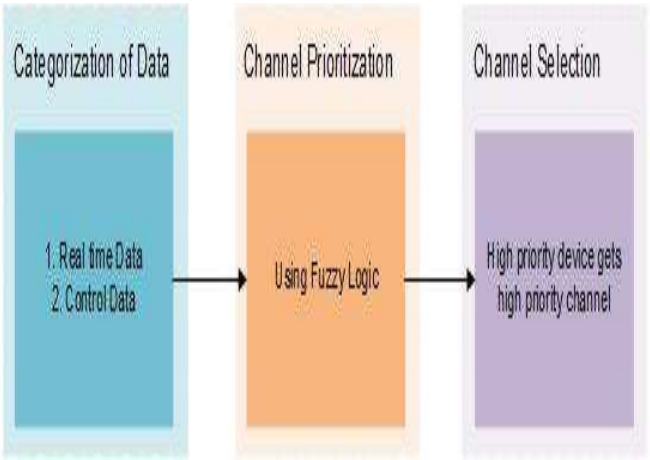


Fig. 2. Channel Allocation Mechanism

The proposed Algorithm consist of the following stages:

1) According to devices and channel usages, we have created two categories for selecting a scheduled priority-based channel allocation.

2) Then, we have created a category of devices that are belonging to the ground station & data fusion center. They are:

a) Real-Time Data (video or high-quality aerial images)

b) Control Data (Monitor, Data Analysis, Billing Information)

3) Then, we have prioritized the channels based on three parameters such as maximum usage of the channel (Maximum Usage), Packet Error Rate(PER) & Signal to Interference Noise Ratio(SINR).

4) This process is done by the Fuzzy-Logic approach, and after the use of the fuzzy logic algorithm, the channels are sorted according to priority. Later the channels are allocated

according to the following method such as higher priority device will get the higher priority channel, and the lower priority device will get the lower priority channel allocation.

IV. SIMULATION RESULTS

In order to verify the proposal & validate it by applying simulation, a well-known simulator MATLAB Simulink is used in this research. Furthermore, to assign the channel, we have developed a fuzzy logic system. This Fuzzy Interface System uses triangle membership functions and the Mamdani inference method, both of which are well-established techniques. The centroid method is used to implement the defuzzification process.

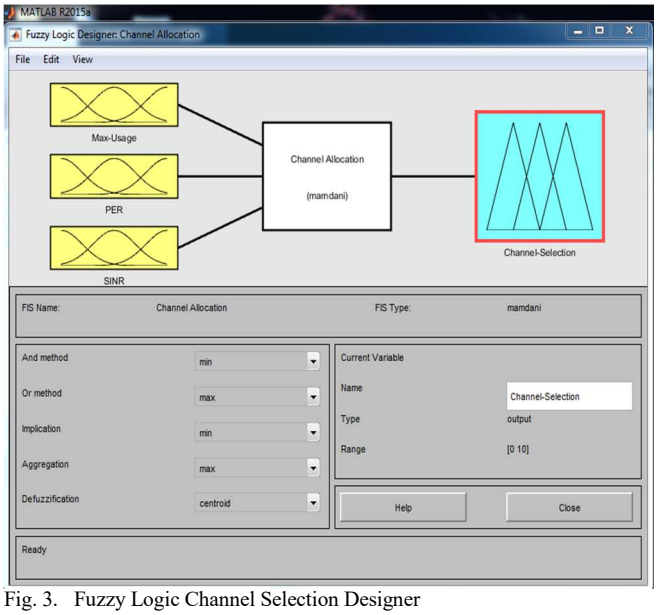


Fig. 3. Fuzzy Logic Channel Selection Designer

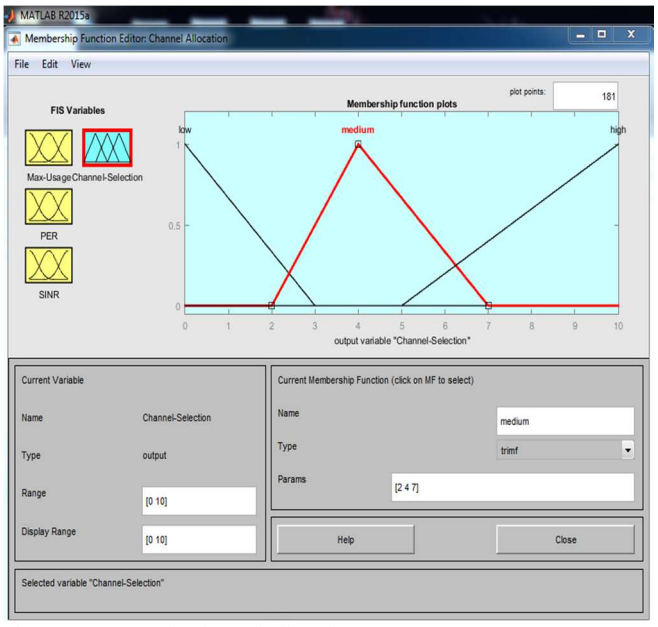


Fig. 4. Fuzzy Logic Channel Allocation Parameters

For this fuzzy logic algorithm paradigm, we have taken three input variables: the Maximum Usage of Channel (Max-Usage), the Packet Error Rate (PER), and the Signal to Interference Noise Ratio (SINR). Each input variable will display a different result depending on the rule condition that is being used. Fuzzy Logic allocates all results together in a scheme. Each input has a triangle membership function. Each triangle displays how the function behaves in three different frequency ranges: low, middle, and high. We have applied several AND logic rules here. As for the regarding rules, channel allocation has displayed a graph to select the best channel for SUs in Cognitive Radio based UAV scenario.

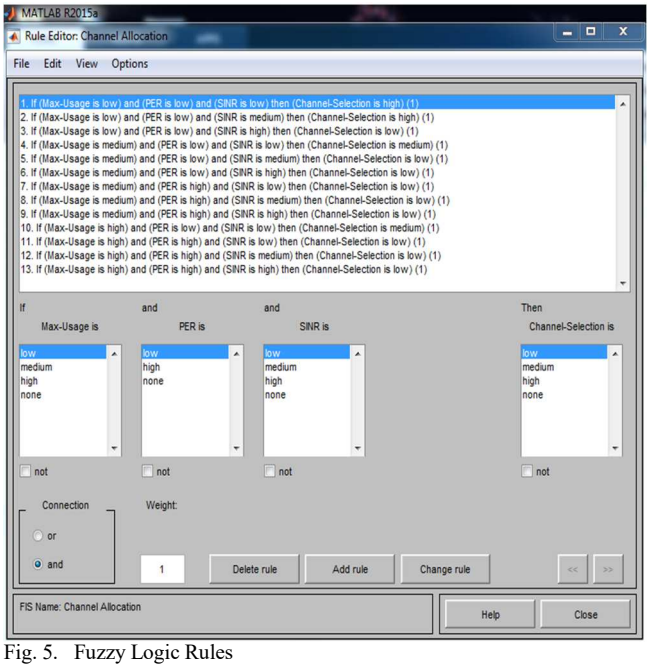


Fig. 5. Fuzzy Logic Rules

Applying thirteen AND rules in the Fuzzy Logic System, we have got various values by changing input parameters. From the output results, if we want to select the effective channel for SUs, then showing eight (8) or more than eight(8) values- will be the best for channel allocation.

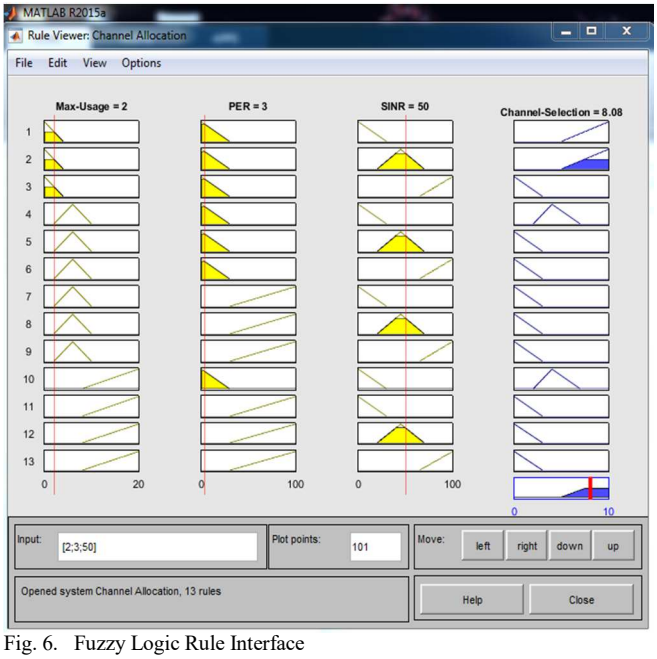


Fig. 6. Fuzzy Logic Rule Interface

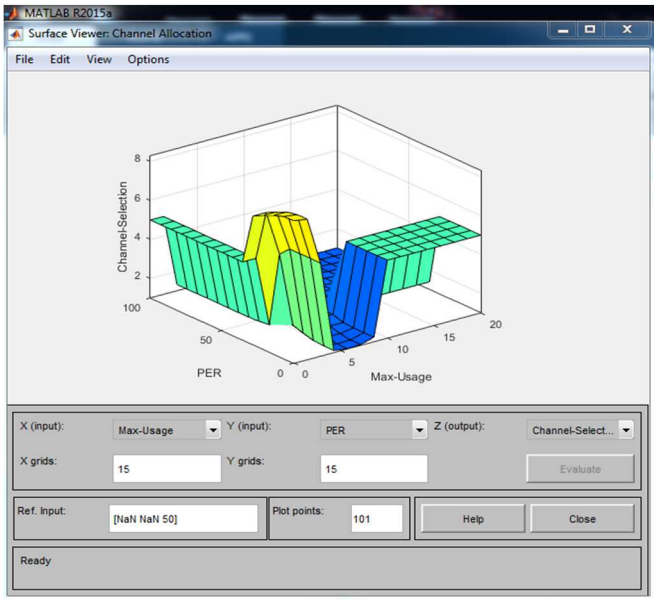


Fig. 7. Fuzzy Logic Surface view of Maximum Usage & PER Interface

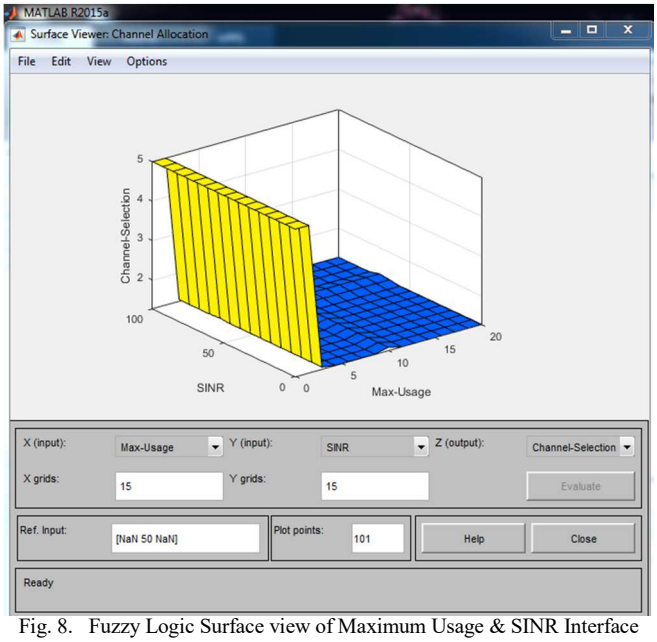


Fig. 8. Fuzzy Logic Surface view of Maximum Usage & SINR Interface

Applying the Fuzzy Logic Interface, we would like to establish the following formula - "higher priority device will get higher priority channel; lower priority device will get the lower priority channel." The simulation results indicate that our assigned method obtains the optimal channel allocation solution.

V. CONCLUSION

Channel Allocation on Cognitive Radio networks among SUs for a modern application has undoubtedly been an intention of interest for researchers. This research paper proposes the fuzzy logic algorithm method to further explore the channel allocation problem in CR Technology when assigned in UAV Networks. In our short knowledge, this is the first research paper applying the fuzzy logic algorithm to find out the best channel for Cognitive Radio SUs in UAV Network. We hope that our research will cover the way for future studies evaluating Channel Allocation in UAV for essential scenarios.

REFERENCES

[1] Lokman Sboui, Hakim Ghazzai, Zouheir Rezki and Mohamed-Slim Alouini "Energy-Efficient Power Allocation for UAV Cognitive Radio Systems" in 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall),Canada.

[2] Dianxiong Liu, Yuhua Xu, Jinlong Wang, Jin Chen, Qihui Wu, Alagan Anpalagan, Kun Xu and Yuli Zhang "Opportunistic Utilization of Dynamic Multi-UAV in Device-to-Device Communication Networks" in 2020 IEEE Transactions on Cognitive Communications and Networking Journal.

[3] Xin Liu, Mingxiang Guan, Xueyan Zhang and Hua Ding "Spectrum Sensing Optimization in an UAV-Based Cognitive Radio" in IEEE Access 2018.

[4] Minh Dat Nguyen, Tai Manh Ho, Long Bao Le, and Andr´e Girard "UAV Placement and Bandwidth Allocation for UAV Based Wireless Networks" in 2019 IEEE Global Communication Conference(GLOBECOM),USA.

[5] Bodong Shang, Vuk Marojevic, Yang Yi, Aly Sabri Abdallah, and Lingjia Liu " Spectrum Sharing for UAV Communications: Spatial Spectrum Sensing and Open Issues in IEEE Vehicular Technology Magazine, June 2020.

[6] Weidong Mei and Rui Zhang "UAV-Sensing-Assisted Cellular Interference Coordination: A Cognitive Radio Approach" in IEEE WIRELESS COMMUNICATIONS LETTERS, VOL. 9, NO. 6, JUNE 2020.

[7] Zheng Chu, Wanming Hao, Pei Xiao and Jia Shi "UAV Assisted Spectrum Sharing Ultra-Reliable and Low-Latency Communications" in 2019 IEEE Global Communication Conference(GLOBECOM),USA.

[8] Yue Ling Che, Sheng Luo and Kaishun Wu "Spectrum Sharing based Cognitive UAV Networks via Optimal beam width Allocation" in ICC2019-2019 IEEE International Conference on Communications(ICC),Shanghai, China.

[9] Jiapan Wen,Rong Yu,Yuhao Wang,Huilin Zhou,Fuhui Zhou "Resource Allocation and Trajectory Design for UAV-Enabled Wideband Cognitive Radio Networks" in 2019 IEEE Globecom Workshops(GC Wkshps),USA.

[10] Mohammad Mozaffari,Walid Saad,Mehdi Bennis,Young-Han Nam and Mérouane Debbah "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems" IEEE Communications Surveys & Tutorials (Volume: 21, Issue: 3, thirdquarter 2019).

[11] Haijun Wang,Haitao Zhao,Jiao Zhang,Dongtang Ma,Jiaxun Li andJibo Wei "Survey on Unmanned Aerial Vehicle Networks: A Cyber Physical System Perspective" in IEEE Communications Surveys & Tutorials (Volume: 22, Issue:2, Secondquarter 2020).

[12] Chaoxing Yan,Lingang Fu,Jiankang Zhang and Jingjing Wang "A Comprehensive Survey on UAV Communication Channel Modeling" in IEEE Access(V7) 2019.

[13] Tevfik Yücek and Hüseyin Arslan "A survey of spectrum sensing algorithms for cognitive radio applications" in IEEE Communications Surveys & Tutorials (Volume: 11, Issue: 1, First Quarter 2009).

[14] Grigore Stamatescu,Dan Popescu and Radu Dobrescu "Cognitive radio as solution for ground-aerial surveillance through WSN and UAV infrastructure" in Proceedings of the 2014 6th International Conference on Electronics, Computers and Artificial Intelligence (ECAI).

[15] Hector Reyes, Nickolas Gellerman, and Naima Kaabouch "A Cognitive Radio System for Improving the Reliability and Security of UAS/UAV Networks in 2015 IEEE Aerospace Conference.

[16] Lokman Sboui, Hakim Ghazzai, Zouheir Rezki and Mohamed-Slim Alouini "Achievable Rates of UAV-Relayed Cooperative Cognitive Radio MIMO Systems" in 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall),Canada.

[17] G. M. D. Santana,Rogers S. Cristo,Catherine Dezan,Jean-Philippe Diguet,Diana P. M. Osorio and Kalinka R. L. J. C. Branco "Cognitive Radio for UAV communications: Opportunities and future challenges" in IEEE 2018 International Conference on Unmanned Aircraft Systems (ICUAS).

Performance Analysis of Mobile Broadband Networks in Ibra City, Oman

Majan Abdullah Al Jahdhami
Department of Electronics and Communication Engineering, College of Engineering, A'Sharqiyah University (ASU),
Ibra 400, Oman
1605114@asu.edu.om

Ayman El-Saleh
Department of Electronics and Communication Engineering, College of Engineering, A'Sharqiyah University (ASU),
Ibra 400, Oman
ayman.elsaleh@asu.edu.om

Abdulraqeb Alhammadi
Department of Electronics and Communication Engineering, College of Engineering, A'Sharqiyah University (ASU),
Ibra 400, Oman
abdulraqeb@ieee.org

Ibraheem Shayea
Electronics and Communication Engineering Department, Faculty of Electrical and Electronics Engineering, Istanbul Technical University (ITU)
34467 Istanbul, Turkey
ibr.shayea@gmail.com

***Abstract*- Mobile broadband (MBB) is one of the critical goals in fifth-generation (5G) networks due to rising data demand. It provides very high-speed internet access with seamless connections. Existing MBB, including third-generation (3G) and fourth-generation (4G) networks, also requires monitoring to ensure good network performance. Thus, mobile network operators (MNOs) regularly monitor their network performance to meet user satisfaction. This paper presents a performance evaluation of existing MBB networks in Ibra city, Oman. The data measurements were collected through drive tests from two MNOs supporting 3G and 4G technologies: Omantel and Ooredoo. Several performance metrics are measured during the drive tests, such as signal quality, throughput (downlink and uplink), ping and handover rates. Measurement results demonstrated that the tested area has been covered with 4G networks and records an average throughput with downlink and uplink approximately 20 Mbps and 15 Mbps, respectively, with a minimum average ping and packet loss of 36.5 ms and 0.14, respectively, overall MNOs.**

Keywords—Quality of experience, QoE, quality of service, QoS, Mobile Broadband, 5G, data rate

I. INTRODUCTION

Mobile broadband (MBB) networks are growing very fast with supporting high-speed internet access. The demand for data is rapidly increasing due to a huge number of users accessing data through several cellular technologies and various types of internet services. Spectrum demand for mobile communication systems is growing due to the increase of wireless technologies and applications. Intelligent technologies are required to fully utilize and monitor current spectrums [1, 2]. Dynamic spectrum access uses to increase spectrum efficiency by adjusting radio resources. However, with an increasing load on existing networks and a user base more and more taking internet connection for granted, performance and quality measurement become essential for all concerned stakeholders. Therefore, 5G networks provide enhanced MBB, which supports high-speed data, video streaming with low latency and seamless mobility. Several factors affect real measurements of MBB performance, such as a variety of user devices, physical impairments, mobility and accessibility configuration [3]. The measurement experiment was carried on the end-users with the support application to collect various network data to verify the quality of service (QoS) and quality of experience (QoE). In [4], the authors highlight the importance of QoE in cellular networks with various radio access technologies (4G, 5G, and beyond), where they provide literature on the most advanced measurement methods in QoE. In addition, the QoE is further investigated by different metrics and models for web QoE estimation [5]. The web QoE helps mobile network operators (MNOs) understand their customers' usage service patterns, perceive quality, and point towards areas to improve. However, in research works [6, 7], the authors investigate the QoS and QoE by conducting an experimental study of the current MBB supporting 3G and 4G networks in Malaysia. A specific application installed on a smartphone handset collected the drive test data from several rural and urban regions. The measurement data of three MNOs are associated with several performance indicators such as coverage, latency, satisfaction, and speed for two MBB services: web browsing and video streaming. The work in [8] analyzed data measurement using several key performance indicators (KPIs) in 4G networks, such as signal quality and download throughput. The drive test considered the actual road traffic conditions at a vehicle speed of 30km/h. The experimental results demonstrated that the achieved throughput leads to different profiles in terms of time evolution.

In this paper, we investigate and evaluate the performance of existing MBBs (Omantel and Ooredoo) in Ibra City. Several KPIs were used to analyze the MBBs, such as signal quality, throughput, ping rate and handover rate. This investigation study provides valuable data measurements that can be used for future network improvements and pinpointing during the deployment of 5G networks. The rest of this paper is organized as

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

follows: Section II describes the methodology and experimental design. The performance of the MBBs is analyzed and evaluated in Section III. Section IV provides study limitations and 5G trends. Section V concludes the paper.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

This section presents the test methodology of the measurement campaign to evaluate the existing MNOs with several performance metrics relevant to the user experience. The data collection was carried on a commercial android application developed by Gyokov Solutions called "G-NetTrack" installed in two Samsung Galaxy handsets [9]. Besides, this application has been tested in terms of performance metrics, recording logfile, continuity of testing and steadability. Also, this application supports drive tests for outdoor scenarios and provides a wide range of features such as map visualization, cell scanning, loading cells, cell measurement for serving and neighbor cells. It has been used in numerous research papers, which considers a reliable application for collecting data measurements. Fig. 1 visualizes the information of data measurement over time.

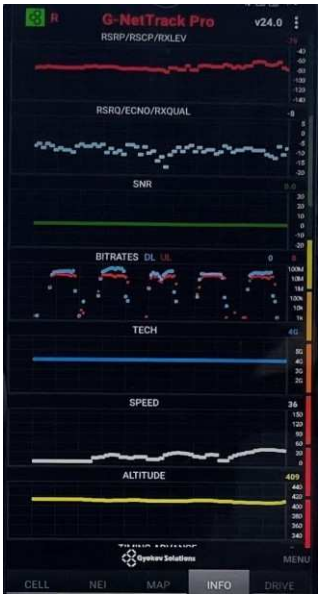


Fig. 1. Example of measurement information by G-NetTrack

Fig. 2. A general methodology of data collection and analysis

The data measurements were performed in Ibra city with coordinates of 22°41′N 58°33′E, and its population is approximately 163,473 people. In this regard, the MBB performance of two MNOs (Omantel and Ooredoo) is investigated and evaluated. However, this research aims to analyze and understand the MBB performance where it does not benchmark the performance of MNOs. Thus, the existing MNOs were labelled as X and Y in the discussed and demonstrated results. The prepaid subscriber identification module (SIM) cards were used each MNO with the same data package to ensure apple-to-apple comparison. Fig. 2 displays the general methodology of data measurements and analysis. The two mobile devices were fixed on a mobile phone stand holder inside a car as shown in Fig. 3. Then, the mobile devices collect the data from the MNO towers and then store them on the devices' memory that could be used for analysis after completing the data campaign. Fig. 4 displays the experimental testbed (measurement area) where the data measurements were taken on the orange route.

Fig. 3. Data measurement during the drive test.

Fig. 4. Tracking route of the measurement in Ibra City.

Fig. 5 shows the flowchart of the general data measurement. The drive test starts with setup the app setting several

parameters of data sequences such as ping time and URL, data rate time (downlink (DL) and uplink (UL)) and tested file with one gigabyte in size. Once these parameters are set, the drive test logs the data measurement from the starting point and stops at the ending points. The signal level and signal quality were measured continuously during the drive test. Besides, several events such as handover and cell reselection were recorded during the data measurements. All the above data measurements are associated with the timestamp and GPS coordinates (longitude and latitude).

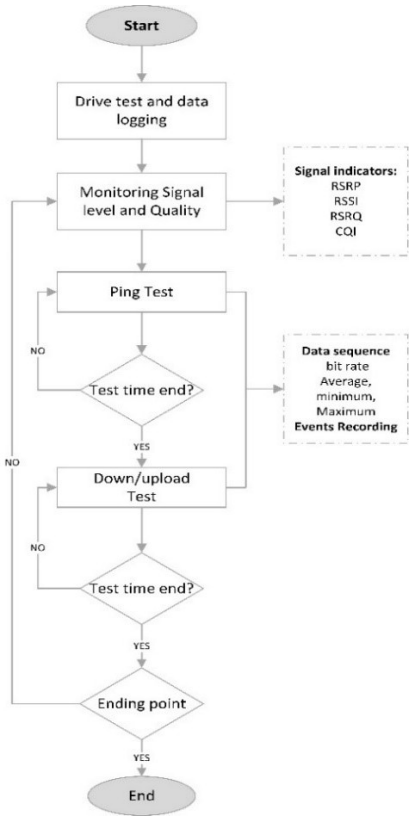


Fig. 5. Flowchart of measurement methodology

III. PERFORMANCE EVOLUTION

Existing MBBs evolution is paramount to ensure that the network provides a high user quality experience. This experimental research was conducted to analyze the actual performance of the current national MNOs in Ibra City. All the measurements were conducted at the same time, operating system and test sequences for a fair comparison. The measurement results are demonstrated and discussed for each KPI. The measurements were collected in the daytime for all MNOs. The car speed was limited up to 70 km/h for all outdoor scenarios. Fig. 6 illustrates the car speeds during the drive test. It can be observed that the speed is maintained below 70 km/h and sometimes reach 0 km/h at traffic lights.

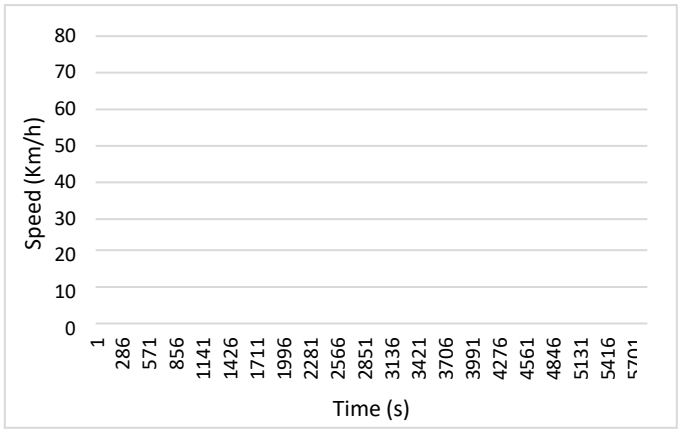
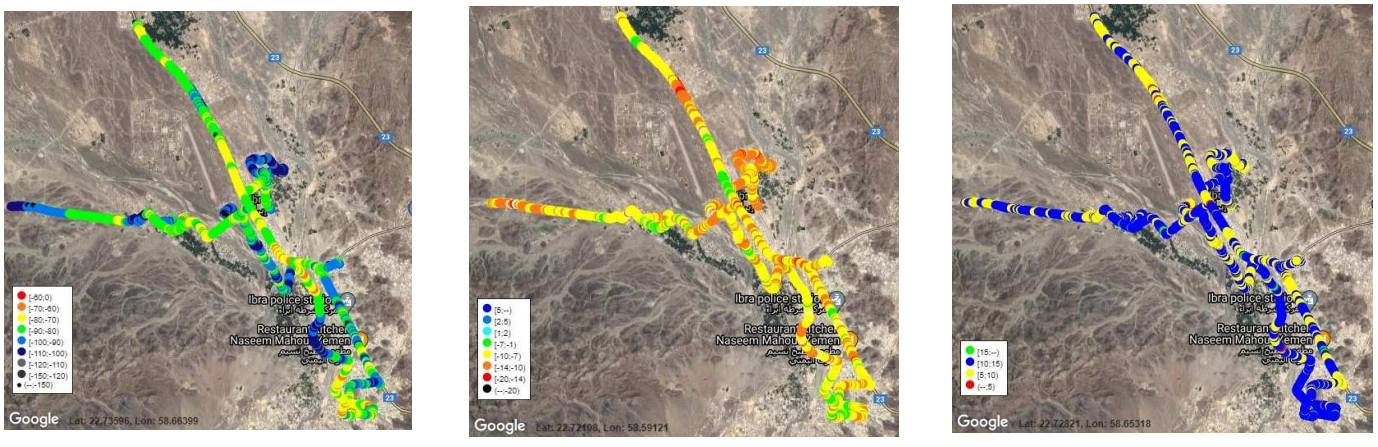


Fig. 6. Car speed versus time

Fig. 7 displays an example of data visualization of one operator only: reference signal received power (RSRP), reference signal received quality (RSRQ) and channel quality indicator (CQI). The route colors represent the measured values of each KPI. In the next subsections, the performance for each MNO has been analyzed and discussed according to these KPIs.



(a) (b) (c)

Fig. 7. Example of signal quality visualization (a) RSRP level, (b) RSRQ and (c) CQI

A. Signal Level and Quality

Fig. 8 illustrates the RSRP level and received signal strength indicator (RSSI) for the two MNOs. The average RSRP and RSSI levels for both operators are between -85 dBm and -83 dBm, and -76 dBm and -73 dBm, respectively. All the MNOs cover the test area with 4G networks.

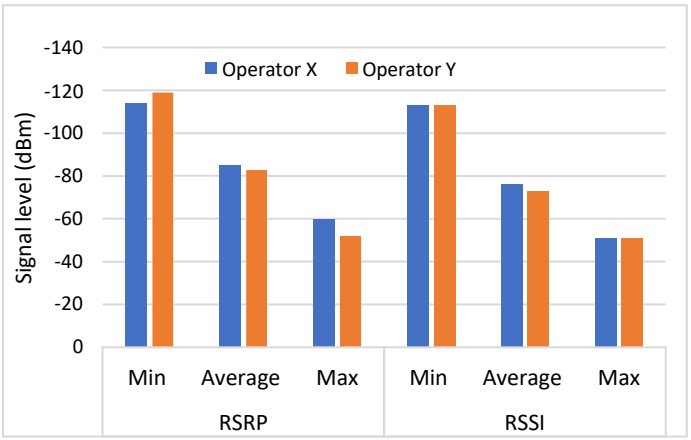


Fig. 8. Levels of RSRP and RSSI of the measured MNOs.

Fig. 9 displays the RSRQ level for the two MNOs. Here, many resource blocks with measured RSRP and RSSI are used over the same bandwidth for each MNOs. RSRQ value helps the base station to decide to perform cell reselection or intra-inter handover. The figure shows the average RSRQ level varying between -9.6 dB and -10.3 dB, where the RSRQ achieves a minimum and maximum value of -5 dB and -20 dB, respectively.

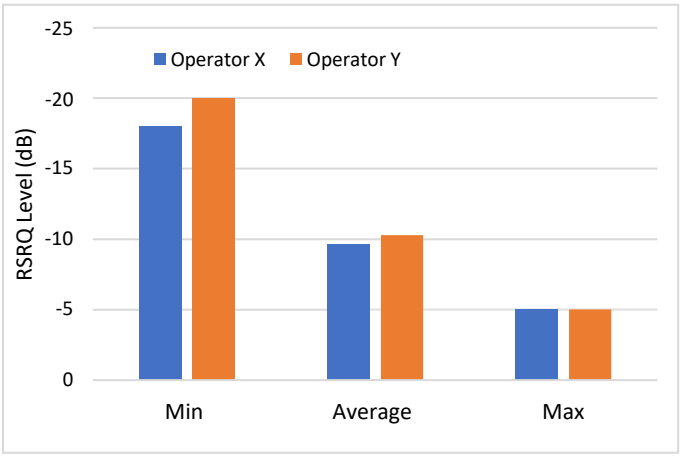


Fig. 9. RSRQ level of the measured MNOs.

Fig. 10 shows the statistics data of CQI for measured MNOs. The CQI is used to indicate the channel quality between mobile devices and evolved NodeBs (eNB). It can be seen that the average CQI levels vary between 10.8 and 8.6 for all MNOs. Both MNOs obtain a maximum value of CQI with 15. The high CQI level indicates to high data rate RSRQ. The CQI is a key metric for LTE systems, where MNOs usually use it to determine the relationship between radio link conditions and throughput.

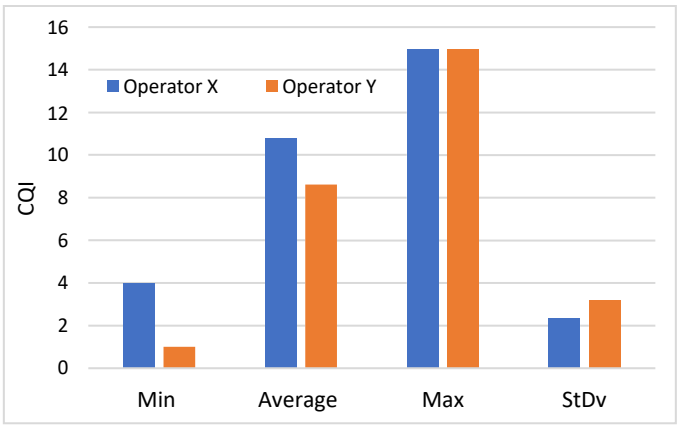


Fig. 10. CQI level of the measured MNOs.

B. Data Rats

Data rate analysis is used to evaluate the internet speed performance for a wireless network, known as connection speed and data transfer rate, and throughput. Fig. 11 displays the data rate: downlink and uplink over all MNOs. It can be seen that the maximum and minimum data rates in DL and UL are approximate 104 Mbps and 22 Mbps, respectively, for both MNOs. However, the average data rate in DL obtained by operator Y is higher compared with operator X.

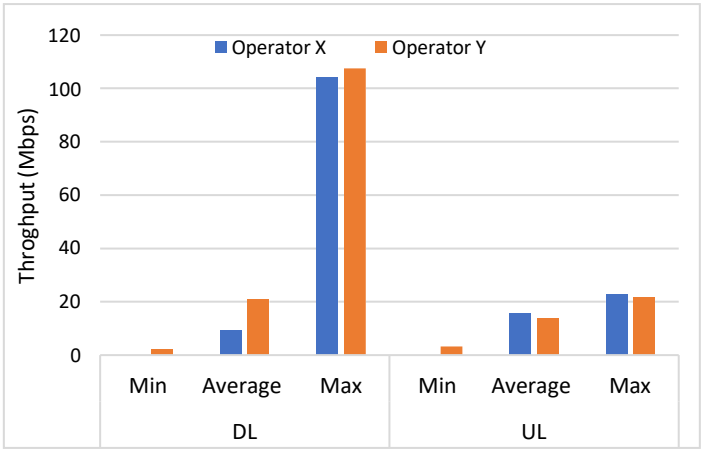


Fig. 11. Data rate (DL and UL) over the MNOs

C. Latency (Ping, Packet Loss)

Ping measures the reaction time of speed connection and recodes as ping count where a lower ping rate is better than a higher rate. In practice, the round-trip time of many deployed 4G networks tends to be in the 30–100 ms range once. Fig. 12 shows the ping rate achieved by the measured MNOs.

The results reveal that all the MNOs achieved an average ping rate lower than 50 ms, indicating a good speed connection. However, operator X obtains a slight rate of ping loss with 2.2 compared with operator Y.

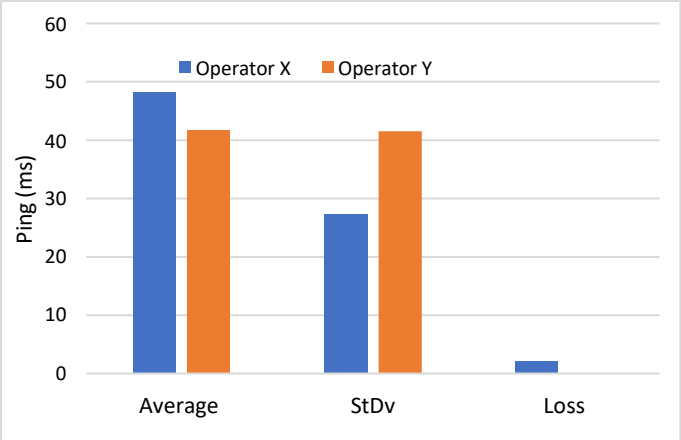


Fig. 12. Ping rate and packet loss over the MNOs

D. Handover

The major rising issue in mobile cellular networks deployment is user mobility, which may increase the handover rates and cause a radio link failure. These conditions directly affect communication quality in terms of long interruption time and throughput degradation [10]. There are two types of handovers: intra-frequency (Intra-eNB) and inter-frequency (inter-eNB) between legacy radio access technologies. The former occurs among base stations that operate in the same network technology. The latter occurs among base stations that operate in different network technologies. Fig. 13 shows the number of handovers for the two MNOs. In our measurement, the 4G networks were the dominant serving network with no recording of 3G networks. This means 4G networks fully cover the measured area. Thus, the recorded handovers are only among 4G networks (Intra-eNB handover). Also, operator X achieves a high number of handovers compared with operator Y.

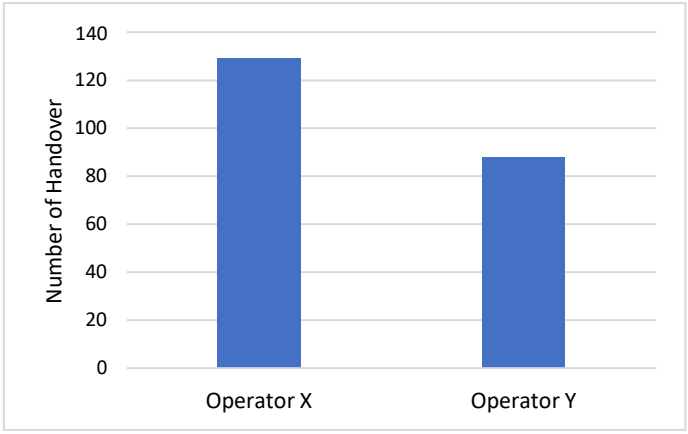


Fig. 13. Handover number over the MNOs.

E. Analysis Summary

The network performance metrics obtained from the two national MNOs of suburban morphology are summarized in Tables 1 and 2. Table 1 summarizes the signal quality RSRP, RSRQ and CQI for both operators. Table 2 summarizes throughput, ping rate and handover rate. Digits that are in bold type refer to best performance. All the MNOs cover the test area with 4G at measurement time with good signal quality and data rates. These demonstrated results of each operator depend on several factors such as bandwidth capacity, number of sites, number of active and type subscribers. These factors have a significant effect on the data rate and consequently degrades the network performance.

TABLE I. SUMMARY OF THE SIGNAL LEVELS AND QUALITIES

	RSRP (dBm)			**RSRQ (dB)**			**CQI**			
MNO	*Min*	*Avg.*	*Max*	*Min*	*Avg*	*Max*	*Min*	*Avg*	*Max*	*Sdv.*
X	-114	-84.90	-60	-18	**-9.62**	-5	4	**10.80**	15	2.34
Y	-119	**-82.70**	-52	-20	-10.27	-5	1	8.62	15	3.20

TABLE II. SUMMARY OF THE THROUGHPUT, PING RATE AND HANDOVER NUMBER

	Throughput (Mbps)				**Ping (ms)**			**Handover No.**
MNO	Avg DL	Avg UL	Max DL	UL Max	Avg	Stdev	Loss	Total
X	9.26	**15.54**	104.01	**22.77**	48.18	27.30	2.18	129
Y	**20.90**	13.80	**107.54**	21.53	**41.71**	41.53	**0.05**	**88**

IV. STUDY LIMITATIONS AND 5G TRENDS

A. Study Limitation

This investigation study presented the performance of existing MBB networks. Several limitations can be discussed as follows:

- Limited drive test: the data measurements were collected in one area, whereas larger areas are better but time-consuming. Thus, this work can be extended to various areas especially populated cities.
- Outdoor scenario: this study can be further extended to include the performance of MBB in an indoor environment like shopping malls with considering single and multi-floor scenarios. Indoor coverage would become much more important due to the high demands of modern usage cases. Thus, further investigation in indoor environments will provide an overview of how the MBBs perform in a complicated internal building's structure that can interfere with radio frequencies.
- One time drive test: this study was considered at the time of a peak hour. The peak time indicates when the network experiences heavily load due to the high demand required in this time. Hence, drive tests during peak-off hours can be included for further investigations.
- MBB services: this study is limited to two types of MBB services: web browsing and file (DL and UL) test due to application limitations. Other services tests such as video streaming and voice can be included in this study, but it will require an application to support these services.

B. 5G Trends

Oman telecommunications regulatory authority (TRA) allocates frequency band 3400 - 3600 MHz to the licensed MNOs to be used for 5G networks with a bandwidth of 100 MHz each MNO. Omantel and Ooredoo are allocated with frequency bands of 3400 - 3500 MHz and 3500 - 3600 MHz, respectively. Recently, both operators have deployed their 5G networks at the non-standalone stage. In addition, they have been launched super-fast and reliable 4G, and 5G fixed wireless (FWA) access. In a trial FWA demonstration, Omantel and Ericsson's ability to deliver multi-gigabit speeds of the FWA using a carrier bandwidth of 800 MHz at 26 GHz. Ooredoo and Nokia initially covered 3,000 homes in city centers with the 5G FWA services using Nokia equipment [11]. In addition, TRA Oman declared that a third operator mobile license (Vodafone) would be in operation soon to support and improve MBB performance.

V. CONCLUSION

This study has presented a performance analysis of the existing national MNOs in Oman. A drive test has been conducted in a suburban area, Ibra, Oman. The existing MNOs have been evaluated by using several performance metrics such as RSRP, RSRQ, CQI, throughput (DL and UL), ping rate and handover number. The results have shown that the area covers 4G networks and records an average throughput with DL and UL data rates of approximately 20 Mbps and 15 Mbps, respectively. The minimum average ping and packet loss of 36.5 ms and 0.14, respectively, overall MNOs. This leads to the current MNOs performing well in this drive test in outdoor environments where the test needs further investigation in indoor environments. In future work, this study can be further extended to include the performance of MBB in an indoor environment like shopping malls. More importantly, the performance of 5G MBB networks will be considered in our future researches.

ACKNOWLEDGMENT

The research leading to these results has received funding from The Research Council (TRC) of the Sultanate of Oman under the Block Funding Program with agreement no. TRC/BFP/ASU/01/2019.

REFERENCES

[1] A. Alhammadi, M. Roslee, and M. Y. Alias, "Fuzzy logic based negotiation approach for spectrum handoff in cognitive radio network," in *2016 IEEE 3rd International Symposium on Telecommunication Technologies (ISTT)*, 2016, pp. 120-124.

[2] M. Roslee, A. Alhammadi, M. Y. Alias, K. Anuar, and P. Nmenme, "Efficient handoff spectrum scheme using fuzzy decision making in cognitive radio system," in *2017 3rd International Conference on Frontiers of Signal Processing (ICFSP)*, 2017, pp. 72-75.

[3] B. J. Lobo, M. R. Alam, and B. E. Whitacre, "Broadband speed and unemployment rates: Data and measurement issues," *Telecommunications Policy,* vol. 44, p. 101829, 2020.

[4] K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, "Quality of experience for streaming services: measurements, challenges and insights," *IEEE Access,* vol. 8, pp. 13341-13361, 2020.

[5] H. Z. Jahromi, D. T. Delaney, and A. Hines, "Beyond first impressions: Estimating quality of experience for interactive web applications," *IEEE Access,* vol. 8, pp. 47741-47755, 2020.

[6] I. Shayea, M. H. Azmi, M. Ergen, A. A. El-Saleh, C. T. Han, A. Arsad*, et al.*, "Performance analysis of mobile broadband networks with 5g trends and beyond: Urban areas scope in malaysia," *IEEE Access,* vol. 9, pp. 90767-90794, 2021.

[7] I. Shayea, M. Ergen, M. H. Azmi, D. Nandi, A. A. El-Salah, and A. Zahedi, "Performance analysis of mobile broadband networks with 5G trends and beyond: Rural areas scope in Malaysia," *IEEE Access,* vol. 8, pp. 65211-65229, 2020.

[8] A. L. Imoize, K. Orolu, and A. A.-A. Atayero, "Analysis of key performance indicators of a 4G LTE network based on experimental data obtained from a densely populated smart city," *Data in brief,* vol. 29, p. 105304, 2020.

[9] G. Solutions. (03-02-2021). *G-NetTrack Pro*. Available: https://gyokovsolutions.com/g-nettrack/

[10] A. Abdulraqeb, R. Mardeni, A. Yusoff, S. Ibraheem, and A. Saddam, "Self-optimization of handover control parameters for mobility management in 4G/5G heterogeneous networks," *Automatic Control and Computer Sciences,* vol. 53, pp. 441-451, 2019.

[11] Ericsson, "Omantel and Ericsson successfully test 5G mmWave in Oman," ed: Press Releases, 2021.

A New Tele-Healthcare System of Heart Sound Signal Classification based on Gated Recurrent Unit

Eka Sari Oktarina
Dept. of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
ekasari91@mail.ugm.ac.id

Noor Akhmad Setiawan
Dept. of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
noorwewe@ugm.ac.id

Igi Ardiyanto
Dept. of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
igi@ugm.ac.id

***Abstract*— The heart has responsibility for pumping blood throughout the body. Heart health is essential to carry out daily activities. PCG is a method of monitoring heart signals that has been used for a long time, but it is rich in information on the characteristics of heart signals. Subjectivity in Phonocardiogram (PCG) heart signal analysis in making decisions when recognizing cardiac signals is a major issue since it might lead to fatal misinterpretation. Many investigations and developments have been made to classify the PCG heart sound signal so that objective examination of the PCG heart signal may be performed. The most essential feature of our contribution is evaluating the performance of the Gated Recurrent Unit to classify heart sound signals in a tele-healthcare system using CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) and Pearson Distance Metric as preprocessing methods. The result shows CEEMDAN algorithm and Pearson Distance Metric used in the decomposition process can separate murmurs and noise from the heart sound signal, while the Shannon Energy increases the quality of the signal. Based on our evaluation, GRU correctly classifies heart sound signals, and it can be seen from the precision produced, the accuracy of 87.2%, a precision of 79% for normal heart sound and 100% for abnormal signals.**

Keywords*— *phonocardiogram, classification, tele-auscultation, deep learning, gated recurrent unit.

I. INTRODUCTION

The human heart is the body's most vital organ, used to pump blood and carry nutrients to all body parts. However, in fact, in a survey conducted by the World Health Organization (WHO), as many as 31% of deaths were caused by coronary heart disease (CHD) [1]. Everyone needs to have regular heart checks. In addition, the advancement of information technology through the usage of the Internet of Things (IoT) has developed in the health sector [2]. It shows the future of tele-healthcare technologies, which are predicted to lower operating costs such as transportation, consultation, and hospital charges. Internet of Things (IoT) based healthcare might enable the sending of heart signals using internet technology to doctors in the isolated areas [3]. IoT technology can also be used to maintain the privacy of patients of different gender with doctors. Another advantage of utilizing IoT is that heart signal data storage can be done for visual analysis [4].

A published research result shows that models and techniques of heart signal transmission have been proposed in several articles, either through the use of the Zigbee project on wireless sensor networks, the use of Bluetooth, or WiFi [5]–[7]. Unlike previous research, with technological developments, making heart signal transmission can be done on a broader scope, namely by using the use of IoT technology carried out by Arijit in 2017, by sending a heart signal with noise that has been suppressed using the Maximum Inter technique-class Distant Maximum Intra-class Close (MIDMIC) over the internet [8]. Another study states that the heart signal to the fetus is successfully transmitted via the internet, making it easier for pregnant women to carry out examinations without the need to come to the hospital [9]. This proves that sending heart signals via the internet can be done even with more than 90% [10] Research continues to develop, making Bin Xiao in 2020 Conduct research on the classification of cardiac signals in neonates in order to detect congenital heart disease in an IoT-based device [11].

Three techniques can be used to carry out cardiac monitoring. An electrocardiogram (ECG) is the most popular technique used to record electronic activity in the heart, while photoplethysmography (PPG) is a technique for calculating blood flow rates [12] using a light sensor as the measurement instrument. Meanwhile, the phonocardiogram (PCG) is a method of recording heart sounds and murmurs that occur during the cardiac cycle. Different noises or beats are produced by cardiac activity. A typical heart sound consists of two beats known as the first (S1) and second (S2) heart sounds (S2). Another heart sound is known as the third heart sound (S3), and there is a fourth abnormal heart sound (S4), known as a murmur, which occurs when the heart is abnormal [13] (illustrated in Figure 1). The PCG and ECG contain more information than the PPG, and the level of complexity in a PCG makes it superior to an ECG in detecting properties of heart signals, especially in detecting murmurs that occur in abnormal heart sounds [14].

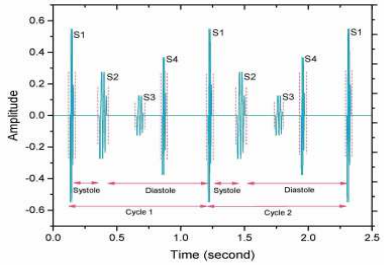


Fig. 1. Illustration of Heart Sound Signal with S1, S2, S3 And S4 [15]

Auscultation refers to the activity of listening for heart sounds, which is performed using an electronic or traditional stethoscope, an old but very effective method of diagnosing several cardiovascular diseases [16]. In the process, recording the heart signal can record the sound signal of the heart and the surrounding noise. The results of the examination, which are based on the doctor's hearing, are

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

also an obstacle in determining the results of heart examinations because they are the result of subjectivity. Then, the analysis of the detection of the characteristics of the heart sound signal automatically becomes very important to do so that there is no misdiagnosis when recording the heart sound signal [4].

Several studies on automatic PCG signal analysis to detect the location and grouping of S1 and S2 signals in normal heart signals and S3 and S4 of aberrant heart signals with diverse objects have resulted from the advancement of technology in classifying, machine learning, and deep learning. In 2016, Tschannen classified heart signals using the CNN algorithm and the physionet data as a dataset. The results revealed that the sensitivity, specification, and scoring levels were 0.96, 0.83, and 0.89, respectively [17]. Chen conducted research using the DNN method with a dataset derived from 311 training data and 16 people as test data in the following year. This study resulted in an accuracy rate of 91.12% [18]. In 2020, Gated Recurrent Unit (GRU) method proposed by Shan Gao can identify heart failure as an initial screening. the gru model can provide an average accuracy of 98% [19].

The complexity of the phonocardiogram signal becomes the biggest challenge in the classification process, so that in 2019, a PCG signal segmentation study was conducted [20]. These studies show promising results, but the time used in carrying out the process tends to be extended. Suyi Li has summarized much research on the classification of cardiac signals in 2018 [21].

Classification of heart sound signals requires several processes, including denoising and signal segmentation. This process is crucial because it will affect the following result in the classification process. Some work has been done in performing cardiac signals using the envelope extraction technique used in heart sound signals [3]. On the other hand, an algorithm called Empirical Mode Decomposition (EMD) is believed to be able to separate PCG signals contaminated by noise and noise by extracting a set of single-component signals [22]. In addition, the development of EMD was achieved by constructing a fully integrated empirical mode decomposition algorithm with adaptive noise to identify the presence of noise in the PCG called CEEMDAN.

This study proposes a novel approach for classifying heart sound signals in a smart healthcare system. The proposed technique comprises of a procedure which classifies heart sound signals using a Gated Recurrent Unit to distinguish between normal and pathological cardiac signals. The signal is sent via the internet to be stored in the cloud so that it can be displayed on a website that has specific access rights so that it can be used responsibly. The CEEMDAN algorithm separates the heart signal from the murmur. It is then segmented using the Shannon energy envelope method.

II. Related Works

The development of IoT technology in healthcare, at least two studies were carried out in 2018 to 2020 in sending heart signals by utilizing IoT technology. One of them is the manufacture of a hardware device used to retrieve heart signal data, then the signal attenuation process is carried out using the wavelet transform displayed on an android mobile which Jia Xin successfully developed in 2018 [23]. Furthermore, in 2020, Bin Xiao succeeded in monitoring heart signals in the fetus for early detection of abnormal heart signals in the fetus. Heart signals are sent via the internet network and stored in the cloud, and then the computer will automatically detect abnormal heart signals in the heart signals that have been sent [11].

In 2020, Suyi Li conducted a review of deep learning to classify heart signals in the last five years. With various classification techniques, there are still many contributions that can be made to improve the research that has been done [20]. Classifying heart signals using various methods has shown promising results, as was done by Sonoor Behbani in 2019. PCG signal segmentation with data from Physionet gave expected prediction results; abnormal S1, S2 was 98.98 percent, 98.78 percent, 98.78 percent, and 98.37 percent .The search process for S1 and S2 with a hybrid algorithm matches the ECG value [21].

Pre-processing is an important phase in the classification process, evidenced by the many studies that have been done before. As has been done by Ren (2018) uses Empirical Mode Decomposition in performing signal elimination to extract the characteristics of the PCG signal [24]. To detect murmurs in the PCG signal, the Complete Ensemble Empirical Mode Decomposition (CEEMD) approach was applied. [25]. In 2021, Jusak conducted a study identifying heart signals and proved that combining CEEMDAN and Pearson Distance Metric in separating murmurs and original heart signals could improve heart sound identification [26].

III. Proposed Model

Tele-Auscultation System Model of the research is shown in Figure 2. This model is similar to an auscultation model proposed in [27], but it is used for transmitting inner heart signals in the shape of the ECG signal. As shown in Figure 2, the system tele-auscultation was constructed using a model multipoint wireless sensor network data communication.

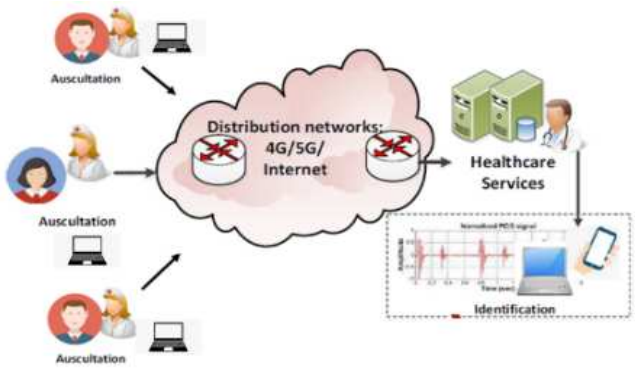


Fig. 2. Tele-Auscultation Design [4]

The main framework of this study is shown in Figure 3 and comprises three steps: pre-processing, segmentation and the classification model.

A. Preprocessing

The pre-processing stage has significant meaning because it will affect the next stage. At this stage, there is noise elimination because when taking heart sound signal data, there is some noise that appears, this is because the sound of the lungs, the sound around the recording environment, the sound of the movement of the stethoscope makes the recorded data have noise [27].

The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique is used in the first stage of pre-processing, as has been done in previous studies [26], [28]. The heart sound signal is decomposed through determination into a function number called the Intrinsic Mode Function (IMF). The IMF produces a single, time-varying frequency component that represents the entire heart sound signal [29], which can be systematically expressed as:

$$(n) = \sum_{k=1}^{K} \overline{IMF_k}(n) \quad (1)$$

Where is the $\overline{IMF_k}(n)$ represent the k-th IMF, and $\overline{IMF_k}(n)$ represent the trend-like final residue. The average of k-th IMF over an ensemble of I observed from signals contaminated by White Gaussian Noise, which noise is controlled by amplitude, is the IMF_k component. The heart sound signal is defined as an N-samples signal and denoted as $\{s(n), n = 1, \dots, N\}$, this is obtained by analyzing the stethoscope's heart signal. The use of CEEMDAN to decompose signals should improve the possibility of separating the S1 and S2. Heart sound signals with unwanted components, such as noise or murmurs. This is very important because in addition to extracting anomalies in the heart sound signal, which are primarily non-stationary signals, but also helpful in separating normal heart sound signals from noise during the auscultation process

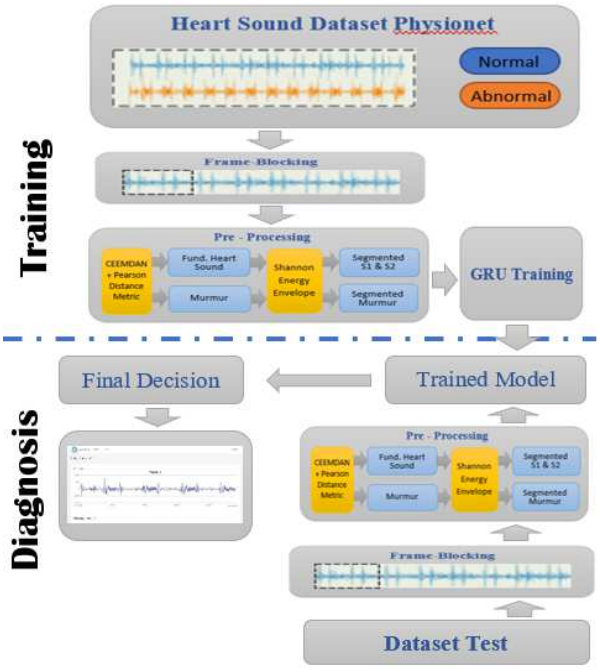


Fig. 3. Design of PCG Heart Sound Signal Classification

B. Heart Sound Separation

However, separating the fundamental heart sound from the murmur or noise is very challenging. The signal will be analyzed using an algorithm known as Pearson distance metric, which was introduced [25] in extracting heart sound signals with murmurs and sounds and is highly significant in detecting the accuracy of the splitting of heart sound signals. This is very difficult to do because of the process of distinguishing the original heart signal and murmur in an automated system, and the Pearson distance metric can do it well [25].

The Pearson distance metric that differentiates IMFs of heart sound, can be expresses as:

$$d_k = \frac{cov(s(n), \overline{IMF_k}(n))}{\sqrt{cov(s(n)).cov(\overline{IMF_k}(n))}} \quad (2)$$

Where d_k is the correlation coefficient related with $\overline{IMF_k}(n)$. Equation (2) describe that the correlation coefficient between the original signal and the $\overline{IMF_k}(n)$ is cross-correlation. So, it can be said that the correlation represents the degree of similarity between the original signal and each $\overline{IMF_k}(n)$. A normalized correlation coefficient is described in $0 \leq |d_k| \leq 1$. Furthermore, the Pearson distance metric proposes to describe the degree of similarity.

$$P_k = 1 - |d_k| \quad (3)$$

The Pearson distance metric (P_k) (3) results show that the smaller the value of the Pearson distance metric shows the close resemblance between the original signal and the $\overline{IMF_k}(n)$. The greater the Pearson value indicates that the two show different things.

C. Shannon Energy Envelope

As a result of the pre-processing phase, the signal will be segmented using the Shannon Energy method. The most popular technique of envelope extraction is Shannon energy. There are four-step to use the Shannon envelope after process pre-processing. Interested readers are encouraged to refer [30] for further explanation related to the Shannon energy.

Step 1. After the pre-processing phase, Sampling frequency decimated the separated fundamental heart sound signals, y_{FS}, by a factor of 4 of order 30.

Step 2. Signal normalization according to (1) so that the variance of the signal is set to 1.

$$y_{FS-norm}(n) = \frac{y_{FS-dec}(n)}{\max_i |y_{FS-dec}(i)|} \quad (4)$$

where $y_{FS-norm}$ are signal normalization, y_{FS-dec} are signal decomposition.

Step 3. Average Shannon energy (E_s) calculation based on (5)

$$E_s = -\frac{1}{N}\sum_{i=1}^{N}(y^2_{FS-norm}(i).\log y^2_{FS-norm}(i)) \quad (5)$$

N is several signal series in every 20 ms. Signal overlapped 10 ms. Then N = 40

Step 4. Calculate of average Shannon energy normalization (P_{avg}) according to (6)

$$P_{avg}(n) = \frac{E_s(n) - M(E_s(n))}{S(E_s(n))} \quad (6)$$

where M and S are mean value and standard deviation.

D. Heart Sound Signal Classification

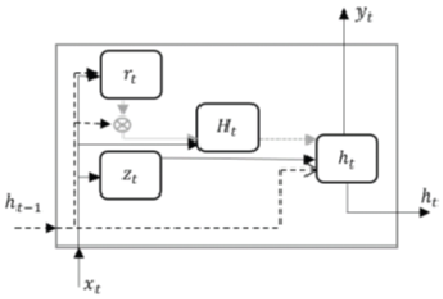


Fig. 4. Gated Recurrent Unit [31]

And the last step, the signal will be clasificated by GRU. The GRU method to analyze heart failure carried out by San Gao in 2020 shows that the algorithm can detect it with high accuracy [19].

GRU simplifies the LSTM structure by making the node to have two gates namely, reset gate ($r^{(t)}$) and update gate ($z^{(t)}$). The update gate determines how much the unit is updated when it is activated, which can be explained in equation (7). The reset gate serves to clear the memory in the previous computation, which is calculated in equation (8).

$$z^{(t)} = \sigma\left(b_z + U_z x^{(t)} + W_z h^{(t-1)}\right) \quad (7)$$

$$r^{(t)} = \sigma\left(b_r + U_r x^{(t)} + W_r h^{(t-1)}\right), \quad (8)$$

while the hidden layer ($h^{(t)}$) is described as follows:

$$h^{(t)} = z^{(t)} h^{(t-1)} + \left(1 - z^{(t)}\right)\tilde{h}^{(t)} \quad (9)$$

Where b represents the bias and U, W are the weight matrices of different gate referring to the subscripts.

Designing the structural design of the compilation unit is very important, and this process will be used as a reference model in learning. The design of the model has a model structure that is arranged to meet the required output weight value. The success of the model can be seen from the values of accuracy and loss. The accuracy value is the value used to determine the success of the model. The loss value is a measure of structural error, and the goal is to minimize it. The desired accuracy result is above 80% because, with this value, the classification process is considered successful. This study used a various hidden layer (HL), epoch and batch sizes to find the best model to the classification of PCG signal has been described in Table 1.

TABLE I. INPUT HIDDEN LAYER

Test	Input Layer	HL 1	HL 2	HL 3	HL 4	HL 5
1	200	2048	2048	1024	2048	2048
2	200	2048	2048	2048	1024	512
3	200	1024	2048	2048	1024	512
4	200	512	1024	2048	1024	512

IV. RESULT AND DISCUSION

In this section, we will use various PCG signals to check the performance of the proposed method. These signals are taken from the physionet database, which is freely accessible on the Internet. The author conducted this study using 2969 data, consisting of 2372 normal PCG signals and 579 abnormal PCG signals. In addition, the data is divided into training, validation and test data sets at a ratio of 80:10:10. Firstly, we We process the heart sound signal data before preprocessing, x(n), as a series of 8000 samples with frequency sampling of 8KHz from Physionet database.

The heart sound signal decomposed using CEEMDAN algorithm resulted in the extraction of the original PCG signal (a0007) with noise and a murmur, as shown in figure 5. In the figure 5, the normal heart sound decomposes into 11 independent IMFs. The index at the IMF shows the effect of the oscillating signal; the lower the index indicates the speed of oscillation at the IMF, on the other hand, the higher the IMF index indicates slower oscillations.

The fourth line of Figure 6 show a two-cycle reconstruction heart sound signal is taken from PhysioNet 2016 database challenge (signal a0007). It is extracted the fundamental heart sound in the fourth row is the result of aggregation of $\overline{IMF_1}(n)\ldots\overline{IMF_2}(n)$ corresponds to Pearson distance metric $P_k \leq 0.8$ as shown in the second row of Figure 6.

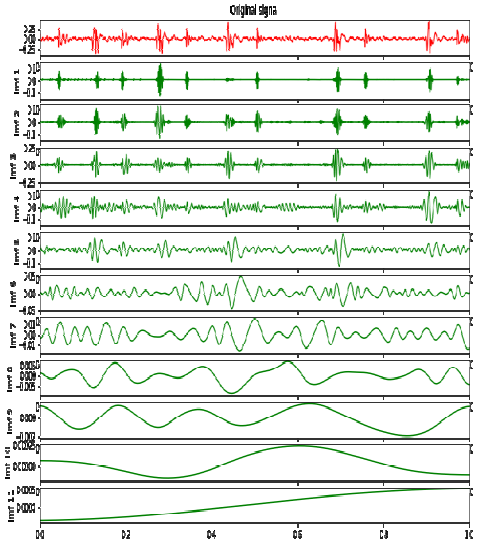


Fig. 5. Decomposition of Normal Heart Sound using CEEMDAN

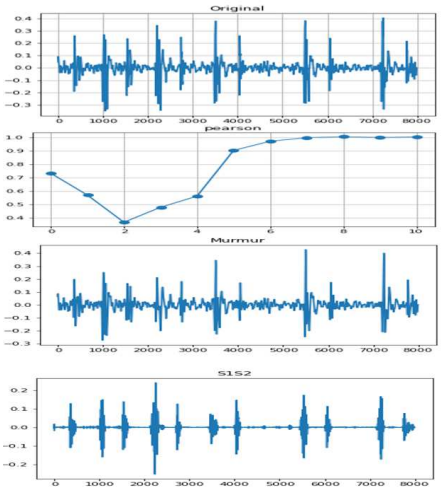


Fig. 6. First row the original heart sound, second row is the Pearson Distance Metric, and the third and fourth row are separated fundamental heart sound from noise and murmur

Figure 8 then the heart signal was segmented using the Shannon Energy Envelope normalization from normal heart sound signal in Figure 7. This algorithm aims to facilitate the detection of fundamental heart sound signals from murmurs and noises.

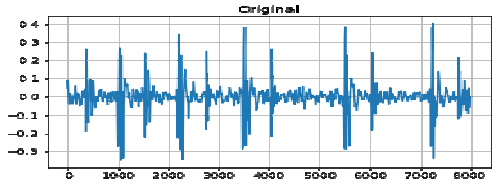


Fig. 7. Normal Heart Sound Signal (a0007)

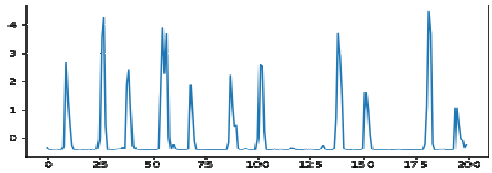


Fig. 8. Result of Sannon Energy Envelope (Normal Signal-a0007)

The results of the segmentation of the PCG signal are used as input for the heart signal classification process using the GRU. Several calculations and observations are made while the training process is running, including accuracy, data loss, and processing time. Test results with several epochs values (100-500), 200 input layers, and hidden layer variations as shown in Table 1. The training results table taken based on a predetermined model can be seen in Table 2.

Table 2 shows the training results of 500 epochs where the final value of the process is in the form of a train accuracy value of 0.979 and a test accuracy of 0.872. If the training process above is plotted in a graph, it will appear as shown in the figure 9. From the graph in Figure 9, the results of the accuracy and loss functions are still fluctuating because the training accuracy value is higher than the accuracy value with the test data.

There are several parameters to measure the success rate of a study. The first parameter is measured by the success rate of the GRU method in classifying data. The success rate of a classification in the medical field is seen from the level of accuracy, sensitivity and specification of the classification results.

The measurement is seen from doing an average of the sensitivity and specifications used to measure the performance of a medical classification method. These measures were used to assess the appropriateness of the proposed method [32]. The success of the training model on PCG signal classification based on the database is proven from the confusion matrix.

The best results of the classification using GRU are conducted on a computer with Intel core I5 processor specifications, 16GB RAM, operating on Windows 10 pro-64 bit, with NVIDIA 1650 4GB. Produces an accuracy of 87.2%, a precision of 79% for normal heart sound and 100% for abnormal signal by using 500 global steps in 179 minutes as revealed in Table 3.

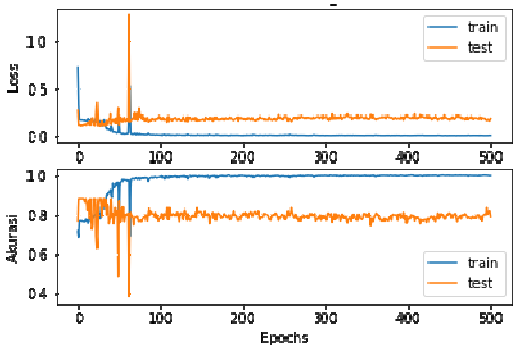


Fig. 9. Result of Training

TABLE II. TRAINING RESULT

Epoch	Accuracy (%)	Loss	Val_Acc	Val_Loss	Time Process (M)
Test 1					
100	0.777	0.222	0.939	0.062	96.42
200	0.8344	0.155	0.967	0.031	120.77
300	0.855	0.15	0.79	0.028	189.107
400	0.861	0.134	0.975	0.023	232.81
500	0.855	0.133	0.978	0.021	193.82
1000	0.858	0.136	0.979	0.019	374.29
Test2					
100	0.845	0.15	0.977	0.022	82.3
200	0.851	0.14	0.978	0.02	111.21
300	0.868	0.129	0.975	0.024	145.07
400	0.855	0.144	0.979	0.02	188.76
500	0.872	0.129	0.979	0.019	176.99
1000	0.865	0.129	0.978	0.021	101.65
Test3					
100	0.855	0.141	0.976	0.022	65.24
200	0.865	0.133	0.976	0.023	63.22
300	0.868	0.128	0.976	0.022	114.31
400	0.858	0.141	0.981	0.018	153.11
500	0.865	0.128	0.979	0.019	123.07
1000	0.865	0.13	0.981	0.019	361.22
Test4					
100	0.807	0.245	0.947	0.079	39.82
200	0.848	0.263	0.969	0.042	53.29
300	0.841	0.157	0.973	0.031	81.11
400	0.858	0.141	0.978	0.02	70.26
500	0.855	0.138	0.979	0.019	87.04
1000	0.865	0.132	0.965	0.021	183.75

TABLE III. METRIC PERFORMANCE OF CLASSIFICATION

	Precision	Recall	F1-Score	Support
Normal	0.79	1.00	0.88	146
Abnormal	1.00	0.75	0.85	150
Accuracy			0.87	296
Macro Avg	0.9	0.87	0.87	301
Weight Avg	0.9	0.87	0.87	301

V. CONCLUSION

Classification of phonocardiogram signals using the Gated Recurrent Unit on a tele-auscultation has been built. The classification results show that the Gated Recurrent Unit can classify heart sound signals well, this can be seen from the precision produced, accuracy of 87.2%, a precision of 79% for normal heart sound and 100% for abnormal signal by using 500 global steps. The results obtained indicate that the use of the CEEMDAN method used in the decomposition process is able to separate murmurs and noise from the heart

sound signal with the average time process 3.39 minutes, and the Shannon Energy envelope is able to perform the segmentation process very well with the average time process 0.006s. Our future research to combine the CEEMDAN and Blind Source Separation (BSS) to identify the fundamental of heart sound signal and murmur as a pre-processing signal.

ACKNOWLEDGMENT

This work has been fully supported by Lembaga Pengelola Dana Pendidikan (LPDP) and Departement of Electrical Engineering and Information Technology Universitas Gadjah Mada. We also acknowledge great support from Dr. Jusak for his suggestions in doing this research.

REFERENCES

[1] T. Khan, "Cardiovascular Diseases," *WHO*, 2020. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

[2] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015, doi: 10.1109/ACCESS.2015.2437951.

[3] I. Puspasari, W. I. Kusumawati, E. S. Oktarina, and J. Jusak, "A New Heart Sound Signal Identification Approach Suitable for Smart Healthcare Systems," *2019 2nd Int. Conf. Appl. Eng.*, pp. 1–6, 2019, doi: 10.1109/ICAE47758.2019.9221752.

[4] E. S. Oktarina, I. Puspasari, and J. Jusak, "Auskultasi Jarak Jauh untuk Pengukuran dan Perekaman Sinyal Suara Jantung," *J. Rekayasa Elektr.*, vol. 14, no. 3, 2018, doi: 10.17529/jre.v14i3.12013.

[5] J. Jusak and I. Puspasari, "Wireless tele-auscultation for phonocardiograph signal recording through Zigbee networks," *APWiMob 2015 - IEEE Asia Pacific Conf. Wirel. Mob.*, pp. 95–100, 2016, doi: 10.1109/APWiMob.2015.7374939.

[6] B. A. Izneid, I. Sukar, M. Ali, and M. Souiyah, "Development of wireless Bluetooth heart rate remote monitoring system," *IET Conf. Publ.*, vol. 2012, no. 614 CP, 2012, doi: 10.1049/cp.2012.2101.

[7] J. Liu, Y. Chen, Y. Wang, X. Chen, J. Cheng, and J. Yang, "Monitoring Vital Signs and Postures during Sleep Using WiFi Signals," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2071–2084, 2018, doi: 10.1109/JIOT.2018.2822818.

[8] A. Ukil and U. K. Roy, "Smart cardiac health management in iot through heart sound signal analytics and robust noise filtering," *IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, vol. 2017-Octob, pp. 1–5, 2018, doi: 10.1109/PIMRC.2017.8292659.

[9] A. Bin Queyam, R. Kumar Meena, S. Kumar Pahuja, and D. Singh, "An IoT Based Multi-Parameter Data Acquisition System for Efficient Bio-Telemonitoring of Pregnant Women at Home," in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2018, pp. 14–15, doi: 10.1109/CONFLUENCE.2018.8442686.

[10] A. Gharehbaghi, A. A. Sepehri, and A. Babic, "An Edge Computing Method for Extracting Pathological Information from Phonocardiogram," *Stud. Health Technol. Inform.*, vol. 262, pp. 364–367, 2019, doi: 10.3233/SHTI190094.

[11] B. Xiao *et al.*, "Follow the Sound of Children's Heart: A Deep-Learning-Based Computer-Aided Pediatric CHDs Diagnosis System," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1994–2004, 2020, doi: 10.1109/JIOT.2019.2961132.

[12] L. M. Sepulveda-Cano, E. Gil, P. Laguna, and G. Castellanos-Dominguez, "Selection of nonstationary dynamic features for obstructive sleep apnoea detection in children," *EURASIP J. Adv. Signal Process.*, vol. 2011, 2011, doi: 10.1155/2011/538314.

[13] S. Ismail, I. Siddiqi, and U. Akram, "Localization and classification of heart beats in phonocardiography signals —a comprehensive review," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, 2018, doi: 10.1186/s13634-018-0545-9.

[14] W. Phanphaisarn, A. Roeksabutr, P. Wardkein, J. Koseeyaporn, and P. Yupapin, "Heart detection and diagnosis based on ECG and EPCG relationships," *Med. Devices Evid. Res.*, vol. 4, no. 1, pp. 133–144, 2011, doi: 10.2147/MDER.S23324.

[15] S. C. S, A. H. Salman, N. Ahmadi, R. Mengko, A. Z. R. Langi, and T. L. R. Mengko, "Automatic Segmentation and Detection of Heart," pp. 103–107, 2015.

[16] D. Setiawan, A. Surtono, and S. W. Suciyati, "Ekstraksi Ciri Suara Jantung Menggunakan Metode Dekomposisi dan Korelasi Sinyal (Dekorlet) Berbasis Jaringan Syaraf Tiruan," *J. Teor. dan Apl. Fis.*, vol. 03, 2015, doi: 10.23960/JTAF.V3I1.1284.

[17] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," *Comput. Cardiol. (2010).*, vol. 43, pp. 565–568, 2016, doi: 10.22489/cinc.2016.162-186.

[18] T. E. Chen *et al.*, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, 2017, doi: 10.1109/TBME.2016.2559800.

[19] S. Gao, Y. Zheng, and X. Guo, "Gated recurrent unit-based heart sound analysis for heart failure screening," *Biomed. Eng. Online*, vol. 19, no. 1, pp. 1–18, 2020, doi: 10.1186/s12938-020-0747-x.

[20] S. Behbahani, "A hybrid algorithm for heart sounds segmentation based on phonocardiogram," *J. Med. Eng. Technol.*, vol. 43, no. 6, pp. 363–377, 2019, doi: 10.1080/03091902.2019.1676321.

[21] S. Li, F. Li, S. Tang, and W. Xiong, "A Review of Computer-Aided Heart Sound Detection Techniques," *Biomed Res. Int.*, vol. 2020, 2020, doi: 10.1155/2020/5846191.

[22] M. A. COLOMINAS, G. SCHLOTTHAUER, M. E. TORRES, and P. FLANDRIN, "Noise-Assisted Emd Methods in Action," *Adv. Adapt. Data Anal.*, vol. 04, no. 04, p. 1250025, 2012, doi: 10.1142/s1793536912500252.

[23] N. Giordano and M. Knaflitz, "A novel method for measuring the timing of heart sound components through digital phonocardiography," *Sensors (Switzerland)*, vol. 19, no. 8, 2019, doi: 10.3390/s19081868.

[24] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A Novel Cardiac Auscultation Monitoring System Based on Wireless Sensing for Healthcare," *IEEE J. Transl. Eng. Heal. Med.*, vol. 6, no. c, p. 1, 2018, doi: 10.1109/JTEHM.2018.2847329.

[25] Y. YIN, Y. Zeng, X. Chen, and Y. Fan, "The internet of things in healthcare: An overview," *J. Ind. Inf. Integr.*, vol. 1, pp. 3–13, 2016, doi: 10.1016/j.jii.2016.03.004.

[26] J. JUSAK, I. PUSPASARI, and W. I. KUSUMAWATI, "A Semi-automatic Heart Sounds Identification Model and Its Implementation in Internet of Things Devices," *Adv. Electr. Comput. Eng.*, vol. 21, no. 1, pp. 45–56, 2021, doi: 10.4316/aece.2021.01005.

[27] D. Gradolewski, G. Magenes, S. Johansson, and W. J. Kulesza, "A wavelet transform-based neural network denoising algorithm for mobile phonocardiography," *Sensors (Switzerland)*, vol. 19, no. 4, pp. 1–18, 2019, doi: 10.3390/s19040957.

[28] P. Flandrin, E. Torres, and M. A. Colominas, "A COMPLETE ENSEMBLE EMPIRICAL MODE DECOMPOSITION Laboratorio de Se ˜ nales y Din ´ amicas no Lineales , Universidad Nacional de Entre R ´ Laboratoire de Physique (UMR CNRS 5672), Ecole Normale Sup ` erieure de Lyon , France," pp. 4144–4147, 2011.

[29] N. E. Huang *et al.*, "The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998, doi: 10.1098/rspa.1998.0193.

[30] Y. Liu, C. C. Y. Poon, and Y. T. Zhang, "Heart sound segmentation algorithm based on heart sound envelogram," *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS'08 - "Personalized Healthc. through Technol.*, vol. 24, pp. 1308–1310, 2008, doi: 10.1109/iembs.2008.4649404.

[31] T. T. H. Le, J. Kim, and H. Kim, "Classification performance using gated recurrent unit Recurrent Neural Network on energy disaggregation," *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 1, no. July, pp. 105–110, 2016, doi: 10.1109/ICMLC.2016.7860885.

[32] A. Gharehbaghi, M. Borga, B. J. Sjöberg, and P. Ask, "A novel method for discrimination between innocent and pathological heart murmurs," *Med. Eng. Phys.*, vol. 37, no. 7, pp. 674–682, 2015, doi: 10.1016/j.medengphy.2015.04.013.

Diagnosis of Aglaonema Plant Disease Using Forward Chaining and Naive Bayes Methods

Heliza Rahmania Hatta
Department of Informatics
Mulawarman University
Samarinda, Indonesia
heliza_rahmania@yahoo.com

Riswandi Syam
Department of Informatics
Mulawarman University
Samarinda, Indonesia
wandyrs88@gmail.com

Dedy Cahyadi
Department of Informatics
Mulawarman University
Samarinda, Indonesia
dedy.cahyadi@unmul.ac.id

Anindita Septiarini
Department of Informatics
Mulawarman University
Samarinda, Indonesia
anindita@unmul.ac.id

Novianti Puspitasari
Department of Informatics
Mulawarman University
Samarinda, Indonesia
novia.ftik.unmul@gmail.com

Masna Wati
Department of Informatics
Mulawarman University
Samarinda, Indonesia
masnawati@fkti.unmul.ac.id

***Abstract*—The aglaonema plant is quite in demand by the public because it is a current trend and can be a business field for ornamental plant lovers. However, the growth of aglaonema plants can be hampered by disease, so a system is needed. This Aglaonema plant disease diagnosis expert system was created to determine the type of disease so that it can be overcome as early as possible. The expert system method used is the Forward Chaining method and the Naive Bayes method. The Forward Chaining method is used to make rules based on symptoms and the Naïve Bayes method to determine the results of the diagnosis. This study used 25 symptoms for 10 diseases and obtained an accuracy value of 76% using 50 data.**

Keywords—aglaonema, expert system, forward chaining, naive bayes

I. INTRODUCTION

Indonesia's climate as a tropical country provides convenience for ornamental plant entrepreneurs whose variety is also found in Indonesia. The ornamental plant business is not impossible to match the vegetable business and the fruit business which is currently still at the top, if it can take advantage of the wide variety of plants and combine them with the right techniques [1].

Currently, ornamental plants are in great demand by the public, because they have become a current trend and ornamental plants have become a business field for ornamental plant lovers, one of the most popular ornamental plants is Aglaonema [2]. The Aglaonema plant or in Indonesia by another name "sri rejeki" is an ornamental plant without flowers but has a variety of leaves that include motifs, shapes, colors, and sizes, this is what makes this plant traded by counting the number of leaves, leaf color and motifs [3].

The appeal of Aglaonema lies in the beauty of the leaf patterns that collaborate with the sparkle of the color. Even though it seems tempting, it doesn't mean Aglaonema's business is without stumbling blocks. One of the problems that often occurs is disease attacks. Diseases attack indiscriminately, both standard and exclusive Aglaonema for tens of millions of rupiah were attacked. Many phenomena occur in Aglaonema farmers and businessmen who do not understand this plant disease. Especially for the "new players" who tend to be "talkative" about this plant business. The problem that then arises is the low quality of the plants, even death. This problem will cause losses, therefore we need a technology that can find out the type of disease against the symptoms experienced [4].

The rapid development of technology is directly proportional to the development in the field of computers. Various innovations continue to be developed to improve the role and performance of computers. One of the systems adopted by computer technology to be used as a disease diagnosis is an expert system [5]. Expert systems are part of artificial intelligence which is a computer program that can use knowledge from experts. In addition, the expert system is able to make inference procedures in order to solve a complex problem. The ability of this system requires an expert or expert to be able to solve the problem. It can be concluded that an expert system is a system that can absorb the knowledge of an expert into a computer program so that the system can solve a problem like an expert [6][7].

The expert system method used is forward chaining and naive bayes. The Forward Chaining method is a search method or forward tracking technique based on existing information and then combines it into a rule to get a conclusion or result [8][9]. Making this expert system begins by first declaring questions from user problems to then the question underlies the conclusion of the problem analysis after calculating the results of the diagnosis using Naive Bayes [9][10].

Several studies have been conducted using the forward chaining method and the naive Bayes method to diagnose diseases in plants. The forward chaining method was used to diagnose diseases in rice [5][11], cocoa [12][13], and corn [14][15] plants. Naive Bayes method is used to diagnose diseases in corn [16][17], palm oil [10], and coffee [18] plants. And the research uses forward chaining and naive bayes methods to diagnose diseases in papaya plants [19]. The diagnosis of disease in aglaonema plants using the forward chaining method was carried out by Ratih [20] but there is no information on how many symptoms were used in her research. Therefore, this study will use forward chaining and naive bayes methods to detect diseases in aglaonema plants, where forward chaining methods are used to create rules and naive bayes methods to diagnose diseases.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

II. METHODS

A. Expert System

An expert system is a computer system or program designed in such a way that it can make decisions like an expert [6][9][21]. An expert system is a computer system or program that can use science, facts, and thinking techniques in making decisions. Based on the information possessed, the expert system can solve problems like an expert or expert in his field [22]. In making the system, the expert system will combine inference rules based on a certain knowledge base provided by one or more experts in a particular field [23]. The combination of these two things is stored in the computer, which is then used in the decision-making process to solve a particular problem [24][25].

B. Forward Chaining

Forward chaining is a search method that starts the search process from a set of data or facts to get a conclusion which is the solution or result of the problem at hand. The way forward chaining works is to choose a rule where the premise part matches the information in the working memory section [9][21]. The mechanism of the forward trace inference method begins by entering a set of known facts, then these facts will be matched with the rules that form the knowledge base of the system. These rules work by using if-then rules that exist in the system. If there are facts that match the if part, then the rule will be executed [26].

C. Naive Bayes

Naïve Bayes is a simple probabilistic classification method that calculates a given set of probabilities by adding up the frequencies and combining values from a given dataset [9][21]. This algorithm uses Bayes' Theorem and that can assume all independent or independent attributes given by the class variable value. Another definition says Naïve Bayes is a classification method that uses probability and statistical methods proposed by the British scientist Thomas Bayes, which can predict future opportunities based on previous experience. The general formula used in Bayes' Theorem with Equation 1.

$$P(H \mid E) = \frac{P(E \mid H).P(H)}{P(E)} \quad (1)$$

description:
p(H|E): Probability of hypothesis H occurs if evidence E occurs
p(E|H): Probability of emergence of evidence E, if hypothesis H occurs
p(H): Probability of hypothesis H regardless of the evidence
p(E): the probability of evidence E regardless/no matter what

To explain how the Naïve Bayes theorem works, it should be noted that the classification process requires some clues to determine what class is suitable for the sample being analyzed. Therefore, Bayes' theorem in the classification is stated in Equation 2.

$$P(C \mid F1...Fn) = \frac{P(C).P(F1...Fn \mid C)}{P(F1...Fn)} \quad (2)$$

Where Variable C represents the class, while the variable F1 ... Fn represents the characteristics of the instructions needed to perform the classification. Then the formula explains that the probability of entering a sample of certain characteristics in class C (Posterior) is the probability of the appearance of class C (before the inclusion of the sample, often called a prior), multiplied by the probability of the occurrence of sample characteristics in class C (also called likelihood), divided by the probability of the appearance of the characteristics of the sample globally (also called evidence). Therefore, the formula can also be written simply as Equation 3.

$$Posterior = \frac{prior * likelihood}{evidence} \quad (3)$$

III. RESULTS

Making rules in this study using forward chaining where all information is collected such as symptoms and diseases. Rules will be formed based on information provided by experts regarding diseases in Aglaonema plants, there are 25 symptoms that can be seen in Table I, for 10 diseases can be seen in Table II.

TABLE I. SYMPTOM NAME

No.	Code	Symptom Name
1.	G01	Soft stems like porridge
2.	G02	Soft leaves like mush
3.	G03	Smells bad
4.	G04	Looks like it was hit by hot water
5.	G05	The edges of the stems and leaves are purplish red which eventually rots and spread throughout the leaves
6.	G06	Brown leaf spot
7.	G07	Pale leaf bones
8.	G08	Rooting rot
9.	G09	Stems and leaves are grayish brown and eventually rot
10.	G10	Brown root
11.	G11	Growing dwarf
12.	G12	The leaves are curling
13.	G13	The leaves turn yellow
14.	G14	Pale leaf color
15.	G15	The leaves are shriveled
16.	G16	There are white spots on the leaves
17.	G17	The leaves are wilting
18.	G18	Colonies by forming rows at the bottom of the leaf bone
19.	G19	The leaves and stems are white
20.	G20	The leaves become hollow
21.	G21	The leaves are ragged
22.	G22	The edges of the leaves are slightly jagged
23.	G23	The edges of the leaves are starting to break
24.	G24	There are blackish white spots on the leaves
25.	G25	White spots spread on the stem

TABLE II. DISEASE NAME

No.	Code	Disease Name
1.	P1	Soft Rot
2.	P2	Fusarium wilt
3.	P3	Botrytis Mushroom
4.	P4	Virus
5.	P5	scale lice
6.	P6	White Flea
7.	P7	Shield Flea
8.	P8	Caterpillar
9.	P9	grasshopper
10.	P10	Black ant

Experts will provide information about the symptoms experienced by each disease in Agloenema plants. Based on the information obtained from the experts, rules will be made using the forward chaining method. Making rules based on these symptoms can be seen in Table III and a picture of data tracking with the forward chaining method can be seen in Figure 1.

TABLE III. RULES USING THE FORWARD CHAINING METHOD

No	Sym ptom	Disease									
		P1	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*
1	G01	√									
2	G02	√									
3	G03	√									
4	G04	√									
5	G05		√								
6	G06		√	√							
7	G07		√								
8	G08		√	√							
9	G09			√							
10	G10			√							
11	G11				√		√				
12	G12				√						
13	G13				√	√	√				
14	G14				√						
15	G15					√	√				
16	G16					√					
17	G17					√	√				
18	G18							√			
19	G19							√			√
20	G20								√	√	
21	G21								√	√	
22	G22								√		
23	G23									√	
24	G24										√
25	G25										√

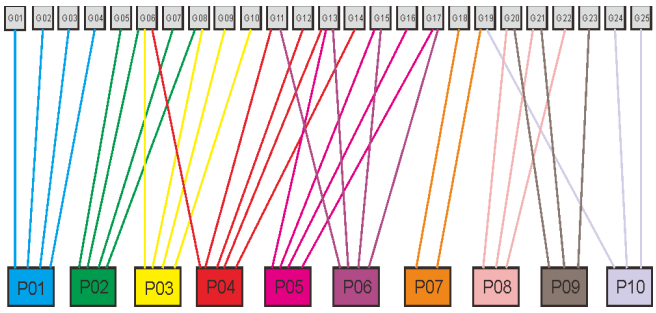


Fig. 1. Data tracking with the forward chaining method

IV. DISCUSSIONS

This expert system for diagnosing diseases of the aglaonema plant uses the forward chaining method as a determination rule and then it will be analyzed using the Naïve Bayes method to obtain diagnostic results. The research to be carried out has stages of an activity to develop an expert system to be built.

Tests are carried out to determine the accuracy of the diagnosis generated by the system using the Naïve Bayes method. For example, the symptoms selected by the user are Purplish-red stems and leaves that eventually rot (G05), brown leaf spots (G06), pale leaf bones (G07), and rotting roots (G08).

A. Calculating the probability of each disease

All possible diseases that are selected can occur, then all diseases will be calculated probabilities.

$$P1(\text{Disease 1}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P2(\text{Disease 2}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P3(\text{Disease 3}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P4(\text{Disease 4}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P5(\text{Disease 5}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P6(\text{Disease 6}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P7(\text{Disease 7}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P8(\text{Disease 8}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P9(\text{Disease 9}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

$$P10(\text{Disease 10}) = \frac{\text{Number of Possible Disease that Apper}}{\text{Total of All Diseases}} = 0.1$$

B. Calculate the probability of the selected symptom in each disease

Select the selected symptom and compare it with the number of symptoms in each disease. The more symptoms selected in one disease, the greater the probability value.

$$P1(\text{Yes} \mid \text{Disease 1}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{4} = 0$$

$$P2(\text{Yes} \mid \text{Disease 2}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{4}{4} = 1$$

$$P3(\text{Yes} \mid \text{Disease 3}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{2}{4} = 0.5$$

$$P4(\text{Yes} \mid \text{Disease 4}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{1}{5} = 0.2$$

$$P5(\text{Yes} \mid \text{Disease 5}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{4} = 0$$

$$P6(\text{Yes} \mid \text{Disease 6}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{4} = 0$$

$$P7(\text{Yes} \mid \text{Disease 7}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{2} = 0$$

$$P8(\text{Yes} \mid \text{Disease 8}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{3} = 0$$

$$P9(\text{Yes} \mid \text{Disease 9}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{3} = 0$$

$$P10(\text{Yes} \mid \text{Disease 10}) = \frac{\text{Number of Symptoms Selected}}{\text{Number of Symptoms in the Disease}} = \frac{0}{3} = 0$$

C. Calculates the probability of the selected symptom in all diseases

Adding up all the scores obtained at point B.

$$\begin{aligned} P(\text{Yes}) = {} & (P1(\text{Disease 1}) * P1(\text{Yes} \mid \text{Disease 1})) + \\ & (P2(\text{Disease 2}) * P2(\text{Yes} \mid \text{Disease 2})) + \\ & (P3(\text{Disease 3}) * P3(\text{Yes} \mid \text{Disease 3})) + \\ & (P4(\text{Disease 4}) * P4(\text{Yes} \mid \text{Disease 4})) + \\ & (P5(\text{Disease 5}) * P5(\text{Yes} \mid \text{Disease 5})) + \\ & (P6(\text{Disease 6}) * P6(\text{Yes} \mid \text{Disease 6})) + \\ & (P7(\text{Disease 7}) * P7(\text{Yes} \mid \text{Disease 7})) + \\ & (P8(\text{Disease 8}) * P8(\text{Yes} \mid \text{Disease 8})) + \\ & (P9(\text{Disease 9}) * P9(\text{Yes} \mid \text{Disease 9})) + \\ & (P10(\text{Disease 10}) * P10(\text{Yes} \mid \text{Disease 10})) \end{aligned}$$

$$\begin{aligned} P(\text{Yes}) = {} & 0.1 * 0 + 0.1 * 1 + 0.1 * 0.5 + \\ & 0.1 * 0.2 + 0.1 * 0 + 0.1 * 0 + \\ & 0.1 * 0 + 0.1 * 0 + 0.1 * 0 + \\ & 0.1 * 0 = 0.17 \end{aligned}$$

D. Probability of YES in Disease to all Diseases

The results of the following calculations are to see how likely the disease is based on the selected symptoms.

$$P1(\text{Disease 1} \mid \text{Yes}) = \frac{P1(\text{Yes} \mid \text{Disease 1}) * P1(\text{Disease 1})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P2(\text{Disease 2} \mid \text{Yes}) = \frac{P2(\text{Yes} \mid \text{Disease 2}) * P2(\text{Disease 2})}{P(\text{Yes})} = \frac{1 * 0.1}{0.17} = 0.588$$

$$P3(\text{Disease 3} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 3}) * P3(\text{Disease 3})}{P(\text{Yes})} = \frac{0.5 * 0.1}{0.17} = 0.294$$

$$P4(\text{Disease 4} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 4}) * P4(\text{Disease 4})}{P(\text{Yes})} = \frac{0.2 * 0.1}{0.17} = 0.117$$

$$P5(\text{Disease 5} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 5}) * P5(\text{Disease 5})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P6(\text{Disease 6} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 6}) * P6(\text{Disease 6})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P7(\text{Disease 7} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 7}) * P7(\text{Disease 7})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P8(\text{Disease 8} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 8}) * P8(\text{Disease 8})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P9(\text{Disease 9} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 9}) * P9(\text{Disease 9})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

$$P10(\text{Disease 10} \mid \text{Yes}) = \frac{P(\text{Yes} \mid \text{Disease 10}) * P10(\text{Disease 10})}{P(\text{Yes})} = \frac{0 * 0.1}{0.17} = 0$$

So based on calculations based on the symptoms of purplish red stem and leaf edges that eventually rot (G05), brown leaf spots (G06), pale leaf bones (G07), and rotting roots (G08) the diagnosis results in Fusarium wilt disease (P2) with the highest bayes value is 0.588.

TABLE IV. EXPERT RESULTS COMPARED TO SYSTEM RESULTS

No	Test Case	Expert Diagnosis Results	System Diagnostic Results	Information
1	G05, G06, G21, G23	P09	P09	Same
2	G05, G06, G19, G24	P10	P10	Same
3	G06, G08, G13, G15	P03	P02	Different
4	G06, G08, G18, G19	P07	P07	Same
..	.....	.....	.....	.....
47	G06, G08, G20, G21	P08	P08	Same
48	G06, G08, G21, G23	P09	P09	Same
49	G11, G12, G18, G19	P07	P04	Different
50	G11, G12, G20, G21	P08	P08	Same

Testing the accuracy of the diagnosis results is done by comparing the results of the diagnosis on the system with the results of the diagnosis from the expert. After getting the test data, then the level of accuracy is measured in percent. This stage of testing the accuracy of the diagnosis results using symptom and disease data obtained from experts. The results of the diagnosis of the system will be compared with the results of the experts. If the results of the diagnosis of the system are the same as the results of the experts, then the data is declared correct or appropriate, data can be seen in Table IV.

Number 3 and number 49 in Table IV are results obtaines between the expert and the system are different. Based on Table III, the symptom in number 3 is most likely disease 2 because G06 and G08 are symptoms of disease 2, while G13 and G15 are not symptoms in disease 2 or disease 3. The difference in results in number 3 can be stated as a human error due to a wrong diagnosis. However, it is different with result number 49, where based on Table III, G11 and G12 are symptoms for disease 4, while G17 and G18 are symptoms for disease 7, so it can be seen at a glance that the results from the expert and the results from the system are both correct. The reason for the difference in results in number 49 is not yet known, so further research is needed to be able to analyze the causes of difference in results. In addition, due to differences in results, the accuracy of the system is not too high. Testing the results of the diagnosis based on Table 4 gets an accuracy value of 76%.

$$\text{Accuracy} = \frac{\text{The Same Amount of Data}}{\text{The Total Amount of Data}} = \frac{38}{50} * 100\% = 76\%$$

V. CONCLUSION

Forward chaining and naive bayes methods can be applied to expert systems to diagnose diseases in Agloenema plants. The symptoms used in the study were 25 symptoms for 10 diseases. The study used 50 data and resulted in an accuracy value of 76%. Future research is expected to use other methods so that the accuracy values can be compared to find out which method is more suitable for use in the expert system for diagnosing aglaonema plant diseases. In addition, it is necessary to further analyze the causes of differences in results from experts and results from the system.

ACKNOWLEDGMENT

We would like to thank the experts in this research, namely Dr. Yesnti Puspita Sari, M.Si and Dr. Linda Oktavianingsih, M. Si, Lecturer of the Department of Biology, Faculty of Mathematics and Natural Sciences, Mulawarman University, Indonesia, who provided information on symptom data from the emergence of diseases in Aglaonema plants.

REFERENCES

[1] C. Byess, "Plant Profits: An Instructional Manual to Hosting a School Plant Sale and Funding Agricultural Education Programs from the Profits," 2020, Accessed: Sep. 02, 2021. [Online]. Available: https://lib.dr.iastate.edu/creativecomponents/577/.

[2] M. Zahara and C. C. Win, "A Review: The Effect of Plant Growth Regulators on Micropropagation of Aglaonema sp.," *J. Trop. Hortic.*, vol. 3, no. 2, p. 96, 2020, doi: 10.33089/jthort.v3i2.58.

[3] O. Goni, M. F. Khan, M. M. Rahman, M. Z. Hasan, F. B. Kader, N. Sazzad, M. A. Sakib, B. Romano, M. A. Haque, and R. Capasso, "Pharmacological insights on the antidepressant, anxiolytic and aphrodisiac potentials of Aglaonema hookerianum Schott," J. Ethnopharmacol., vol. 268, p. 113664, 2021, doi: 10.1016/j.jep.2020.113664.

[4] L. Banos, "Effect of potting media on growth of Aglaonema , fishtail palm and Freycinetia Source :," *jhs.iihr.res.in*, pp. 1–2, 2002, Accessed: Sep. 02, 2021. [Online]. Available: https://jhs.iihr.res.in/index.php/jhs/article/view/231.

[5] A. M. Tambunan, S. R. Siringoringo, R. Aruan, P. I. Aisyah, and D. Sitanggang, "An expert system for diagnosing plant diseases using certainty factor and backward chaining based on android," in *Journal of Physics: Conference Series*, 2019, vol. 1230, no. 1, p. 12075, doi: 10.1088/1742-6596/1230/1/012075.

[6] W. P. Wagner, "Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies," *Expert Syst. Appl.*, vol. 76, pp. 85–96, 2017, doi: 10.1016/j.eswa.2017.01.028.

[7] M. Bohlouli, N. Mittas, G. Kakarontzas, T. Theodosiou, L. Angelis, and M. Fathi, "Competence assessment as an expert system for human resource management: A mathematical approach," *Expert Syst. Appl.*, vol. 70, pp. 83–102, 2017, doi: 10.1016/j.eswa.2016.10.046.

[8] P. Windriyani, S. Kom, S. T. Wiharto, M. Kom, S. W. Sihwi, and S. Kom, "Expert system for detecting mental disorder with forward chaining method," in *Proceedings - International Conference on ICT for Smart Society 2013: "Think Ecosystem Act Convergence", ICISS 2013*, 2013, pp. 79–85, doi: 10.1109/ICTSS.2013.6588068.

[9] H. R. Hatta, F. Ulfah, D. M. Khairina, H. Hamdani, and S. Maharani, "Web-expert system for the detection of early symptoms of the disorder of pregnancy using a forward chaining and Bayesian method," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 11, pp. 2589–2599, 2017, Accessed: Sep. 02, 2021. [Online]. Available: https://repository.unmul.ac.id/handle/123456789/4666.

[10] M. Nababan, Y. Laia, D. Sitanggang, O. Sihombing, E. Indra, S. Siregar, W. Purba, and R. Mancur, "The diagnose of oil palm disease using Naive Bayes Method based on Expert System Technology," in Journal of Physics: Conference Series, 2018, vol. 1007, no. 1, p. 12015, doi: 10.1088/1742-6596/1007/1/012015.

[11] E. Agustina, I. Pratomo, A. D. Wibawa, and S. Rahayu, "Expert system for diagnosis pests and diseases of the rice plant using forward chaining and certainty factor method," in *2017 International Seminar on Intelligent Technology and Its Application: Strengthening the Link Between University Research and Industry to Support ASEAN Energy Sector, ISITIA 2017 - Proceeding*, Nov. 2017, vol. 2017-Janua, pp. 266–270, doi: 10.1109/ISITIA.2017.8124092.

[12] M. M. Yusof, N. F. Rosli, M. Othman, R. Mohamed, and M. H. A. Abdullah, "M-DCocoa: M-agriculture expert system for diagnosing cocoa plant diseases," in *Advances in Intelligent Systems and Computing*, 2018, vol. 700, pp. 363–371, doi: 10.1007/978-3-319-72550-5_35.

[13] V. Ariandi, H. Kurnia, Heriyanto, and H. Marry, "Expert system for disease diagnosis in cocoa plant using android-based forward chaining method," in *Journal of Physics: Conference Series*, Dec. 2019, vol. 1339, no. 1, p. 012009, doi: 10.1088/1742-6596/1339/1/012009.

[14] M. Saiful and A. Muliawan Nur, "Application of Expert System with Web-Based Forward Chaining Method in Diagnosing Corn Plant Disease," in *Journal of Physics: Conference Series*, 2020, vol. 1539, no. 1, doi: 10.1088/1742-6596/1539/1/012019.

[15] M. Saiful and A. Muliawan Nur, "Application of Expert System with Web-Based Forward Chaining Method in Diagnosing Corn Plant Disease," in *Journal of Physics: Conference Series*, May 2020, vol. 1539, no. 1, p. 012019, doi: 10.1088/1742-6596/1539/1/012019.

[16] B. Budianto, I. Fitri, and W. Winarsih, "Expert System for Early Detection of Disease in Corn Plant Using Naive Bayes Method," *J. Mantik*, vol. 3, no. 36, pp. 308–317, 2020, Accessed: Sep. 02, 2021. [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/568.

[17] I. R. Munthe, B. H. Rambe, R. Pane, D. Irmayani, and M. Nasution, "Expert System for Early Detection of Disease in Corn Plant Using Naive Bayes Method," *J. Mantik*, vol. 3, no. January, pp. 31–38, Feb. 2019, Accessed: Oct. 20, 2021. [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/568.

[18] I. D. Rafi Syahputra, Agung Triayudi, "Application Of Expert System To Diagnose Pests And Diseases In Coffee Plant Using Web-Based Naïve Bayes," *J. Mantik Vol. 3 Number 4, Febr. 2020, pp. 383-392 E-ISSN 2685-4236*, vol. 3, no. 4, pp. 444–450, 2020, Accessed: Oct. 20, 2021. [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/576/372.

[19] W. E. Sari, Y. E. Kurniawati, and P. I. Santosa, "Papaya Disease Detection Using Fuzzy Naïve Bayes Classifier," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, Dec. 2020, pp. 42–47, doi: 10.1109/ISRITI51436.2020.9315497.

[20] R. A. Imaniar, "Sistem Pakar Untuk Mendiagnosa Hama Penyakit Tanaman Aglaonema," *J. Ilmu Komput.*, vol. 1, no. 1, pp. 70–79, Mar. 2017, doi: 10.33060/jik/2012/vol1.iss1.8.

[21] S. Maharani, N. Dengen, G. Y. Saputra, D. M. Khairina, and H. R. Hatta, "Expert system applications for early diagnosis teeth and oral disease in children," in *ICITACEE 2015 - 2nd International Conference on Information Technology, Computer, and Electrical Engineering: Green Technology Strengthening in Information Technology, Electrical and Computer Engineering Implementation, Proceedings*, 2016, pp. 87–91, doi: 10.1109/ICITACEE.2015.7437776.

[22] D. D. Woods, "Decomposing automation: Apparent simplicity, real complexity," in *Automation and Human Performance: Theory and Applications*, 2018, pp. 3–17.

[23] J. Martinez-Gil, "Automated knowledge base management: A survey," *Computer Science Review*, vol. 18. Elsevier, pp. 1–9, 2015, doi: 10.1016/j.cosrev.2015.09.001.

[24] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Model. User-adapt. Interact.*, vol. 27, no. 3–5, pp. 393–444, Dec. 2017, doi: 10.1007/s11257-017-9195-0.

[25] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *Int. J. Inf. Manage.*, vol. 48, pp. 63–71, 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[26] E. Davis and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Communications of the ACM*, vol. 58, no. 9. Association for Computing Machinery, pp. 92–103, Sep. 01, 2015, doi: 10.1145/2701413.

Indonesian Digital Wallet Sentiment Analysis Using CNN And LSTM Method

Ananda Affan Fattahila
School of Computing
Telkom University
Bandung, Indonesia
affanfattahila@student.tekomuniversity.ac.id

Fendi Irfan Amorokhman
School of Computing
Telkom University
Bandung, Indonesia
fendiirfan@student.tekomuniversity.ac.id

Kaenova Mahendra Auditama
School of Computing
Telkom University
Bandung, Indonesia
kaenova@student.tekomuniversity.ac.id

Kurniadi Ahmad Wijaya
School of Computing
Telkom University
Bandung, Indonesia
kurniadiwijaya@student.tekomuniversity.ac.id

Ade Romadhony
School of Computing
Telkom University
Bandung. Indonesia
aderomadhony@telkomuniversity.ac.id

***Abstract*— In this paper, we present a deep learning based approach to sentiment analysis on Indonesian digital wallet review. We collected the reviews of the following digital wallets: DANA, OVO, Link Aja, and Sakuku by obtaining the data from Google Play Store. Sentiment analysis from comments on Google Play Store is used to extract user opinions on various aspects: accessibility, transaction, account, service, and performance. We applied the CNN LSTM architecture on performing the sentiment analysis. The experimental results show that the training accuracy is 87% and the validation accuracy is 83%. Based on the data and analysis results of the four digital wallet applications, namely DANA, OVO, LinkAja, and Sakuku, the sentiments tend to be negative towards the categories we have created, including transactions and access, services, accounts, and performance.**

Keywords—sentiment analysis, digital wallet, CNN, LSTM

I. INTRODUCTION

The Digital Era is a time when information is quick to scale, easy to obtain, and disseminate using digital technology or a technology that uses a computerized system connected to the internet [1]. The COVID-19 pandemic has impacted many sectors such as industry, economy, government, education, etc. This also makes system changes that cause some activities to use quota restrictions until they are done online. The impact of the COVID-19 pandemic also demands that many things can be done digitally or electronically. As a result, technology has become very popular because it can help human needs and facilitate human work [2].

The increasingly developing technology provides many conveniences in the economic sector, transportation, information, education, to the comfort of digital transactions, as we can see in the current development of digital wallets that can be done quickly with smartphones. The increasing use of e-commerce today affects consumer habits in making payments [3].

The development of digital wallets as digital wallets used by people to facilitate online and offline transactions are increasingly popular. In 2018, Bank of Indonesia recorded that the number of transactions made using digital wallets in Indonesia reached Rp 23.3 trillion. This value is expected to grow with the popularity of cashless payment methods to Rp 355.7 trillion in 2023 [4]. In February 2019, Go-Pay still ranked first, followed by OVO, DANA, LinkAja, and Sakuku [4]. Digital wallet technology is indeed easy to use because. Thanks to this technology, humans do not need to be in direct contact to make payments. However, behind the ease of using digital wallet technology, it can be seen from several cases, such as in the Go-Pay application, sometimes there are still problems such as the balance not increasing after transfer from a bank [5]. In fact, there are still several other cases that occur, of course can be harmful to the company if not resolved.

In today's highly competitive market, the step towards success for any company, including digital wallet-based companies is normal. Several digital wallet companies have widespread reviews, starting from Twitter, Facebook, google play store, and many others. One way to do sentiment analysis is to collect existing reviews, where the data can be a form of feedback from real users. By analyzing the sentiments in consumer reviews, it can be classified whether the consumer's perception of the product is positive or negative.

Thus, the sentiment analysis approach is considered one of the best ways to track or know consumers' feelings. This research is expected to give an advantage to a company over its competitors in the market. The information obtained from the analysis can assist companies in answering their questions, such as how someone reacts or feels after using an existing feature.

Digital wallets still have a lot of criticism from the public because this technology is still new, so improvements need to be made. This is also supported by ratings on several digital wallet applications such as Go-Pay, OVO, LinkAja, Sakuku, which still vary, and not all of them receive a 5-star rating. To gather the improvements suggestion from the public, we conducted a sentiment analysis experiment on several Indonesian digital wallets review using the CNN-LSTM method.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

We chose to use CNN-LSTM, started with CNN in order to reduce the length of the incoming input data and continue with LSTM in order to learn from the previous data and LSTM has a memory block that will determine which value will be selected as the relevant output for the given input.

II. RELATED WORK

The emergence of new technology that everyone can use makes consumers communicate more often with service providers or products to make high use of services that can be used online in daily activities [9]. Purchasing online by using a service, the quality of the service itself is something that consumers will consider. Consumers expect that the service of an application product can provide good quality and satisfaction. This is also related to digital wallet products between consumers and digital wallet companies to communicate transactions. When a transaction occurs at a digital wallet company, evaluating the consumer experience in using the service will feel satisfied and build a sense of trust, commitment, and loyalty of consumers using a digital wallet application [10].

Sentiment Classification is a public work with the rapid development of information technology to deal with public sentiment. Public sentiment is essential data in the research domain to analyze the emotional point of view and the author's opinion on various issues [6]. From this, it can be seen where public sentiment is precious for correcting an error or deficiency contained in digital wallet products for the convenience of consumers.

Machine learning and deep learning approaches are used to classify public emotions. In a deep learning study, it is much better than machine learning algorithms to conduct public sentiment. Indicated by the accuracy of machine learning testing and training 81% to 90%, while deep learning got 85% to 97% [7].

Sentiment analysis is a fundamental technique for exploring public emotions and is a multidisciplinary research area in developing social media interactions and dealing with big data [8]. Sentiment analysis from comments on google play store is used to extract user opinions on various views such as application system, design, performance, and other issues. This paper uses the CNN-LSTM approach. Convolutional Neural Network (CNN) is one type of neural network that is very good when faced with spatial data.

The Convolutional layer is only used for current connections from the previous layer, local neurons are connected to neurons in the next layer. This method helps increase knowledge at the layer to understand more generally the incoming input. In LSTM, there are several steps: define network, compile network, fit network, evaluate network, and make a prediction. Using LSTM can determine a value that will be used as output relevant to the given input.

Long Short Term Memory (LSTM) is a neural network development that can be used for time-series data modeling [11]. LSTM can also be interpreted as a neural network with an adaptable architecture, so that its shape can be adjusted, depending on the application that is made. In LSTM, there are several steps: define network, compile network, fit network, evaluate network, and make a prediction. Using LSTM can determine a value that will be used as output relevant to the given input.

III. DATA COLLECTION AND SENTIMENT ANALYSIS

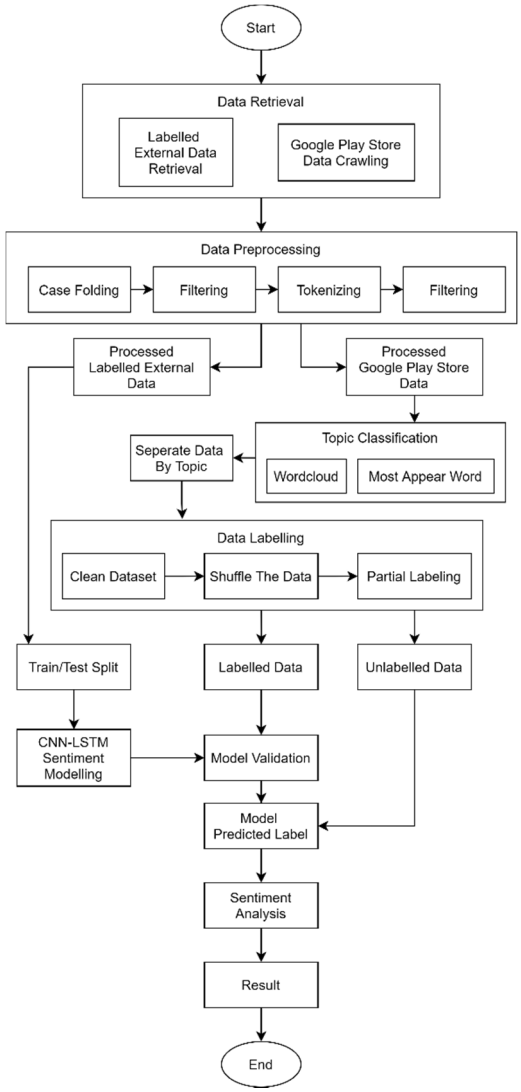


Fig. 1. Workflow of Indonesian Digital Wallet Sentiment Analysis using CNN-LSTM Method.

A. Dataset

The review data collection is taken from the Google Play Store at the data crawling stage (Fig. 1). The reviews of digital wallet sentiment data that will be taken are reviews of digital wallet applications from large companies such as OVO, DANA, LinkAja, and Sakuku digital wallet applications. We used the google-play-scraper library to retrieve reviews on the top digital wallet applications on the play store, with the crawling requirement is only Indonesian reviews that will be gathered. In data retrieval, we tried to collect as much data as possible until the scraper cannot find data according to the specified requirements. The requirements that we determine in doing scraping are the review language used according to the Google Play Store and the relevance of the review. Table I shows review examples. We retrieved a data with a total of 2102 reviews on OVO, 3711 reviews on DANA, 1524 reviews on LinkAja, and 678 reviews on Sakuku. We do manual sentiment labeling on these data for making training and testing data. In total we successfully labeled 200 reviews data. Using Kappa Level of

Agreement, we can justify the reliability of the manual annotate data. The value of kappa we got for the data is 0.94 .

TABLE I. REVIEW EXAMPLES

Digital Wallet	Crawling Review Sample
Link Aja	*Aplikasi yang sangat simple, memudahkan kita untuk melakukan transaksi. Aplikasi ini sangat memudahkan kita untuk pembayaran pinjaman online, token listrik, bayar gojek, bayar grab food & bayar di Bukalapak. Benar-benar aplikasi yg sangat lengkap & ga bikin ribet buat bayar ini itu.* / A very simple application, make us easy to do the transaction. This application is very helpful to do the online lending payment, electricity token, online transportation payment (gojek, grab food), e-commerce (Bukalapak) payment. It's really a complete application, easy to pay anything.
DANA	*Kenapa si dana skrng menurun amat system nya, saya biasa bayar tagihan wifi, pulsa listrik, beli paket data pakai dana suka ilang fitur² pembayaran nya aneh skrng* / Why dana system performance is now decreasing. I usually pay the wifi bills, electricity pulse, data packet, now I could not find those features, strange.
OVO	*Semenjak di update salah kode security terus, padahal uda benar kode security-nya, Error ini Aplikasi OVO nya min. Jdi ngk nyaman dalam bertransaksi lewat OVO ini.* / After updated, I got security code error message forever, while I believe the code is correct. It must be an error on OVO application. It's really an inconvenience experiment having a transaction via OVO.
Sakuku	*sakuku sangat membantu, hanya saja limit lnya terlalu kecil. harapan saya ada tambahan lagi fitur fitur menarik untuk pengalaman yang lebih maksimal* / sakuku is very helpful, the only downside is the limit is very low. I hope there will be additional interesting features for better user experience.

Other than the review data retrieved from Google Play Store, we also retrieved another data from IndoBenchmark[1] and IndoLEM [2] that has a similar data for sentiment prediction. We employed data augmentation because the datasets from IndoBenchmark and IndoNLU have been used to train and benchmark more advance models like IndoBERT [13] which now is one of state-of-the-art models in Bahasa Indonesia Language Model.

B. *Preprocessing*

In our research, after collecting the dataset, a data pre-processing process is carried out which has a flow as shown in Fig. 2.

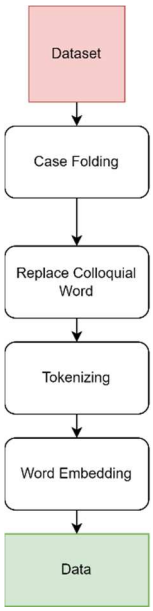


Fig. 2. Workflow of Data Preprocessing.

C. *Aspect Classification*

The Digital Wallet application is not only related to people's satisfaction when transacting. There are several other metrics that can affect people's satisfaction in using the application. At this stage we try to find out what things can affect a user to have negative sentiments and positive sentiments. We use the review data that has been collected to see what words can be classified into certain aspects.

In determining these aspects, we use the word cloud method for each Digital Wallet. We will then analyze the results of the word cloud and agree on what aspects and keywords that allow for a review to be classified into these aspects. We decided 5 aspects: Accessibility, Transaction, Account, Service, and Performance, which in each aspect has its own keywords for a review. The keywords of each aspect can be seen in Table 2.

[1]The IndoBenchmark data we retrieved can be accessed on https://github.com/indobenchmark/indonlu/tree/master/dataset/smsa_doc-sentiment-prosa

[2]The data from IndoLEM we retrieved can be accessed from https://github.com/indolem/indolem/tree/main/sentiment/data

TABLE II. ASPECT AND KEYWORDS

Aspect	Keywords
Accessibility	*Login, Register, OTP, Kode, Pin*
Transaction	*Transfer, Pembelian, Pembayaran, Saldo, Topup, Top, Up, Beli, Listrik, Transaksi, Biaya, Potongan, Uang, Beli*
Account	*Upgrade, Verifikasi, KTP, Premium, Akun*
Service	*Email, WA, Whatsapp, Telpon, CS, Telfon, Respon, pelayanan*
Performance	*Lambat, Lag, Cepat, Performa*

After determining the aspects and keywords, we categorized the reviews. The categorization was performed by keyword matching, so that if a similar word is found in the keywords in the review, it will be classified in the aspect that has those keywords. Based on the approach, a review can be classified into many aspects. An example can be seen in Table III.

TABLE III. EXAMPLE OF RIVIEWS ASPECT CLASSIFICATION

Reviews	Aspect Classification
[Original] aduh jadi min tiba login sandi salah benar loh coba bekukan jam maksudnya ya perbaiki akun mengirim uang ribu [Translation] Argh so *min* at login password wrong right though try freeze hour means yes fix account send money thousand	Accessibility: login Transaction: uang Account: akun

D. Sentiment Analysis

Before we built the model, we performed several pre-processing: replace colloquial word and typo word using kamus_alay [3], case folding, tokenization, embedding. stopword removal. To perform the classification, we used CNN-LSTM approach. The model parameters were shown on Table IV and the architecture is shown on Fig. 3.

TABLE IV. THE CNN-LSTM PARAMETERS

Parameter	Value
Batch Size	60
Epoch	20
Conv Filter 1D	filters=32, kernel_size=2, padding='same', activation='relu'
Max Pooling 1D	pool_size=2
LSTM Size	128, dropout=0.2, recurrent_dropout=0.2
Layer Dense (1)	64
Layer Dense (2)	16
Dropout	0.5

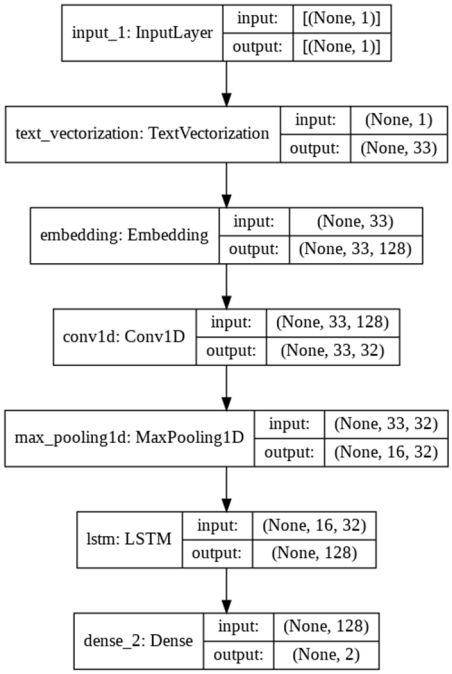


Fig. 3. The CNN-LSTM architecture.

E. Model Evaluation

We evaluated the CNN-LSTM model using the following metrics: accuracy, loss, precision, recall, and F-1 score. We divide the dataset by 20% for the test dataset and 80% for the train dataset where we have chosen a random_state of 10. Meanwhile, the testing dataset comes from the results of scraping reviews on the google play store. The labeling for each review was performed by two of the authors, with any disagreement was solved by discussion. Table V shows the model performance on training, validation, and testing dataset of the negative sentiment label or positive sentiment.

[3] Kamus_Alay can be accessed on https://bit.ly/3G8Mcfb and https://bit.ly/2ZGNazx

TABLE V. MODEL PERFORMANCE

Data	Accuration	Loss	Precision	Recall	F1
Traning	0.87	0.4	0.89	0.89	0.89
Validation	0.83	0.27	0.82	0.82	0.83
Testing	**0.81**	0.44	0.81	0.81	**0.77**

The training evaluation on each epoch: accuracy, loss, and F-1 score were shown on Table V. We can conclude that the model is quite good based on the validation evaluation. After that, we try to test it on our Testing Data which got an F1 Score and Accuracy with a value of 0.77 and 0.81.

F. Sentiment Prediction on Indonesian Digital Wallet

After predicting and classifying the reviews into 5 main aspects: Accessibility, Transaction, Account, Service, and Performance, we performed sentiment prediction of Indonesian Digital Wallet reviews.

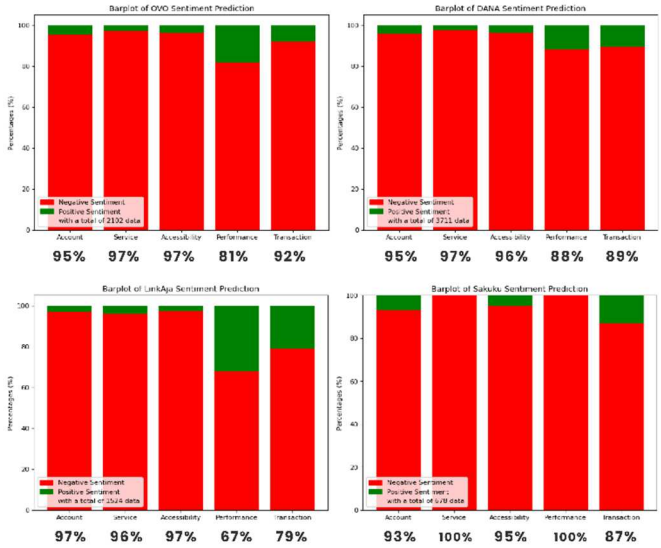


Fig. 4. Sentiment Predictions Results with a Negative Sentiment Percentages.

Fig. 4 shows the reviews of the Indonesian Digital Wallet on the Google Play Store are predicted mostly negative. It can be shown from the bar plot of LinkAja have a percentage on each aspect: Account, Service, Accessibility, Performance, Transaction respectively are 97%, 96%, 97%, 67%, 79% of negative sentiment on the review data on each aspect. We averaged the sentiment from each aspect to get the overall sentiment of each Indonesian Digital Wallet. Respectively Indonesian Digital Wallet: OVO, DANA, LinkAja, Sakuku get overall negative sentiment with a percentages to value 92.4%, 93%, 87,2%, and 95%. Therefore, the sentiment of Google Play Store reviews on each Digital Wallet with every aspects we defined and the overall percentages are tend towards negative sentiment.

IV. CONCLUSION

Our research uses the combined CNN-LSTM architecture to predict the sentiment analysis of reviews or reviews of digital wallet applications consisting of DANA, OVO, Link Aja, and Sakuku. The designed CNN-LSTM model produces 81% accuracy. Based on the review dataset that has been collected, predictions will be made to provide an overview of the digital wallet company.

Based on the experimental results of the four digital wallet applications, namely DANA, OVO, Link Aja, and Sakuku, the sentiments tend to be negative towards the categories we have created, including transactions and access, services, accounts, and performance. However, of the four digital wallet applications, most reviews are in the transaction category, which includes the following keywords: transfer, *pembelian*/buying, *pembayaran*/payment, *saldo*/balance, topup, top, up, *beli*/buy, *listrik*/electricity, *transaksi*/transaction, *biaya*/fee, *potongan*/reduction, *uang*/money. This could be considered as an input for digital wallet companies to continue to grow to serve consumers well. This can be done by optimizing the primary transaction because it has the most reviews, followed by application accounts, access, service, and application performance for the trust and business progress of the related digital wallet company.

There are still rooms for improvement to produce better model. In our approach, we found that Indonesian are tend to use colloquial language on writing the reviews. This condition may affect our model performances because of inconsistence use of language. This model can also be reused for other sentiment analysis application that applied to Indonesian text.

REFERENCES

[1] Admin , Mendidik Anak di Era Digital | Gerakan Literasi Nasional. Kemdikbud.go.id, https://gln.kemdikbud.go.id/glnsite/mendidik-anak-di-era-digital/, 2019.

[2] A. Josi, L. A. Abdillah and Suryayusra. Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah. Jurnal Sistem Informasi, Volume 5,Nomor 2, September 2014, hlm 159-164, 2014.

[3] A. Safarudin, L. Kusdibyo and W. Senalasari, Faktor-Faktor Pembentuk Loyalitas Generasi Z dalam Menggunakan Financial Technology E-Wallet, Prosiding The 11th Industrial Research Workshop and National Seminar Bandung, 26-27 Agustus 2020.

[4] Anggraeni, L., Go-Pay Jadi E-Walletl Paling Populer di Indonesia - Medcom.id. Available at: https://www.medcom.id/teknologi/news-teknologi/0Kv9o5pk-go-pay-jadi-dompet-digital-paling-populer-di-indonesia, 2019.

[5] CNN Indonesia, Konsumen Gopay Keluhkan Saldo Tak Bertambah Usai Top Up. teknologi. Available at: https://www.cnnindonesia.com/teknologi/20191026152757-185-443111/konsumen-gopay-keluhkan-saldo-tak-bertambah-usai-top-up, 2019.

[6] T. T. Mengistie and D. Kumar, "Deep Learning Based Sentiment Analysis On COVID-19 Public Reviews," 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021, pp. 444-449, doi: 10.1109/ICAIIC51459.2021.9415191.

[7] M. Khan and A. Malviya, "Big data approach for sentiment analysis of twitter data using hadoop framework and deep learning," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020, pp. 1-5.

[8] I. Kandasamy, W. Vasantha, J. M. Obbineni, and F. Smarandache, "Sentiment analysis of tweets using refined neutrosophic sets," Computers in Industry, vol. 115, p. 103180, 2020.

[9] R. P. J. Kingshotta, P. Sharmaa, and H. F. L. Chungb, "The Impact of Relational Versus Technological Resources on E-Loyalty: A Comparative Study Between Local, National, and Foreign Banks," Industrial Marketing Management, 2018.

[10] H. Luthfiah, M. A. Agmalaro, "Prediksi temporal untuk kemunculan titik panas di kabupaten rokan hilir riau menggunakan long short term

memory rnn hafshah luthfiah,", Departemen Ilmu Komputer, Institut Pertanian Bogor, 2018.

[11] Fajri Koto, Afshin, Rahimi, Jey Han Lau, and Timothy Baldwin, (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In Proceedings of the 28th COLING, December 2020.

Optimization of Fuzzy Support Vector Machine (FSVM) Model in Multiple Metric Spaces

Dita Fadma Ristianti
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, Indonesia
Dita1907050001@webmail.uad.ac.id

Sugiyrto Surono
Dept. Mathematic FAST
Ahmad Dahlan University
Yogyakarta, Indonesia
sugiyarto@math.uad.ac.id

Joko Eliyanto
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, Indonesia
joko1907050003@webmail.uad.ac.id

***Abstract*— Fuzzy membership function was introduced into the Support Vector Machine (SVM) resulting in modifications. Selecting the correct membership function is an important step in the Fuzzy Support Vector Machine (FSVM) method. One of the general criteria for selecting fuzzy membership is determined by the distance between a point and its fixed center category. This study aims to develop the SVM method into Fuzzy SVM (FSVM) with several distance functions that are applied to the Early Stage Diabetes data which collects 520 data. The distance functions used include Euclid, Canberra distance, Minkowski distance, Chebyshev distance, Minkowski Chebyshev distance, and Bray-Curtis distance where this distance function is used to determine the best distance that can be seen from the results of accuracy, specificity, g-means which is best for viewing diabetes risk. The results of this comparison show that the FSVM method with several distance functions is more than the SVM method. Where the FSVM method at the Canberra distance with a penalty value of $C = 2^5$ is the best distance to see the risk of diabetes, based on the results of specificity = 100%, g-means = 86.91%, and accuracy = 85.26% is superior to the SVM method at the penalty value $C = 2^{10}$ with specificity = 69.36%, g-means = 77.31%, and accuracy = 79.49%. Although the FSVM method produces an evaluation value at sensitivity = 75.53%, it is lower than the SVM method with a sensitivity value = 86.17%.**

Keywords—Support Vector Machine (SVM), Fuzzy Support Vector Machine (FSVM), Membership Function, Metric

I. INTRODUCTION

The last ten years, machine learning methods have been developed to aid the classification without being bound by the assumptions, and to provide greater flexibility in data analysis, but still have the accuracy and ease of use are high. Machine learning methods that have been developed one Support Vector Machine (SVM) [1]. Vapnik said, [2] defined the Support Vector Machine (SVM) method as a new machine learning method. The SVM method finds an optimal global solution, by mapping the training data to a high-dimensional space, then in a high-dimensional space it will look for a classification that maximizes the margin between the two data classes [3]. The concept of SVM is an effort to find the best hyperplane, which is used as a separator between the two classes at the input[4]. SVM is one of the featured methods of machine learning because it has good performance in completing the classification and predict cases. The principle of SVM is to find the optimal classification model or set of separators from the classification data trained by the algorithm to divide the data set into two or more different classes. These classes can help predict classes based on new data [5].

However, in the application of SVM there are many distractions that could make the data sample is not ideal. Therefore, the Fuzzy membership functions are introduced into the SVM. FSVM is very effective in many real-time applications such as credit risk evaluation, text categorization and others [6] [7] [8] [9] [10]. The facts prove that FSVM is better than SVM in dealing with noise and can effectively eliminate the influence of noise on SVM [11]. The main problem in the FSVM model is the creation of appropriate memberships to minimize outlier effect data points.[12], [13], [14] and [15] selecting the correct membership function is an important step in the FSVM method. One common criterion for selecting Fuzzy membership is determined by the distance between the point and the central category and equipment[11],[16]. "Euclidean" distance is a common metric for FSVM. As an alternative method, several distance functions are proposed to measure the distance from each point to the center of the class, this distance function will be used to determine the best point.

Utilization data mining is not limited to science and technology, but in the world of healthcare data mining is often used to treat the buildup of medical data. SVM method can be used as a reference to predict and diagnose a particular type of disease using methods that can be applied. Diabetes is a disease in the form of a metabolic disorder characterized by blood sugar levels that exceed normal limits [17] which occurs because the pancreas does not produce enough insulin (a hormone that regulates blood sugar or glucose), or when the body cannot effectively use the insulin it produces[18]. Diabetes is not an infectious disease, but WHO data shows that the percentage of non-communicable diseases in 2004 which reached 48.30% was greater than the number of presentations of infectious diseases, which was 47.50%. Even non-communicable diseases are the number one cause of death in the world (63.50%) (Islam, Ferdousi, Rahman, & Bushra, 2020). (Garnita, Society, Studies, Society, & Indonesia, 2012). Many people with diabetes are not aware of the disease, especially, because of the lack of information in the community about diabetes symptoms. Symptoms of early characteristics of people with diabetes are often referred to as triaspoli (polyuria, polydipsia, and polyphagia). This study aims to develop the SVM method into Fuzzy SVM (FSVM) with several distance functions applied to Early Stage Diabetes data which collected 520 data from Sylhet Diabetes Hospital, Sylhet Bangladesh. The distance functions used include Euclid, Canberra distance, Minkowski distance, Chebyshev distance, Minkowski Chebyshev distance, and Bray-Curtis distance where this distance function is used to determine the best distance that can be seen from the results of accuracy, specificity, g-means which is best for viewing diabetes risk. This study also tried to experiment with developing the SVM method into FSVM using various distances. This is one of the

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

novelty elements offered in this study compared to other studies. The results of the proposed method will compare the SVM method with Fuzzy SVM with several distance functions.

II. METHODS

A. Support Vector Machine (SVM)

Support vector machines (SVM) is a supervised learning method, first introduced by Vapnik in 1995 together with Bernhard Boser and Isabele Guyon [19]. [20] [6] Support Vector Machine (SVM) is a classification method that works by finding a hyperplane with optimum margins. Hyperplanea is a data dividing line between classes. Margin (m) is the distance between the hyperplane and the closest data in each class. The hyperplane can be represented as $\mathbf{w}^T x_i - b = 0$. Where x_i is the data, $y_i \in \{-1, +1\}$ is the class label of x_i, **w** is the weight vector of size *(px1)*, and **b** is the position of the plane relative to the center of the coordinates or better known as bias scalar value. The formula for the SVM optimization problem for linear classification is

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\left[\sum_{i=1}^{n}\xi_i\right] \quad (1)$$

by combining the two functions separator for both classes, then it can be represented in the inequality as follows:

$$y_i\left(x_i^T\mathbf{w}+b\right)-1-\xi_i \geq 0$$
$$y_i\left(x_i^T\mathbf{w}+b\right) \geq 1-\xi_i$$

ξ is a slack variable ξ has been added to the model for classifying data that can not be separated linearly. Where C is the major parameters that determine the penalty due to errors in classification (misclassification) data.

To determine the optimal hyperplane above it is possible to change the shape of the primal into shape Quadratic Programming (QP). Thus the optimization problem can be solved by the Karush-Kuhn-Tucker (of the summit) and formulated into a formula lagrange

$$L = \frac{1}{2}\mathbf{w}^2 + \mathrm{C}\sum_{i=1}^{\mathrm{n}}\xi_i - \sum_{i=1}^{\mathrm{n}}\alpha_i\left[y_i\left(x_i^T\mathbf{w}+b\right)-1+\xi_i\right] - \sum_{i=1}^{\mathrm{n}}\mu_i\xi_i \quad (2)$$

where α_i dan μ_i are *Lagrange Multiplier*. By minimizing *L* with **w** , b , and ξ ,

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{\mathrm{n}}\alpha_i y_i x_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i x_i$$
$$\frac{\partial L}{\partial b} = \sum_{i=1}^{\mathrm{n}}\alpha_i y_i x_i = 0$$
$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_{\mathrm{i}} = 0 \Rightarrow \alpha_i - \mu_{\mathrm{i}} = \mathrm{C}$$

with $\xi_i \geq 0, \alpha_i \geq 0, \mu_i \geq 0,\ \alpha_i\left[y_i\left(x_i^T\mathbf{w}+b\right)-1+\xi_i\right]=0,$ $\mu_i\xi_i = 0$ Thus obtained the dual problem

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad (3)$$

where $\alpha = (\alpha_i, \ldots, \alpha_n)$ is a non-negative Lagrange multiplier vector. By completing the above quadratic optimization α_i so that obtained $w = \sum_{i=1}^{n}\alpha_i y_i x_i$. Based on KKT conditions, is term bias

$$b = y_i - \sum_{i=1}^{n}\alpha_i y_i x_i \quad (4)$$

can also be computed for any supporting vector (observation that the corresponding α_i is greater than zero). The sample point x_i is classified based on the sign of its classification function as follows,

$$f(x) = sign\left(\mathbf{w}^{\mathrm{T}}\left(x_i\right)+b\right) \quad (5)$$

For the non-linear separable in feature space, kernel function $K\left(x_i, x_j\right) = \Phi\left(x_i\right)^T \Phi\left(x_j\right)$ is used to find hyperplane in a higher dimensional space, where $\Phi\left(x_i\right)$ is a non-linear mapping function.

B. Fuzzy Support Vector Machine (FSVM)

In the classification of soft intervals, the value of parameter C should not be too large or too small to ensure the effect of the classifier [11]. Training given S, where dimana $S = \left\{\left(\mathrm{x}_i, y_i, s_i\right)\right\}_{i=1}^{N}$, x_i is a sample of size n, $y_i \in \{+1, -1\}$ stating grade (+1 for positive class and -1 for negative class), and s_i is the fuzzy membership. So, the objective function is written as follows,

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\left[\sum_{i=1}^{n}s_i\xi_i\right] \quad (6)$$

by combining the two functions separator for both classes, then it can be represented in the inequality as follows:

$$y_i\left(\mathbf{\Phi}x_i^T\mathbf{w}+b\right)-1-\xi_i \geq 0$$
$$y_i\left(\mathbf{\Phi}x_i^T\mathbf{w}+b\right) \geq 1-\xi_i$$

where **w** is the vector weighting on local decisions, *b* stated bias, $\mathbf{\Phi}x_i$ a nonlinear function that maps x_i into space features high dimensional in which areas a better

decision can be found, *C* is a regularization parameter chosen beforehand to control the trade-off between margins classification and misclassification costs. Non-negatif variables ξ_i is slack variable states of x_i on SVM, while $s_i\xi_i$ is a error size with different weights according to s_i.

To solve quadratic optimization, the Lagrange Equation is as follows,

$$L=\frac{1}{2}\mathbf{w}^2+\mathrm{C}\sum_{i=1}^{\mathrm{n}}s_i\xi_i-\sum_{i=1}^{\mathrm{n}}\alpha_i\left[y_i\left(x_i^T\mathbf{w}+b\right)-1+\xi_i\right]-\sum_{i=1}^{\mathrm{n}}\mu_i\xi_i \tag{7}$$

where α_i and μ_i are *Lagrange Multiplier*. By minimizing L with to $\mathbf{w}, b$, and $s_i\xi_i$:

$$\frac{\partial L}{\partial \mathbf{w}}=\mathbf{w}-\sum_{i=1}^{\mathrm{n}}\alpha_i y_i x_i=0\rightarrow\mathbf{w}=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{\mathrm{n}}\alpha_{\mathrm{i}}y_i=0$$

$$\frac{\partial L}{\partial \xi_i}=s_iC-\alpha_i-\mu_{\mathrm{i}}=0\rightarrow\alpha_i-\mu_{\mathrm{i}}=s_iC$$

$$L=\sum_{i=1}^{\mathrm{n}}\alpha_{\mathrm{i}}-\frac{1}{2}\sum_{i=1}^{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\alpha_i\alpha_i y_i y_i x_i^T x_i \tag{8}$$

C. Fuzzy Membership Function u for FSVM

Ding Xiaokang [11] explains that FSVM models adopting the conventional method of calculating membership, which determines the class centers by averaging all of the samples. By using the distance from each sample point to the center of the class as d_i, then the membership function can be expressed as:

$$s_i=\begin{cases}1-\dfrac{d_{i+}}{r_{i+}+\beta}, y_i=+1\\ 1-\dfrac{d_{i-}}{r_{i-}+\beta}, y_i=-1\end{cases} \tag{9}$$

$$s_i=\begin{cases}1-\dfrac{\left\|x_i^+-x_{cen}^+\right\|}{max\left(\left\|x_i^+-x_{cen}^+\right\|\right)+\beta}\rightarrow y_i=+1\\ 1-\dfrac{\left\|x_i^+-x_{cen}^+\right\|}{max\left(\left\|x_i^+-x_{cen}^+\right\|\right)+\beta}\rightarrow y_i=-1\end{cases} \tag{10}$$

Where δ is positive value used to avoid *s* to zero, while *d* represents the Euclidean distance from each sample to the class center.

β = constant to avoid $s_i=0$

$d_{i+}=\left\|x_i^+-x_{cen}^+\right\|$

$d_{i-}=\left\|x_i^--x_{cen}^-\right\|$

$r_{i+}=\max d_{i+}$

$r_{i-}=\max d_{i-}$

$x_{cen}^+=$ positive sample center

$x_{cen}^-=$ negative sample center

$x_i^+=$ labelled sample $y_i=1$

$x_i^-=$ labelled sample $y_i=-1$

This function indicates that the closer to the center of the class, the greater the value of membership, and the smaller the contrary.

D. Metric

1. Minkowski Distance

The Minkowski distance is a generalization of the distance matrix, defined as follows:

$$d_{\min}\left(x,y\right)=\left(\sum_{i=1}^{d}\left|x_i-y_i\right|^r\right)^{\frac{i}{r}},r\geq1 \tag{11}$$

where r is a Minkowski parameter, at Euclidean (r = 2) and Manhattan (r = 1) distances. Metric conditions are met as long as p is equal to or greater than 1[21].

2. Chebyshev Distance

The Chebyshev distance is the variance of the Minkowski distance where,

$$p\rightarrow\infty$$
$$d_{cbc}\left(x,y\right)=\max_{k=1}^{n}\left|x_i-y_i\right| \tag{12}$$

where x_i and y_i are nilai the values of x and y in dimension n [21]

3. Canbera Distance

The Canbera distance is given as follows:

$$d(x,y)=\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i|+|y_i|} \quad (13)$$

Canberra distances can perform very well, significantly better than the most used Manhattan and Euclidean distances, as shown [22] This distance tests the sum of the series of fractional differences between the coordinates of a pair of vectors [23].

4. Minkowski Chebyshev Distance

Rodriguez [24] brings up a new distance, namely the combination of the Minkowski and Chebyshev distances. The combination of the Minkowski and Chebyshev distances is shown in the following definition:

$$d_{(w_1,w_2,d)}\left(\bar{x},\bar{y}\right)=w_1 d_{mkw}\left(\bar{x},\bar{y}\right)+w_2 d_{cheb}\left(\bar{x},\bar{y}\right) \quad (14)$$

Or

$$d_{(w_1,w_2,r)}\left(\bar{x},\bar{y}\right)=w_1\left(\sum_{i=1}^{d}|x_i-y_i|^r\right)^{\frac{i}{r}}-\max_{k=1}^{n}|x_i-y_i|,1\le k\le n \quad (15)$$

where x_i and y_i are the value to $-i$ on two vectors $\bar{x}$ and $\bar{y}$, and vice versa on the dimension n

5. Bray-Curtis Distance

The Bray-Curtis distance, sometimes also called the Sorensen distance, is commonly used in ecology and environmental sciences. This distance view space as a lattice that is similar to the distance of a city block. The Bray-Curtis distance has the nice property that if all coordinates are positive, the value is between zero and one. If both objects are at zero coordinates, the Bray-Curtis distance is not specified. [23]

$$d(x,y)=\sum_{i=1}^{n}\frac{|x_i-y_i|}{x_i+y_i} \quad (16)$$

where,

$d=$ distance between x and y

$x=$ cluster center data

$y=$ data on attributes

E. Classification Evaluation

The actual data and the predicted data from the classification model are presented using a cross tabulation (Confusion matrix), which contains information about the actual data class represented in the row matrix and the predicted data class in the column[19].

Accuracy is an evaluation matrix that is very important to assess the performance of an overall classification results. The higher the classification accuracy of classification techniques also means that the performance is getting better. [25] explained that the evaluation of the performance of a classifier in the imbalance class can be measured using the G-mean. Sensitivity is a performance measure to measure the positive class or the accuracy of the positive class. Specificity is a performance measure to measure the negative class or the accuracy of the negative class.

Table 1. Confusion Matrix

Actual	**Predicted**	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Information:

TP : *True Positive* (the number of correct predictions in the positive class)

FP : *False Positive* (the number of wrong predictions in the positive class)

FN : *False Negative* (the number of incorrect predictions in the negative class)

TN : *True Negative* (the number of correct predictions in the negative class)

Accuracy

Accuracy assesses the overall effectiveness of the algorithm by estimating the correct value of the class label. The Accuracy Value is stated as follows

$$Accuracy=\frac{TN+TP}{TN+TP+FN+FP} \quad (17)$$

Sensitivity (SE)

Sensitivity is a performance measure to measure the positive class or the accuracy of the positive class. The sensitivity value states how many positive class samples are correctly labeled. The sensitivity value is stated as follows.

$$Sensivity=\frac{TP}{TP+FN} \quad (18)$$

Specificity (SP)

Specificity is a performance measure to measure the negative class or the accuracy of the negative class. The specificity value states how many samples of the negative class are correctly labeled. The specificity value is stated as follows.

$$Specifity = \frac{TN}{TN + TP} \quad (19)$$

G-means (GM)

[26] said that the g-mean value was used to evaluate the performance of the algorithm on imbalanced data problems. G-means is the product of the prediction accuracy for both classes which includes accuracy in the positive class (sensitivity) and accuracy in the negative class (specificity). This value shows the balance between the classification performance of the majority and minority classes. poor performance in positive sample prediction will result in a low G-means value as well as for the negative class. The g-means value is expressed as follows.

$$g - mean = \sqrt{Sensivity\ x\ Specifity} \quad (20)$$

III. RESULT AND DISCUSSION

In this study the type of data used is secondary data obtained from the official page through https://archive.ics.uci.edu/ml/datasets.php. Data collected in the article amounted to 520 using questionnaires data taken directly from the patient's Hospital ethical standards institutions in which research is conducted and ethical approval was obtained from the Hospital Diabetes Sylhet, Bangladesh Sylhet. The factors that influence the risk of diabetes are 16 as the x_i variable and the y variable as the class label of the x_i variable with members {1,-1}, where 1 is for the class that is not at risk of developing diabetes and -1 for the class that is at risk of developing the disease diabetes. The steps in conducting the analysis in this study are as follows

a) Exploration to see the characteristics of the data.

b) Divide the data into training and testing data.

c) SVM classification on the training data and evaluate the classification performance on the test data.

d) FSVM classify the training data using Euclidean metrics, Canberra, Minkowski, Chebyshev, Minkowski, Chebyshev, and Bray-Cutris and evaluate the classification performance on the test data.

- Calculates Euclid, Canberra, Minkowski, Chebyshev, Minkowski-Chebyshev, and Bray-Cutris matrices from data points to class center.
- Calculate the value of membership function

e) Comparing the performance of SVM and FSVM classification with several matrix models to see the best classification results.

Before the SVM modeling data is divided into training and testing. In this study, the data used amounted to 520 cases divided into training data of 70%, namely 364 cases and testing data of 30%, namely 156 cases. This SVM method uses a polynomial kernel with different C penalty values to see the best accuracy results.

Table 2. Classification

MODEL	C	SE	SP	GM	Accuracy
SVM	2^1	84,04%	67,74%	75,45%	77,56%
	2^5	74,47%	62,90%	68,44%	69,87%
	2^{10}	**86,17%**	**69,36%**	**77,31%**	**79,49%**
FSVM-1	2^1	72,34%	100%	85,68%	83,97%
	2^5	72,34%	100%	85,05%	83,33%
	2^{10}	71,28%	100%	84,43%	82,69%
FSVM-2	2^1	74,47%	100%	86,30%	84,62%
	2^5	**75,53%**	**100%**	**86,91%**	**85,26%**
	2^{10}	71,28%	100%	84,43%	82,69%
FSVM-3	2^1	73,40%	100%	85,68%	83,97%
	2^5	72,34%	100%	85,05%	83,33%
	2^{10}	71,28%	100%	84,43%	82,69%
FSVM-4	2^1	72,34%	100%	85,05%	83,33%
	2^5	71,28%	100%	84,43%	82,69%
	2^{10}	71,28%	100%	84,43%	82,69%
FSVM-5	2^1	72,34%	100%	85,05%	83,33%
	2^5	72,34%	100%	85,05%	83,33%
	2^{10}	71,28%	100%	84,43%	82,69%
FSVM-6	2^1	72,34%	100%	85,05%	83,33%
	2^5	72,34%	100%	85,05%	83,33%
	2^{10}	71,28%	100%	84,43%	82,69%

Table 3. Classification Performance The results of the SVM classification performance at different C penalty values resulted in the values of sensitivity, specificity, G-means, and accuracy. On the SVM classification, it can be seen that the best classification performance evaluation is given by a penalty value of $C = 2^{10}$ with an evaluation value of sensitivity 86.170%, specifity 69.355%, G-means 77.307% and accuracy 79.487%. FSVM Classification at the Euclidean Distance (FSVM-1), the best classification performance evaluation results are given by a penalty value of $C = 2^1$ with the same evaluation values, namely sensitivity 72.340%, specifity 100%, G-means 85.676% and accuracy 83.974. FSVM Classification at the Canberra Distance (FSVM-2), it can be seen that the best classification performance evaluation results are given by a penalty value of $C = 2^5$ with an evaluation value of 75.532% sensitivity, 100% specificity, 86.909% G-means and 85.256% accuracy. FSVM Classification at the Minkowski Distance (FSVM-3), it can be seen that the best classifi ation performance evaluation results are given by a penalty value of $C = 2^1$ with an evaluation value of 73.404% sensitivity, 100% specificity, 85.676% G-means and 83.974% accuracy. Furthermore, the results of the evaluation of the FSVM classification at the Chebyshev distance will c be given.

FSVM Classification at the Chebyshev Distance (FSVM-4), it can be seen that the best classification performance evaluation results are given by a penalty value of $C = 2^1$ with evaluation values of 72.340% sensitivity, 100% specificity, 85.053% G-means and 83.333% accuracy. Furthermore, the results of the evaluation of the FSVM classification at the Minkowski-Chebyshev distance will be given. FSVM Classification at the Minkowski-Chebyshev Distance (FSVM-5), it can be seen that the best classification performance evaluation results are given by a penalty value of $C = 2^1$ with an evaluation value of 72.340% sensitivity, 100% specificity, 85.053% G-means and 83.333% accuracy. Furthermore, the results of the evaluation of the FSVM classification at the Bray-Curtis distance will be given. Similarly in the FSVM-4 and FSVM-5, the results of the best classification performance evaluation in FSVM-6 are given by a penalty value of $C = 2^1$ with the same evaluation values, namely sensitivity 72.340%, specifity 100%, G-means 85.053% and accuracy 83.333%.

From the performance of SVM and FSVM classification with several distance functions, it can be seen that the results of Fuzzy SVM give the best results to see the risk of diabetes. It can be seen in Table 2. the sensitivity value (SE) of the SVM method is superior with the highest percentage of 86.17% at $C = 2^1$ while the FSVM method with several distance functions gives the highest percentage of 75.53% at $C = 2^5$. However, in terms of specificity (SP), g-means (GM), and accuracy for all C penalty values, the FSVM method with several distance functions is very superior to the SVM method. The specificity value (SP) of the FSVM method with several distance functions gives an average percentage result of 100% while the SVM method has the highest specificity (SP) value with a percentage of 69.36% at $C = 2^{10}$, the value of g-means (GM) method FSVM with several distance functions gives the highest percentage of 86.91% at C= 2^5 while the SVM method has the highest g-means (GM) $C = 2^5$alue with a percentage of 77.31%. The accuracy value of the FSVM method with several distance functions gives the highest percentage of 85.256% at $C = 2^5$.

IV. CONCLUSION

In this paper, a method for developing SVM into FSVM has been presented with several distance functions including Euclid distance, Euclid distance, Canberra distance, Minkowski distance, Chebyshev distance, Minkowski Chebyshev distance, and Bray-Curtis distance where this distance is used to determine the best distance that can be seen from the results. the best accuracy, sensitivity, specificity, g-means. We applied the FSVM method with multiple distance functions to the Early Stage Diabetes data. The results of this comparison show that the FSVM method with several distance functions is more than the SVM method. Although the sensitivity (SE) value of the SVM method is superior, for the value of specificity (SP), g-means (GM), and accuracy on all C penalty values, the FSVM method with several distance functions is very superior to the SVM method.

ACKNOWLEDGMENT

Thank you to the leaders of Ahmad Dahlan University (UAD) and the UAD Postgraduate Program, and the UAD Mathematics Education Study Program who have facilitated this research. Thank you also said to reviewers who have provided input and comments so that this article becomes more qualified.

REFERENCES

[1] B. Scholkopf and A. Smola, Learning with Kernel :Support Vector Machines, Regulerizatiom, Optimization, and Beyond. 2002

[2] Vladimir. N. Vapnik., "The Nature of Statistical Learning," *Theory*. 1995

[3] L. Evans, N. Lohse, and M. Summers, "A fuzzy-decision-tree approach for manufacturing technology selection exploiting experience-based information," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6412–6426, 2013

[4] W. Purnami, A. M. Regresi, and L. Ordinal, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *J. Sains Dan Seni Its*, vol. 1, no. 1, 2012

[5] W. Purnami, A. M. Regresi, and L. Ordinal, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *J. Sains Dan Seni Its*, vol. 1, no. 1, 2012

[6] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 6, pp. 820–831, 2005

[7] Z. Zhiwang, G. Gao, and Yong Shi, "Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors," *Eur. J. Oper. Res.*, vol. 1, no. 237, pp. 335–348, 2014

[8] T.-Y. Wang and H.-M. Chiang, "Fuzzy support vector machine for multi-class text categorization," *Inf. Process. Manag.*, vol. 4, no. 43, pp. 914–929, 2007

[9] M. Perez, D. M. Rubin, L. E. Scott, T. Marwala, and W. Stevens, "A hybrid fuzzy-SVM classifer, applied to gene expression profling for automated leukaemia diagnosis," in *Conference: EduPath ConferenceAt: Cape Town, South Africa*, 2008, pp. 3–5

[10] N. Ö. Özcan and F. Gürgen, "Fuzzy support vector machines for ECG arrhythmia detection," *Proc. - Int. Conf. Pattern Recognit.*, pp. 2973–2976, 2010

[11] X. K. Ding, X. J. Yang, J. Y. Jiang, X. L. Deng, J. C. Cai, and Y. Y. Ji, "Optimization and analysis on fuzzy SVM for objects classification," *J. Inf. Hiding Multimed. Signal Process.*, vol. 9, no. 6, pp. 1421–1429, 2018

[12] X. Jiang, Z. Yi, and J. C. Lv, "Fuzzy SVM with a new fuzzy membership function," *Neural Comput. Appl.*, vol. 15, no. 3–4, pp. 268–276, 2006

[13] A. Shilton and D. T. H. Lai, "Iterative fuzzy support vector machine classification," *IEEE Int. Conf. Fuzzy Syst.*, pp. 1391–1397, 2007

[14] W. M. Tang, "No TitleFuzzy SVM with a New Fuzzy Membership Function to Solve the Two-Class Problems," *Neural Process. Lett.*, vol. 3, no. 34, pp. 209–219, 2011

[15] H. Li, F. Qi, and S. Wang, "A comparison of model selection methods for multi-class support vector machines," *Lect. Notes Comput. Sci.*, vol. 3483, no. IV, pp. 1140–1148, 2005

[16] C. F. Lin and S. De Wang, "Fuzzy support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002

[17] Kementrian kesehatan republik indonesia, InfoDATIN, Pusat Data dan Informasi Kementrian Kesehatan RI. 2020

[18] Kementrian kesehatan republik indonesia, InfoDATIN, Pusat Data dan Informasi Kementrian Kesehatan RI. 2018

[19] J. Han and M. Kamber, Data Mining Consepts And Techniques 3 edition. San Fransisco. 2012

[20] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. 1981

[21] B. Tirozi, D. B., and E. F, Introduction To Computational Neurobiology And Clustering. 2007

[22] M. Kokare, B. N. Chatterji, and P. K, "Biswas. Comparison of similarity metrics for texture image retrieval," *TENCON 2003*, vol. 2, pp. 571–575, 2003

[23] Jelínek. Jiří, Metrics in Similarity Search. 2007

[24] E. O. Rodrigues, "Combining Minkowski and Chebyshev: New Distance Proposal and Survey of Distance Metrics Using k-Nearest Neighbours Classifier," 2018

[25] K. Lokanayaki and A. Malathi, "A Prediction for Classification of Highly Imbalanced Medical Dataset Using Databoost.IM with SVM," 2014

[26] Q. Li, Y. Wang, and S. H. Bryant, "A novel method for mining highly imbalanced high-throughput screening data in PubChem," *Bioinformatics*, vol. 25, no. 24, pp. 3310–3316, 2009

Outlier Detection Using K-Means Clustering with Minkowski-Chebyshev distances for Inquiry-Based Learning Results in Students Dataset

Endang Wahyuni
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, INDONESIA
endang1907050001@webmail.uad.ac.id

Sugiyarto Surono
Dept. Mathematic FAST
Ahmad Dahlan University
Yogyakarta, INDONESIA
sugiyarto@math.uad.ac.id

Joko Eliyanto
Master of Mathematics Education
Ahmad Dahlan University
Yogyakarta, INDONESIA
joko1907050003@webmail.uad.ac.id

***Abstract*— Outlier appears as an extreme value but often contains very important information, so it is necessary to be studied whether the data remains used or issued. Outlier detection is a hot topic for the study. Increasing new technologies and various applications cause increased requirements of outlier detection. The Outlier method is successfully applied in various fields, namely: economy, business, health, space, geology, and education. Implementation of an outlier of analysis in the education field is often applied to the evaluation of the learning model. Inquire -based learning model is an important component in educational renewal. Learning by this method encourages learners to learn mostly through their active involvement. This study aims to discuss outlier detection by using the K-Means Clustering method on the inquiry-based learning results in students. This study detects outliers with the K-means method using Minkowski-Chebyshev distance. The result of the proposed method will be compared with the extremes of standard deviation (ESD), Box-Plot, and K-Means Clustering using Euclidean distance. The outlier detection results using K-Means Clustering with the Minkowski-Chebyshev and Euclidean distance produce the same result that can detect 3 data as an outlier that is the student with the ID number 7 Exam Value 7.5, ID number 42 Exam Value 9.0, and ID number 72 with the value of 13.5. While the ESD method and Box-Plot are unable to detect any outlier.**

Keywords — Outlier Detection; K-Means Clustering; Inquiry based learning; ESD; Box-Plot;

I. INTRODUCTION

Data that have large volumes, various types of data, and very fast data speeds called Big Data [1] [2] [3]. Variable data can be represented as variables, while variables are seen as dimensions [2] clustering is one of the unsupervised learning methods, wherein terms of data cluster or value do not have a target or have no class label [4]. Clustering is the process of grouping the data into groups or clusters. Each cluster has data that has a high resemblance, and between clusters has a low resemblance [4].

The Outlier Detection method is the most important method of data analyzing, such as decision making, grouping, and pattern classification. Outlier is defined as observations that do not match the overall pattern of grouping [1] [2] [3] [5] [6]. Outlier detection is an important subject in data mining. Outlier detection is widely used to identify and eliminate ordinal or irrelevant objects of data set [2] [6] [7] [8] [9]. However, the main challenge of outlier detection is increased complexity due to diverse datasets and dataset size [7]. The outlier factor of the cluster determines the degree of difference from a cluster of the whole dataset [10]. Implementation of outlier analysis in the education field is often applied to the evaluation of the learning model. Inquire-based learning (IBL) model is an important component in educational renewal. The Inquiry learning model is a process of learning-oriented to the activity of the learners in the process of investigation and discovery of solutions from the issue issued [15]. Inquiry learning puts learners as a subject of learning. Learners play a role to discover the core of the material, while educators play a participant as actors and act as facilitators, motivators for learners. The IBL model stage consists of three stages: producing hypothesis and investigations and discovery and reflection. Based on the relevance of each Inquiry-based Learning step with the indicator of critical thinking ability of student mathematics. In addition, Inquiry Learning focuses on students inactivating in an investigation that leads to students to logical generalizations. These activities have the potential in facilitating the enhanced critical thinking ability. Learning models that give students more freedom have the potential to produce very varied results. Including the occurrence of outliers. This study attempts to detect outliers in the inquiry-based learning dataset. The existing clustering method has fixed outlier determination criteria. This often makes data that is actually quite different from other data not detected as outliers. A method that has a flexibility threshold is needed to determine outliers. The K-means clustering method was chosen to be developed as an outlier detection because of its ability to group data based on the similarity between the data. The greater the difference in a dataset member, it can be concluded that the data is an outlier. The value constraint that determined whether a dataset member was an outlier or not is called a threshold. In the k-means method, the threshold setting is very possible and depends on the distance function used. This study also tried to experiment with the development of an outlier detection method on K-means clustering using Minkowski-Chebyshev distances. This is one element of the novelty offered in this research compared to other studies. The results of the proposed method are then compared with the commonly used outlier detection methods, ESD, Box-Plot rule, and K-Means Clustering with Euclidean distance.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

II. METHODS

A. Outlier Detection

Outlier detection is obtained data that appears with extreme values both univariate and multivariate. The extreme is a far or different value at all with most other grades in its group [2]. Outlier detection is an important subject in data mining. Outlier detection is widely used to identify and eliminate ordinal or irrelevant objects of data set [2] [6] [7] [8] [9]. However, the main challenge of outlier detection is when increased complexity due to diversity dataset and dataset size [7]. The outlier factor of the cluster determines the degree of difference from a cluster of the whole dataset [10]. Outlier detection is used to find fraudulent data. Researchers research data groupings and Outlier Detection process [6] 10]

B. K-Means Clustering

The K-Means clustering the data that exist into some clusters with the criteria in the same cluster has the same characteristics, and have different clusters with data in other clusters. This algorithm is most widely used among all clustering algorithms due to its efficiency and its simplicity [8]. In this algorithm, the number of clusters is assumed to remain.

Suppose D is a dataset with a number of n rows, eg $C_1, C_2, \cdots, C_m$ a is a separate cluster inside D. Then the error function is defined as follows:

$$E = \sum_{i=1}^{m} \sum_{x \in C_i} d\left(\vec{x}, \mu(C_i)\right) \quad (1)$$

$\mu(C_i)$ = cluster center (centroid) of C_i

$d\left(\vec{x}, \mu(C_i)\right)$ = the distance between data $\vec{x}$ to the centroid of cluster $\mu(C_i)$

- Euclidean Distance

The Euclidean Distance is one of the distance calculation methods used to measure the distance of 2 (two) fruit points in Euclidean Space (covering the two-dimensional euclidean field, three dimensions, or even more). To measure the level of similarity of the data with the Euclidean Distance formula used the following formula:

$$d(\vec{x}, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (2)$$

where $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ are two input vectors with quantitative features m. In the Euclidean distance function, all features contribute the same at the function value [11] [12].

- Minkowski-Chebyshev Distance

Rodrigus (2018) raises a new distance that is a combination of Minkowski and Chebyshev. The combination of Minkowski and Chebyshev distance is shown in the following definitions:

$$d(w_1, w_2, r) = w_1 (\sum_{i=1}^{m} |x_i - y_i|^r)^{\frac{1}{r}} + w_2 (max_{i=1}^{m} |x_i - y_i|), 1 \leq k \leq m \quad (3)$$

where $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ are two input vectors with quantitative features m.

C. Box-Plot-Rule

The standard Box-Plot, each has a top(U) and bottom(L) limit, defined as:

$$U = Q_3 + 1.5(Q_3 - Q_1)$$

$$L = Q_1 - 1.5\left(Q_3 - Q_1\right) \quad (4)$$

the value that falls outside the limit is considered as outliers. This rule has a higher chance to detect false outliers than typical informal test [13]. Here Details How to Specify the Restrictions that can be seen in Figure 1.

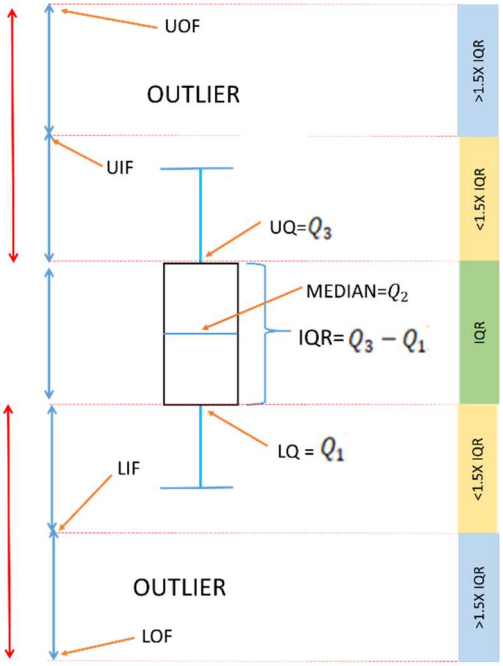


Fig. 1. Determining the Boundaries on The Box-Plot-Rule Rules

D. ESD Test

Standard Deviation is a test based on the extreme studentized deviation (ESD) which is quite good at detecting outliers in a random normal sample. Defined x_j as outlier when:

$$G_j = \max_i \left\{ \frac{|x_i - \bar{x}|}{s_j} \right\}, i = 1, \cdots, m \quad (4)$$

x_j is declared as outlier when $\frac{|x_i - \bar{x}|}{s_j}$ is largest. In principle, if G_j doesn't exceed the critical value, then no x_j need not be selected. If this test finds outliers, further outlier testing is carried out by deleting observations x_j and repeating the process on the remaining n-1 observations [14].

E. Inquiry -based learning

Inquire-based learning is oriented to the activity of the learners in the process of investigation and discovery of solutions from the issue issued [15]. According to [16] Learning outcomes is an overview of the ability of learners in meeting the stage of learning achievement in a basic

competence (KD). Learning outcomes can be used as a benchmark or criteria in achieving educational goals. Inquire learning puts learners as a subject of learning. Learners play a role to discover the core of the material, while educators play a participant as actors and acts as facilitators, motivators for learners. Inquiry-based learning is an important component in educational renewal. Because learning with this method encourages learners to learn mostly through their active involvement, learners are expected to take the initiative, they are trained how to solve the problem, make decisions, and acquire skills. Educators encourage learners to have real and creative experiences.

Inquire-based learning is an important component of educational reform. Because learning whit this method encourages students to learn mostly through their active involvement, students are expected to take the initiative, they are trained how to solve problems, make decisions, and acquire skills/ educators encourage students to have real and creative experiences.Standard Deviation is a test based on the extreme studentized

III. RESULT AND DISCUSSION

In this study, the type of data used is secondary data obtained from the official page https://archive.ics.uci.edu/ml/index.php. The data used are the learning outcomes of undergraduate students using the Deeds e-learning (Digital Electronics Education and Design Suite) with an inquiry-based learning model (IBL) [17]. Deeds is a simulation application for e-learning in digital applications. This application provides learning materials through a special browser for students, and students are asked to solve various problems with different levels of difficulty. This application has been effective, for more than ten years, in teaching and improving student learning outcomes because it provides a highly interactive simulator. Digital electronics courses are organized in separate theory and laboratory sessions where students work with Deeds simulators. The inquire-based student learning process is applied in the context of education with the Deeds simulator.

At the beginning of each session, problem-solving exercises are given to students. For each exercise, students follow a learning process that involves understanding a given problem and dividing it into various tasks, making observations in a simulated environment, conducting experiments to find the answers, and finally explaining and justifying their solutions and the methods used. The final exam questions discuss the concept of each practice session. The exam questions consist of 6 practice sessions with a total of 16 description questions with each item having a different value according to the weight of the question.

At this stage is the process of selecting the relevant data. Irrelevant data will be removed. The data is selected and selected according to the calculation, where irrelevant data will be removed from the study, so it must be selected first which data will be used and which will be removed from the study. Many students do not take the exam, so the student data will be deleted from the exam results. Of the 115 students who attended the training, 22 students did not take the exam and only 93 students who took the final exam. So that 22 students who do not take the final exam will be removed or deleted. The results of student selection are in table 1.

TABLE I. THE RESULTS OF STUDENT SELECTION

Student ID	Score	Student ID	Score	Student ID	Score
1	94,5	37	30,0	72	95,0
2	44,0	38	41,5	73	49,0
3	85,0	39	83,5	74	18,0
4	30,0	41	98,0	75	87,5
5	38,5	42	22,5	76	93,5
6	82,0	44	97,5	77	66,5
7	78,0	45	46,0	78	78,0
8	8,5	46	22,0	79	84,5
9	18,5	47	28,0	80	51,5
10	59,0	48	71,5	81	87,0
11	60,0	49	30,5	82	13,5
12	40,5	51	9,0	83	52,0
13	90,0	52	36,0	85	82,0
14	64,0	53	70,5	86	94,0
15	67,5	54	39,0	87	74,5
16	67,0	55	36,5	88	96,0
17	97,0	56	84,0	89	16,5
18	62,0	57	23,0	91	66,5
19	50,0	58	48,0	92	35,0
20	97,5	59	40,0	93	66,0
22	40,0	60	16,5	94	92,5
24	70,5	61	57,5	95	52,5
25	57,5	62	43,0	96	77,5
27	74,5	63	69,0	98	66,5
28	79,5	64	20,5	99	17,0
29	97,0	66	86,0	100	83,0
30	55,5	67	92,0	101	32,0
32	75,5	68	91,0	102	31,5
33	30,5	69	23,5	103	18,5
34	18,0	70	60,5	104	92,0
36	79,0	71	90,5	106	71,0

After the data selection process, the next step is to visualize the distribution plot of the data that has been selected. The results of the data plot can be seen in Figure 2. In Figure 2 we cannot easily see the outliers. However, as an evaluation, the value of outliers on learning outcomes is very important to obtain.

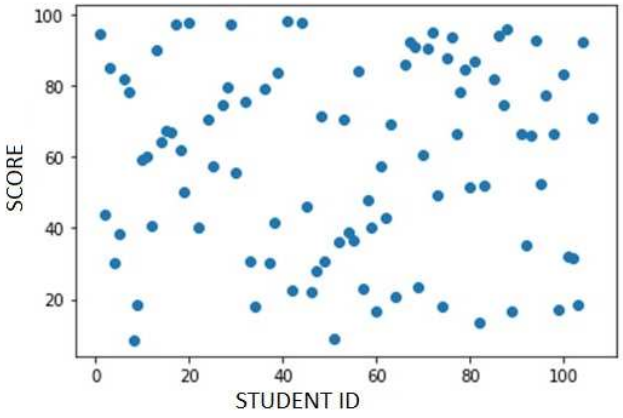


Fig. 2. The Plot of Data Selection Results

Here are examples of Outlier Detection Steps using K-Means with Minkowski-Chebyshev Distance

Step 1: Calculate the cluster center of the initial matrix that is already raised at the start of the centroid is 9.0

Step 2: Defining the Minkowski-Chebyshev distance by using equation (3)

Step 3: Determining the distance between the early centroid with the data

Step 4: Calculating new centroid. Retrieved new centroid: 58.693548387096776

Step 5: Repeat step 3 and step 4 to convergent conditions, when the Centroid value does not change.

Step 6: Defining the outlier by determining the threshold magnitude. By determining that the threshold that so a data that has the value, nature, or different characteristics than the object is generally referred to as an outlier. The threshold amount used is (max distance- min distance) * 90%. The threshold value obtained 44,8983870967742

Step 7: Determining data that is the datal outer is said to be outlier if the distance between the centroid with data> threshold magnitude. So that we get 3 data which are outliers including students with the number ID 7 Exam Value 8.5, ID number 42 exam value of 9.0, and ID number 72 with value 13.5. which can be seen in Figure 3.

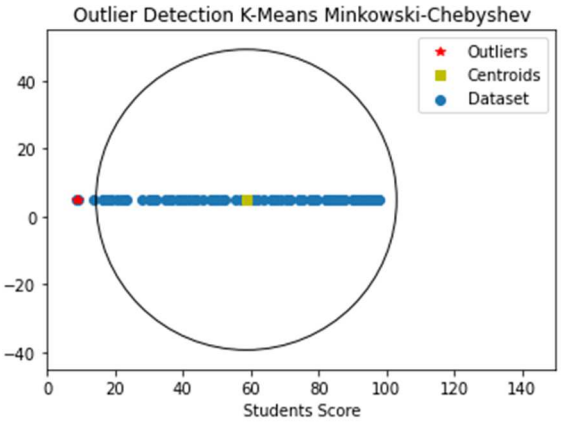


Fig. 3. . Plot of Outlier Detection Results using K-Means Clustering with Minkowski-Chebyshev distance

Figure 3 is a plot of outlier detection results using the K-Means Clustering method with Minkowski-Chebyshev distance. The threshold used is (max distance- min distance) * 90%. Based on the results of the threshold value, this method is able to detect outliers. Based on the picture, it can be seen that there are 3 pieces of data that are outliers. Namely students with ID number 7 with a test score of 8.5, ID number 42 with a test score of 9.0, and ID number 72 with a score of 13.5. The same steps were also carried out on the Outlier Detection Method using K-Means with a Euclidean distance.

The evaluation of outlier detection on the inquiry-based learning results in students using K-Means Clustering with the Minkowski-Chebyshev distance with the Comparative of ESD, Box-Plot and K-Means Clustering with Euclidean distance methods in Table 2. Apparently that the method filed can run well and able to produce the outlier as needed while the other two methods are ESD and IQR do not work. The computational speed of the method is also relatively faster than the IQR method and has a number of lesser iterations than ESD. In the table can also be seen that outlier analysis using K-Means with the Minkowski-Chebyshev and Euclidean distance produces the same result. Although at different speeds.

TABLE II. OUTLIER DETECTION EVALUATION RESULTS

Method	Number of iterations	Time	Result	Information
K-Means Clustering using Euclidean Distance	2	0.0027	Able to detect 3 data as outlier	Student ID 7 Exam Value 8,5; Student ID 42 Exam Value 9,0; and Student ID 72 Exam Value 13,5
K-Means Clustering using Minkowski -Chebyshev Distance	2	0.0039	Able to detect 3 data as outlier	Student ID 7 Exam Value 8,5, Student ID 42 Exam Value 9,0, and Student ID 72 Exam Value 13,5
ESD	10	0.0005	Unable to detect outliers	
Box-Plot	1	0.0120	Unable to detect outliers	

IV. CONCLUSION

In this paper we have presented several methods to detect any outlier such as using K-Means Clustering with Minkowski-Chebyshev and Euclidean distance, as well as comparative methods are ESD and Box-Plot. We apply the method to detect outliers on student data results by S1 students using E-Learning Deeds (Digital Electronics Education and Design Suite) with Inquiry-based learning model (IBL). The results of this comparison show that of the four methods of outlier detection using K-Means Clustering with Euclidean distance is superior to other methods, although the K-Means Clustering method with Minkowski-Chebyshev distance can also produce the number of outliers and the same number of iterations but at different speeds. While the ESD method and Box-Plot in this case can't afford to detect any outlier. For further research we intend to explore the outlier detection for other issues, namely for multivariate data.

ACKNOWLEDGMENT

Thank you to the leaders of Ahmad Dahlan University (UAD) and the UAD Postgraduate Program, and the UAD Mathematics Education Study Program who have facilitated this research. Thank you also said to reviewers who have provided input and comments so that this article becomes more qualified.

REFERENCES

[1] V.Chandola, A. Banerjee, and V. Kumar, "Survey of Anomaly Detection," ACM Comput. Surv., vol 41, noo.3, pp. 1-72, 2009

[2] B. Marr, "Big Data In Practice" United Kingdom: TJ International Ltd, Padatow, Cornwall, 2016

[3] V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies" Artif. Intell. Rev., vol. 40 no.2, pp.85-126, 2004

[4] J. Han, M. Kamber, and J. Pei, "Data Mining Consepts and techniques",2006

[5] S. Kim "Variable Selection and Outlier Detection for Automated K-Means Clustering" vol. 22, no. 1, pp. 55-67, 2015

[6] J.B. Macqueen “Soe Methods for Classification and Analysis of Multivariate Observations” Symp. Math. Stat. Probab., vol.1 pp. 281-297, 1967

[7] V. Bhatt, M. Dhakar, and B. K. Chaurasia ““Filtered Clustering Based on Local Outlier Factor in Data Mining,” Int. J. Database Theory Appl., vol. 9, no. 5, pp. 275–282, 2016

[8] T. Christopher, “A Study of Clustering Based Algorithm for Outlier Detection in Data streams,” no. March, pp. 194–197, 2015

[9] Y. Erdem and C. Ozcan, “K-Means Clustering on Apache Spark,” no. 7, pp. 86–90, 2017

[10] C. Sumithiradevi and M. Punithavalli, “Enhanced K-Means with Greedy Algorithm for Outlier Detection,” Int. J. Adv. Res. Comput. Sci., vol. 3, no. 3, pp. 294–297, 2012.

[11] J. Ren, “A detection algorithm of customer outlier data based on data mining technology,” vol. 33, no. Febm, pp. 272–278, 2017

[12] J. Han, M. Kamber, and J. Pei, Data Mining Consepts And Techniques, 3rd ed. San Fransisco: Morgan Kaudmann, 2012

[13] M. Kuppusamy and K. S. Kannan, “Comparison of methods for detecting outliers,” Int. J. Sci. Eng. …, no. January 2013,

[14] B. Iglewicz and D. C. Hoaglin, How to Detect and Handle Outliers, 16th ed. United States of America §: ASQC Quality Press Publications Catalog, 1993

[15] B. Joyce and M. Weil, “Attaining concepts: The basic thinking skills,” Model. Teach., pp. 161–178, 2003.

[16] W. Sanjaya, Strategi Pembelajaran Berorientasi Standar Proses Pendidikan. Jakarta: Kencana Perdana Media Group, 2008

[17] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, “A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator,” Springer Int. Publ. Switz., pp. 352–366, 2015

Deep Learning in Image Classification using Dense Networks and Residual Networks for Pathologic Myopia Detection

Zein Rasyid Himami
Department of Mathematics
Universitas Indonesia
Depok, Indonesia
zein.rasyid@sci.ui.ac.id

Alhadi Bustamam
Department of Mathematics
Universitas Indonesia
Depok, Indonesia
alhadi@sci.ui.ac.id

Prasnurzaki Anki
Department of Mathematics
Universitas Indonesia
Depok, Indonesia
prasnurzaki.anki@sci.ui.ac.id

***Abstract*— Half of the population in the world predicted will have myopia and one-tenth of the population will have high myopia. Pathologic myopia is the most dangerous form of myopia that can lead to vision loss permanently. The definition of it was updated as the unusual cases were reported. The latest definition of pathological myopia is eyes with posterior staphyloma or myopic maculopathy equal to or higher than category 2 in META-PM. Detection of pathologic myopia requires a high cost because of insufficient specialists worldwide. To produce an efficient cost, artificial intelligence for health care is rapidly adopted. Several ophthalmology studies have been conducted using retinal fundus photographs such as diabetic retinopathy, cataract, age-related macular disease, and pathologic myopia. Nevertheless, pathologic myopia detection has still been a scarce resource due to the unstandardized definition yet. In this study, a public dataset is used. There are 612 images available distinguished into two classes: normal eye and pathologic myopia eye. The augmentation technique was used to create a robust model. ResNet and DenseNet architecture are performed on two different preprocessing and splitting data. Each model also used three variations of the optimizers such as SGD, RMSprop, and Adam to work out which optimizer performs better and fine-tune the learning rate each time the model stops improving. The results showed that the best model on this proposed method provides accuracy, sensitivity, and specificity of 97%, 93%, and 100%. It performed on DenseNet architecture with normalization and standardization preprocessing, 70:20:10 type of data split, and adam optimizer.**

Keywords*— *convolutional neural networks, image classification, pathologic myopia, retinal fundus.

I. INTRODUCTION (*HEADING 1*)

Myopia is a burden health problem worldwide which increases continuously in decades. In metropolitan areas of East Asia, 80% to 90% of juveniles graduating upper grades have myopia, whereas one-tenth to two-tenths of them have high myopia [1]. If it remains untreated, myopia may lead to pathologic myopia, the most dangerous structure of myopia. In [2], Brien et al. performed a meta-analysis and systematic review of myopia and high myopia. It is estimated that in 2050, half of the world population, approximately 4.8 billion citizens, will have myopia, and one-tenth of those will have high myopia, approximately 938 million citizens. It is a highly prominent issue since mild myopia could even be a risk factor for other ocular disorders [1]. Consequently, governments and health workers keep an eye on the prevalence of myopia [2,3] which has been a significant issue in the world [4].

Pathologic myopia, for a long time, had not been standardly defined. It is often confusing to distinguish between high myopia and pathologic myopia. These two forms are highly diverse. In [5], Ohno-Matsui elaborated the proposed definition from Curtin and Duke-Elder. First, Duke-Elder used degenerative myopia structure to define pathologic myopia, the form of myopia with degenerative changes that mainly occur in the globe's posterior pole. Then, Curtin evidenced that the measurements used in myopic level classification, such as axial length and refractive error in eyes, were doubtful barometers to detect pathologic myopia. He suggested the morphology of the posterior staphyloma instead. Hence, the definition of pathologic myopia is a myopic eye with pathologic lesions in the posterior fundus. In contrast, a high stage of myopic refractive error in the eye belongs to high myopia.

A meta-analysis was developed for pathologic myopia (META-PM) due to reports appearing that pathologic myopia occurs in non-highly myopic eyes [6] and emmetropic eyes [7] so that the definition is updated, that is, eyes with the presence of posterior staphyloma or eyes with myopic maculopathy equal to or higher than category 2 in the META-PM. The eyes with myopic maculopathy are described with structural changes to the retina and choroid, exaggerated elongation of the axial length, and deformation of the sclera [8].

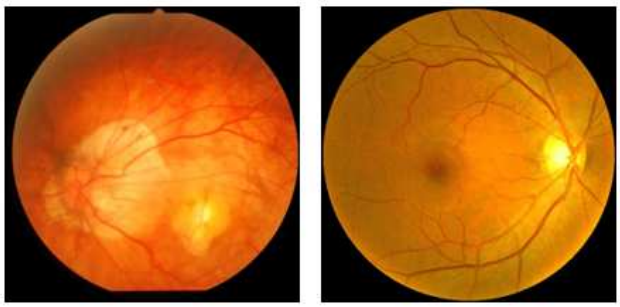


Fig 1. Example of retinal fundus from ODIR

Pathologic myopia is usually progressive and irreversible so that it is identified as one of seven primary diseases that cause blindness in adults in the United States [9]. Early

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

detection accompanied by prompt treatment and active follow-up will prevent patients from suffering immutable vision loss. However, well-skilled pathologic myopia experts are globally still insufficient. On the other hand, myopic maculopathy or pathologic detection is a complicated diagnosis for general eye care doctors such as general ophthalmologists and optometrists. Therefore, automated detection using artificial intelligence will help doctors identify pathologic myopia and produce it more cost-effectively worldwide and locally from retinal fundus photographs. Fundus images are noninvasive and easy to access due to widely primary eye care setting implementation globally. Images of pathologic myopia eye and normal eye are shown in Fig. 1.

Deep learning approaches are sophisticated solutions in image analysis and have widely been applying to preprocess image data. Convolutional Neural Networks (CNN), a part of deep learning, constantly gets enormous attention over the years because of having outstanding performance in image preprocessing. Prior studies in disease detection from retinal fundus images have been conducted, such as diabetic retinopathy [10,11], cataract [12], age-related macular degeneration [13]. However, pathologic myopia detection from retinal fundus images is still a scarce resource. Jiang Liu et al. used PAMELA through the SVM approach and achieved an accuracy of 87.5% [14]. In [9], Devda and Eswari used CNN to classify the disease and achieved highly competitive results. Ran Du et al. [8] performed Efficient-Net to three different datasets and achieved accuracy 92%, 78%, and 88%, respectively. Namra Rauf et al. also achieve an accuracy of detection of 95% [15]. In the future, it is expected that a multidisciplinary collaboration between medical science and mathematics can be established with the help of this program. It enables the authenticity of the research, and the model's performance is tested based on appropriate data in the field [16] [17].

In this study, augmentation and preprocessing were performed to scale up the quantity and the quality of the dataset as well as the robustness of the model. Two varieties of preprocessing were used to find out the most suitable for the models. We used CNN architecture, particularly ResNet-50 and DenseNet-201, with fine-tuning the optimizers such as Root Mean Square Propagation (RMSprop), Stochastic Gradient Descent (SGD) with momentum, and Adaptive Moment Estimation (Adam) and the learning rate each time the model stop improving on two split variants. Then, the models built were evaluated based on the accuracy, specificity, and sensitivity.

II. MATERIALS AND METHOD

In this section, the researchers conducted several phases of the proposed method illustrated in Fig. 2.

A. Acquisition of Image Data

The dataset used in this research is open-access dataset from Ocular Disease Intelligent Recognition (ODIR) collected by Shang gong Medical Technology in China that is available at Kaggle [18]. The dataset consists of 612 color images divided into binary classes, normal and pathologic myopia, each with 306 fundus photographs. These fundus images have an invariable resolution, that is 512×512 pixels.

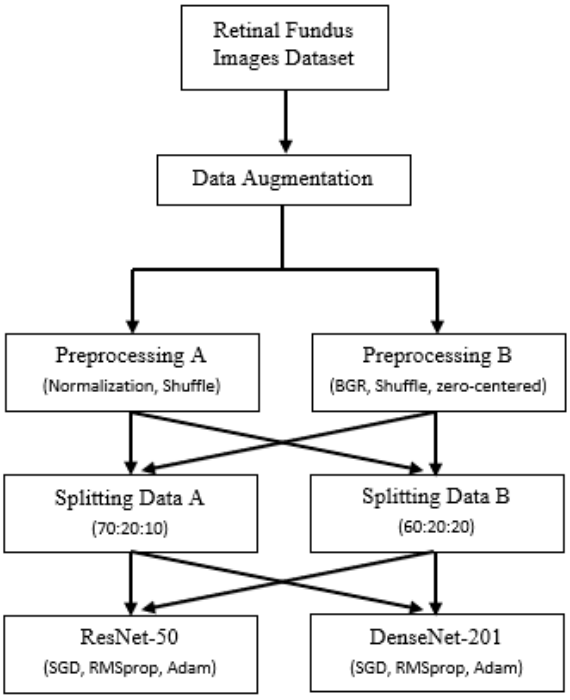


Fig 2. The proposed method

B. Data Preprocessing

Retinal fundus photographs in ODIR have red, green, blue (RGB) channels. We performed two types of preprocessing, that is preprocessing A and preprocessing B. In preprocessing A, the input pixels values of images are scaled between 0 and 1. Then normalization is applied to each color channel concerning the ImageNet dataset as pre-trained weights are set. On the other hand, in preprocessing B, the input images are converted from RGB to BGR. Each color channel is zero-centered concerning the ImageNet dataset as pre-trained weights are set, and scaling is not performed.

C. Data Augmentation

Overfitting is a state of the well-trained model that could not perform well on testing data. To prevent this, the data is augmented to add up the number of fundus images to better identify the model [19]. In this research, the image augmentation technique was randomly zoomed, flipped, rotated, and shifted.

D. Data Split

Splitting the data is required to evaluate the models simply. In this study, we split data into three sections: training, validation, and testing. The training data will be delivered to let the model learn, whereas the validation data is used to check the stability of the model learning. After we achieve the best model, we perform the model to testing data to evaluate it. Two methods used are splitting with a ratio of 70:20:10 and 60:20:20. In the first case, the split data is obtained, that is, 426 training, 124 validations, and 62 testing. In the second one, 364 training, 124 validations, and 124 testing were obtained from 612 data.

E. Feature Extraction and Classification

CNN [20] is mainly used in the computer vision field, which solves image classification cases. This method is generally divided into two main parts, that is feature extraction and fully connected layers. In feature extraction mainly consists of a convolutional layer and pooling layer followed by an activation function. The convolutional layer is a filter of input that helps learn the pattern in terms of image preprocessing. Its operation is:

$$s(t) = (x * w)(t) \quad (1)$$

Where x is an input, w is a kernel, and $s(t)$ is an output, also called a feature map. In image input, x constitutes the pixel values in the shape of the multidimensional array. Padding and stride are parameters primarily used in the convolutional layer. These help the convolution not to lose the pixels that might occur. We add up extra pixels around the boundary of the image. The values usually are set to zeros, which is called padding. Stride is the number of rows and columns traversed per slide when computing the cross-correlation.

The feature map is delivered to the pooling layer to be extracted. There are two popular pooling used, max pooling and average pooling. The most popular non-linear activation function used is Rectified Linear Unit, ReLU. All neurons that have negative values were transformed to zero. The equation is written as follows:

$$f(x) = max(0, x) \quad (2)$$

The feature map or output of feature extraction is usually a multidimensional array. It needs flattening before flowing to a fully connected layer. Two models performed in this study were DenseNet and ResNet architecture because they offer the different forward and backward propagation compared to the prior architecture. For better understanding, the notion of those would be explained as follows.

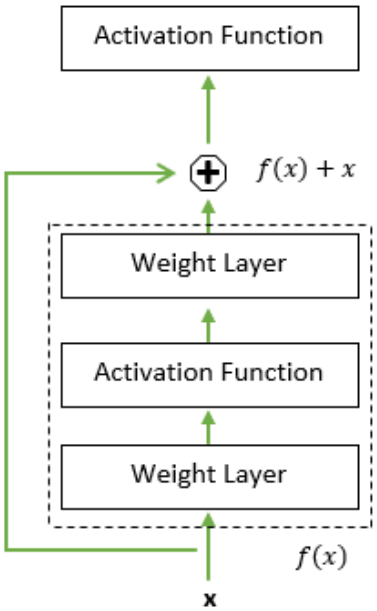


Fig 3. The residual learning in block

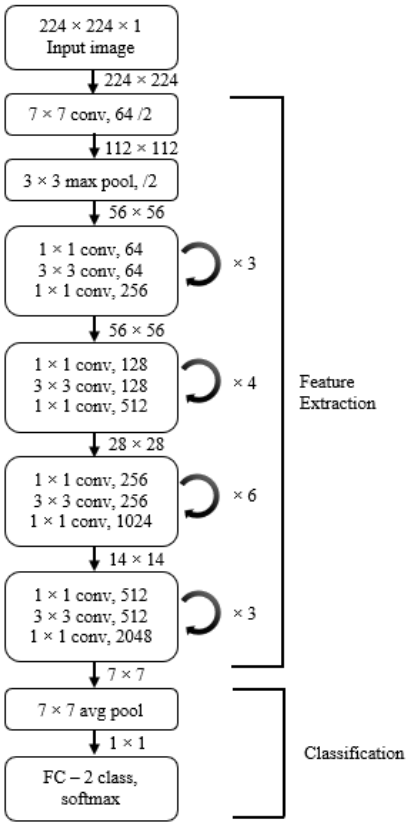


Fig 4. The illustration of ResNet-50 architecture

- ResNet-50

He, Zhang, Ren, and Sun [21] developed CNN architecture called Residual Networks (ResNet). This method, in 2015, won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition. ResNet solved the problem of vanishing gradient. To do so, ResNet adopted residual learning by a shortcut connection in blocks so that the gradient can flow directly through the shortcut backward. On the other hand, it can also accelerate the convergence of deep neural networks forwards. The residual block, see Fig. 3, is formally defined as follows:

$$y = F(x, \{W_i\}) + x \quad (3)$$

Where W_i is the weight of the convolutional layer i, x and y are the input and output of the layers. The function $F(x, \{W_i\})$ represents the residual mapping to be learned. The type we used is ResNet with 50-layers deep (ResNet-50), as shown in Fig. 4.

- DenseNet-201

Huang, Liu, Maaten, and Weinberger [22] developed CNN architecture called Dense Networks (DenseNet). DenseNet adopted a direct connection from any layer to all subsequent layers. As a consequence, the layer will receive cumulative knowledge from all preceding layers. The cumulative knowledge is combined by concatenation, unlike ResNet by summation. The input in the dense block, see Fig. 5, is defined as follows:

$$x_l = H_l[(x_0, x_1, \cdots, x_{l-1})] \quad (4)$$

Where $[x_0, x_1, \cdots, x_{l-1}]$ refers to the concatenation of feature maps produced in layers $0, 1, \cdots and\ l-1$. By this connection, DenseNet requires fewer parameters as the model will not learn redundant feature maps. The type we used is DenseNet with 201-layers deep (DenseNet-201), as shown in Fig. 6.

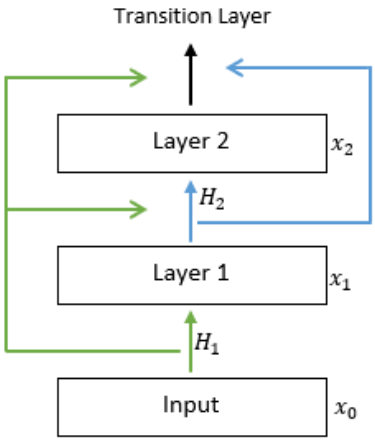


Fig 5. The dense block

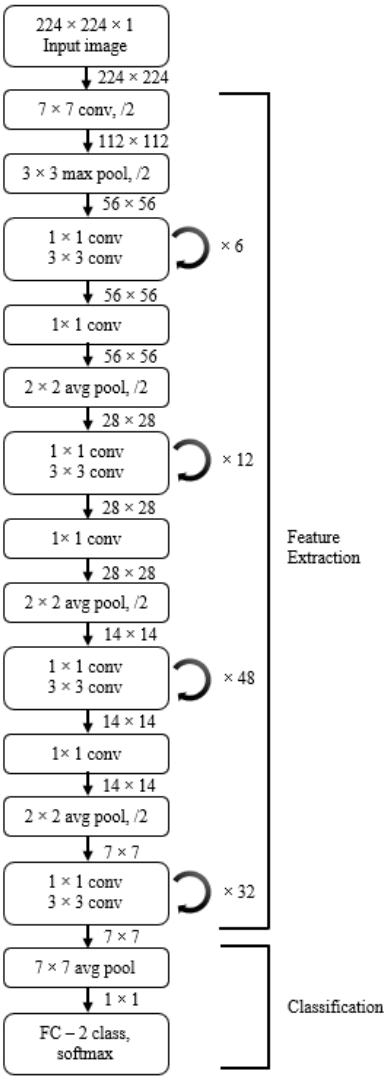


Fig 6. The illustration of DenseNet-201

F. *Optimizers*

One of the hyperparameters tuned in this study was optimizer. Optimized are method or technique used to change the attributes of our model such as weights, learning rate, and sort of things to achieve better result with reducing the lossess. There are three optimizers performed in this study that is SDG, RMSprop, and Adam.

SDG is a well-known optimization technique, yet it has relative high computation time when training data. The momentum parameter is meant to speed up and accelerate learning. Thus, it handles high curvatures, noisy gradients, and small but steady gradients [23]. On the other hand, RMSprop is an improved Root Propagation (Rprop) extensively used to keep a moving average of the square gradient of each weight [24]. The last, Adam is a combination or aggregate of RMSprop and momentum. It has the benefits such as the ability of handling loose gradient in noise problem, less memory and time cost [25]. These three methods were compared to examine its performance in models.

G. *Performance Analysis*

The model's performance can be assessed using metric evaluation. We utilized a confusion matrix in this study. It is a two-dimensional matrix indexed by actual and prediction classes. The measurement of this method includes true positive (TP), false positive (FP), true negative (TN), and false negative (FN), see Table I. TP and TN are a situation in which the trained model accurately predicts the class. On the contrary, FP and FN mean that they inaccurately predict the class.

TABLE I. CONFUSION MATRIX

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	TP	FN
	Negative	FP	TN

In this study, the model will solve the binary classification case. The metrics used are accuracy, sensitivity, and specificity. Accuracy works well on the balanced dataset to calculate the number of correct predictions. The accuracy equation can be written as follows:

$$\text{Accuracy} = \frac{TN+TP}{TN+FN+TP+FP} \quad (5)$$

Sensitivity and specificity are usually used in diagnostic tests. Those mathematically describe the accuracy of a test informing the presence or the absence of a condition. Their formulas can be written as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (7)$$

III. RESULTS

Dataset used from ODIR has 612 color fundus images distinguished into normal class and PM class. Before entering the models, the data augmentation and preprocessing are done. Two compositions were performed to have a ratio of 70:20:10 and a ratio of 60:20:20 for the training, validating, and testing set.

This section shows and addresses the results derived by the proposed methodology, particularly using ResNet and DenseNet architecture. The models were trained on 30 epochs and 16 batch sizes. Here are the charts which display the best model learning on training and validation in each epoch, see Fig. 7. The models ware performed with SDG, RMSprop, and Adam optimizer and the loss function of binary-cross entropy. When a metric has stopped improving, we reduce the learning rate with the lowest point = $1.5e^{-5}$. The most effective performance of optimizers was adam, see Fig. 8. It was slightly better than RMSprop and SDG.

From the composition of the 70:20:10 dataset, the DenseNet model outperformed accuracy, sensitivity, and specificity with scores of 97%, 93.5%, 100%, respectively. The results of the model are summarized in Table II. We compare the outperformed model used in this study to the other prior research in Table III.

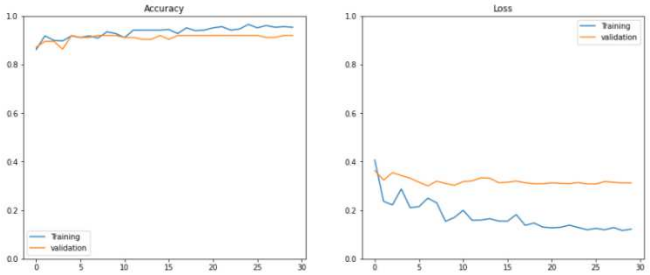


Fig 7. The accuracy and the loss of DenseNet

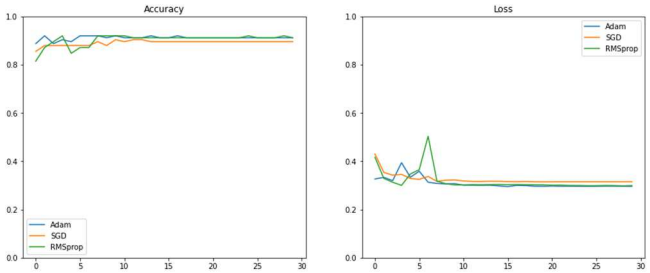


Fig 8. The comparison of each optimizer based on validation accuracy and loss

TABLE III. COMPARATIVE EVALUATION TO PRIOR RESEARCH

Method	Dataset	Accuracy
PAMELA with SVM [17]	SERI	87.5%
Efficient-Net [11]	ACC	92.08%
	PALM	78.06%
	SEED	88.2%
CNN Custom [18]	PALM	95%
The proposed method		
DenseNet-201	ODIR	97%

IV. CONCLUSION

In the eye care setting, an ophthalmologist requires a long time to diagnose pathologic myopia. Hence, reliable technology based on artificial intelligence is necessary to assist ophthalmologists in analyzing disease from retinal fundus photographs. The proposed models in this research, ResNet-50, and DenseNet-201, were trained with two different preprocessing and splitting data. Those also were tuned with three types of optimizers. The accuracy, sensitivity, and specificity obtained for all models concluded that DenseNet-201 with adam gives the best results in the composition data of 70:20:10 and preprocessing with suffle, normalization, and standardization.

In the future, we might reproduce all the models from prior researches and then perform it all to the same dataset to see the fair comparison and the robustness of the models. we also could cooperate with the ophthalmologists to expand the classification of PM such as distinguishing PM classes based on myopic maculopathy from META-PM. Besides, the detection of other diseases from retinal fundus photographs could be conducted using deep learning. Thus, the developers can deploy the model to so-called tools to help doctors prescreen and diagnose diseases.

ACKNOWLEDGEMENTS

This research was supported by the PPI Q2 research grant from the University of Indonesia with contract number NKB-589/UN2.RST/HKP.05.00/2021. The authors deliver a huge appreciation to colleagues from the Directorate of Research and Community Engagement University of Indonesia and Data Science Centre Department at the Faculty of

TABLE II. THE SUMMARY OF ALL MODELS EVALUATION WITH ADAM

Architecture	Preprocessing	Splitting data	Accuracy	Sensitivity	Specificity
ResNet-50	Normalization, Shuffle	70:20:10	84%	67.7%	100%
		60:20:10	85%	77.4%	91.9%
	BGR, Shuffle, zero-centered	70:20:10	95%	93.5%	96.7%
		60:20:10	88%	87.1%	88.7%
DenseNet-201	Normalization, Shuffle	**70:20:10**	**97%**	**93.5%**	**100%**
		60:20:10	91%	90.3%	91.9%
	BGR, Shuffle, zero-centered	70:20:10	87%	80.6%	93.5%
		60:20:10	78%	67.7%	88.7%

Mathematics and Natural Sciences who advanced expertise and insights to cultivate this research in numerous ways.

REFERENCES

[1] I. G. Morgan *et al.*, "The epidemics of myopia : aetiology and prevention sc," *Prog. Retin. Eye Res.*, 2017, doi: 10.1016/j.preteyeres.2017.09.004.

[2] B. A. Holden *et al.*, "Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050," *Ophthalmology*, vol. 123, no. 5, pp. 1036–1042, 2016, doi: 10.1016/j.ophtha.2016.01.006.

[3] A. R. Rudnicka *et al.*, "Global variations and time trends in the prevalence of childhood myopia, a systematic review and quantitative meta-analysis: Implications for aetiology and early prevention," *Br. J. Ophthalmol.*, vol. 100, no. 7, pp. 882–890, 2016, doi: 10.1136/bjophthalmol-2015-307724.

[4] B. Holden, *The impact of myopia impact of increasing and myopia prevalenceof myopia*, no. March 2015. .

[5] K. Ohno-matsui, *Atlas of pathologic myopia*. 2020.

[6] N. K. Wang *et al.*, "Clinical characteristics of posterior staphylomas in myopic eyes with axial length shorter than 26.5 millimeters," *Am. J. Ophthalmol.*, vol. 162, pp. 180-190.e1, 2016, doi: 10.1016/j.ajo.2015.11.016.

[7] K. Ohno-matsui, "Proposed classification of posterior staphylomas based on analyses of eye shape by three-dimensional magnetic resonance imaging and wide-field fundus imaging," *Ophthalmology*, pp. 1–12, 2014, doi: 10.1016/j.ophtha.2014.03.035.

[8] R. Du *et al.*, "Deep learning approach for automated detection of myopic maculopathy and pathologic myopia in fundus images," *Ophthalmol. Retin.*, pp. 1–10, 2021, doi: 10.1016/j.oret.2021.02.006.

[9] J. Devda and R. Eswari, "Pathological myopia image analysis using deep learning," *Procedia Comput. Sci.*, vol. 165, no. 2019, pp. 239–244, 2019, doi: 10.1016/j.procs.2020.01.084.

[10] A. Salma, A. Bustamam, A. Rama, and A. Arus, "Diabetic retinopathy detection and classification using googlenet and attention mechanism through fundus images," vol. 12, no. 14, pp. 590–597, 2021.

[11] R. Amalia, A. Bustamam, A. R. Yudantha, and A. A. Victor, "Diabetic retinopathy detection and captioning based on lesion features using deep learning approach," pp. 1–18, 2021.

[12] E. Sudarsono, A. Bustamam, and P. P. Tampubolon, "An optimized convolutional neural network using diffgrad for cataract image classification," *AIP Conf. Proc.*, vol. 2296, no. November, 2020, doi: 10.1063/5.0030746.

[13] Y. Peng *et al.*, "DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565–575, 2019, doi: 10.1016/j.ophtha.2018.11.015.

[14] J. Liu *et al.*, "Detection of pathological myopia by pamela with texture-based features through an SVM approach," vol. 1, no. 1, pp. 1–11, 2010.

[15] N. Rauf, S. O. Gilani, and A. Waris, "Automatic detection of pathological myopia using machine learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-95205-1.

[16] P. Anki and A. Bustamam, "Measuring the accuracy of LSTM and BiLSTM models in the application of artificial intelligence by applying chatbot programme," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 197–205, 2021, doi: 10.11591/ijeecs.v23.i1.pp197-205.

[17] P. Anki, A. Bustamam, and R. A. Buyung, "Looking for the link between the causes of the COVID-19 disease using the multi-model application," *Commun. Math. Biol. Neurosci.*, vol. 2021, pp. 1–17, 2021, doi: 10.28919/cmbn/6128.

[18] "ocular-disease-recognition-odir5k @ www.kaggle.com." [Online]. Available: https://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k.

[19] I. Goodfelow, Y. Bengio, and A. Courville, "Deep learning," pp. 1–3, 2016, [Online]. Available: http://www.deeplearningbook.org/front_matter.pdf.

[20] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *arXiv Prepr. arXiv2106.11342*, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *\Proc. -30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017,* vol. 2017-January, pp. 2261-2269, 2017, doi: 10.1109/CVPR.2017.243.

[23] S. Ruder, "An overview of gradient descent optimization algorithms," pp. 1–14, 2016, [Online]. Available: http://arxiv.org/abs/1609.04747.

[24] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural networks Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

Transfer Learning-based Mobile-focused Automated COVID-19 Detection from Chest X-ray

M. Aldiki Febriantono
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
m.aldiki@binus.ac.id

Ridho Herasmara
Electrical Engineering, Faculty of Science and Technology
Raden Rahmat Islamic University
Malang, Indonesia
ridho.herasmara@uniramalang.ac.id

Anita Rahayu
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
anita.rahayu@binus.edu

***Abstract*—COVID-19 Pandemic is still a global issue that threatens global health. To combat the pandemic, testing activities has been the first line of defense. However, increasing number of infections resulted in insufficient number of laboratory kits to perform the test. One potential testing method is using transfer learning for automated detection of COVID-19 from chest x-ray image. We create a model used pretrained model of MobileNetV3Large as a feature extractor, and a custom classification layer. We train the model on dataset consisting of chest x-ray image from 10,192 healthy cases, 3,616 COVID-19 cases, 1,345 Viral Pneumonia cases, and 6,012 Lung Opacity cases. The model achieved macro-average accuracy performance of 89.08%, F1 score of 88.10%, Precision of 91.95%, Sensitivity of 85.51%, and Specificity of 95.26%. Comparison with previous models trained on smaller dataset showed that achieved performance is lower and indicates previous research's model won't be able to maintain its performance when evaluated on larger sets of data.**

Keywords—covid-19, chest x-ray, transfer learning, mobilenetv3,

I. INTRODUCTION

COVID-19 Pandemic is still a global issue which threatens global health. Since the beginning of the pandemic, testing and tracing has been the first line of defense against the spread of infection [1]. However, exponential increase of number of infection cases has resulted in insufficient number of laboratory kits available, creating a further challenge to perform test and trace [2]. This challenge provides opportunity for alternative testing methods to fill this gap.

To meet the challenge of limited laboratory kits available, several experimental testing methods has been proposed. The novel methods proposed for rapid testing, including detection of volatile organic compound (VOC) found in breath [3], as well as the gargling method [4]. Some studies proposed the use of automated tools to aid in the diagnostic, specifically by radiological examination and convolutional neural networks [5-8]. The use of radiological examination is attractive since it didn't tap into resources required to perform current test, hence it carries potential to increase testing capacity.

Computed tomography (CT) scan is the first consideration for radiography examination. It is more effective in detection of COVID-19 in patients. However, increasing number of patients, and subsequent rise in examinations performed, already put additional strain on radiology department. Therefore, even though it is less sensitive in detecting early-stage pulmonary involvement, it is considered that Chest X-Ray is a more suitable option [9].

CXRs itself is considered less sensitive in diagnosing early-stage disease. However, rapid spread of COVID-19 globally makes its use increasingly considered. Research had begun the use of deep learning technique to detect COVID-19 from patient CXR image. Current challenge facing this technique is limited availability of the dataset. This limited availability forced some researchers to create their model using a previously trained networks [10-18]. This transfer learning technique attempts to transfer domain knowledge of pretrained network, in attempt to solve problem in related domain, which in this case in an image classification problem. Although these approaches are using state-of-the-art networks, they are still trained on limited dataset, and their results are still difficult to generalize and maintained when faced on larger datasets. Therefore, approaches involving transfer learning needs to be verified on a larger dataset. The dataset itself needs to consist of data from healthy, and pneumonia cases, to prevent disregard of intraclass variation [19]. Furthermore, another challenge also remains to develop algorithm capable of identifying patient with COVID-19 apart from viral pneumonia cases. This problem is emphasized, since the COVID-19 itself share some common characteristics with viral pneumonia [5].

Aside from challenge from data availability, some of the requirement for the new method is for it to be able to be deployed on limited capability hardware, such as mobile devices. Some techniques such as MorphNet [25] offers technique to learn possible smaller structure from existing networks to make it compatible with limited-capability hardware. Another approach provides pretrained networks that was designed to conform to such constrain from the beginning. One of such networks is MobileNet [20]. Its latest iteration was further developed through technique of Network Architecture Search (NAS) which yield MobileNetV3 which is higher in accuracy, but comparable in latency to the original MobileNet [21].

Therefore, this research attempted to solve those problems using available potential solution. This research used a transfer-learning based approach to create a mobile-focused neural network classifier for automated detection of COVID-19 from chest X-ray images, using a broader set of training data that includes normal, COVID-19, lung opacity, and viral pneumonia cases. This research specifically used the MobileNetV3Large [21] as the pretrained base model.

II. METHODS

A. Dataset

In this research dataset was used to train and validate the neural network models. Dataset that was used contains images

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

of chest X-ray images that was made publicly available from previous data collection effort [22][23]. This dataset is significantly larger from datasets used on previous research [10-18]. This dataset also consists of more classes of images, consisting of COVID-19, Lung Opacity, Normal, and Viral Pneumonia cases, which is not available on previous dataset. The dataset contains images in portable network graphic (PNG) format, with a resolution of 256-by-256 pixels over 3 color channels. The details for number of data for each case in the datasets is shown in Table I.

TABLE I. DATASET IMAGES COMPOSITION

No	Case	No. of Images
1	Normal	10192
2	COVID-19	3616
3	Viral Pneumonia	1345
4	Lung Opacity	6012

B. *Preprocessing and Augmentation*

Before the dataset is used for training, several techniques were implemented to perform data preprocessing and augmentation. Previous research has shown that implementing a preprocessing and augmentation can improve training time, generalization ability, as well as model performance, especially in medical application [24].

For the preprocessing, image rescaling is performed. Original image size of 256-by-256 pixels over 3 channels resized into original image size used for input layer of the pretrained base, MobileNetV3Large, of 224-by-224 pixels over 3 channels. As the pretrained base model contains a rescaling layer [21], there were no rescaling nor normalization performed before inputting the data to the model.

After preprocessing done, the data is then augmented. Augmentation of the data included applying transformation to the image data that were used to train the network. Transformations performed were rotating the image data within 90 degrees range, image zoom range of 0.2, shearing range of 0.2, and feature wise centering and standard deviation-based normalization. Fig. 1 shows the sample image after preprocessing and augmentation.

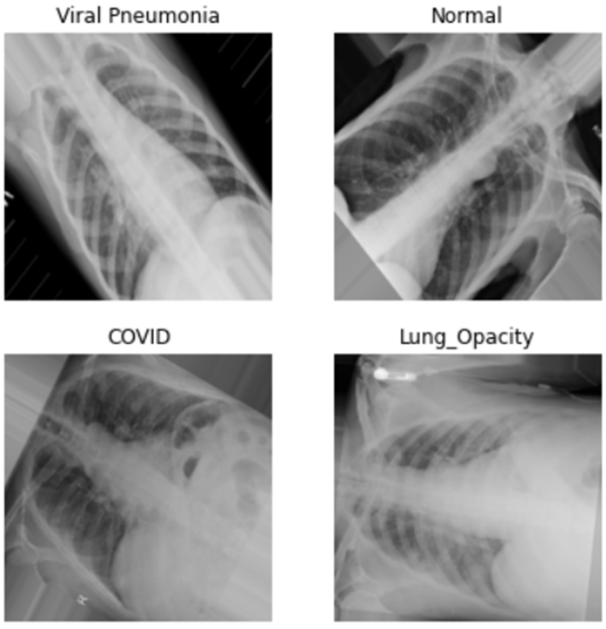


Fig. 1. Sample of Preprocessed and Augmented Data

C. *Transfer Learning*

Transfer learning focused on storing previously learned knowledge from one domain and attempt to apply it to different, but related problem [26]. This research attempted transfers learning to overcome problem related with image classification for automated detection of COVID-19 from chest X-ray. We took a base neural network models that has been trained on image classification problem, to be used as feature extractor part of the new model. Then, we connect output of this feature extractor base part to our custom classification head, to create a new neural network model for this problem.

The feature extractor used is MobileNetV3Large, which has been previously trained on collection of image dataset named *ImageNet* [21]. This model takes input image with dimension of 224x224x3, and outputs a 1-dimensional vector data with the size of 1280. This model consists of 4,226,432 parameters. Parameters from the feature extractor is set to frozen and will not be updated during the training process.

For the structure of the classification head, we created three layers. First, a fully connected layer consisted of 256 fully connected units is created. Second, a dropout layer is created and connected to the first layer output, to prevent overfitting problem [27]. Third, a classification layer consisting of 4 units (one for each possible class) is created and connected to the output of the dropout layer. This classification layer is then connected to the output from the feature extractor part. And finally, to determine the predicted class, output from the classification layer is passed through a *softmax* activation function. This model consists of 4,555,396 parameters in total. Subtracting previously frozen layer from the feature extractor, total trainable parameters are 328,964 parameters. Fig. 2 shows the structure of the model.

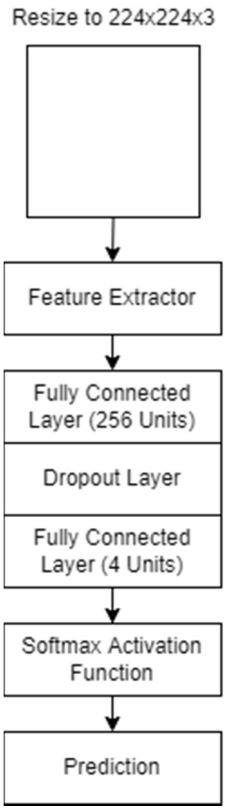


Fig. 2. Outline of the neural network structure

Steps taken by each image is as follows. First, an image data of Chest X-ray is resized into size of 224x224x3. This image is then being fed into the feature extractor, which is based on the MobileNetV3L neural network model. This model outputs an array of 1280 in size. This array is then fed into the classification layer, which outputs the prediction by the network. The prediction is in an array of 4 in size, with highest value being the predicted case, one from the 4 cases available in the dataset.

D. Training and Validation Setup

Training and validation process involved splitting training data into two sets. First set of the data was used to train the network, while the second set of the data was be used for validation process. The split was performed at 80:20 ratio, with 80% of the data being used for training process, while 20% of data be used for the validation process. The split yield 16933 image data for training, and 4232 for validation purpose. Preprocessing was applied for both data used for training and validation, but augmentation techniques were performed only on the training data. As the dataset contains four distinct class of images, labeling process was performed by creating a categorical enumeration using a one-hot scheme.

For each training steps, batch size was set to 32. With over 16933 data for training, this batch size resulted in 530 steps of training for each training epoch. Training epoch set to 600 epochs. Optimizers used is Adam [28], with learning parameter of learning rate 0.001, beta 1 of 0.9, beta 2 of 0.999, and epsilon of 1.10^{-7}. To perform the loss calculation, a categorical cross-entropy loss function is used.

Training and validation were performed on workstation with 4 cores intel CPU, 16 GB RAM, and a 640 CUDA cores GPU with 4GB vram. Programming is performed on python with TensorFlow library, through jupyter notebook kernel. For efficiency, training is performed with tensorflow-gpu module. The training worker parameter is set to 8.

E. Performance Metrics

To measure performance, model will be evaluated using validation set, computing for macro-average of Accuracy, F1 score, Precision, Specificity, and Sensitivity. Accuracy is ratio of correct prediction to total prediction made. F1 score measure balance between precision and recall. Precision measure ratio of true positive prediction to total positive predictions made. Sensitivity measure ratio of true positive to true positive and false negative. Specificity measure ratio of true negative to true negative and false positive.

III. Results and Discussion

A. Training Results

Training has been performed on the neural network classifier, and each training epoch lasted between 104 s and 130 s, for a total training time over 600 epochs of 17 hours 43 minutes and 27 second. Loss value and accuracy value is recorded for each epoch of the training both for the training data and on validation data. Fig. 3 shows progression of model loss value during validation phase of the training. Fig 4. Shows progression of accuracy value during validation phase of the training. This training approach is stochastic by nature, hence trainings, even with the same training parameters, might yield slightly different result.

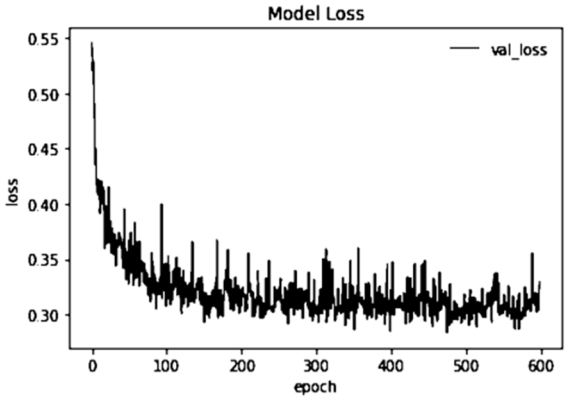


Fig. 3. Model loss value during validation phase of the training

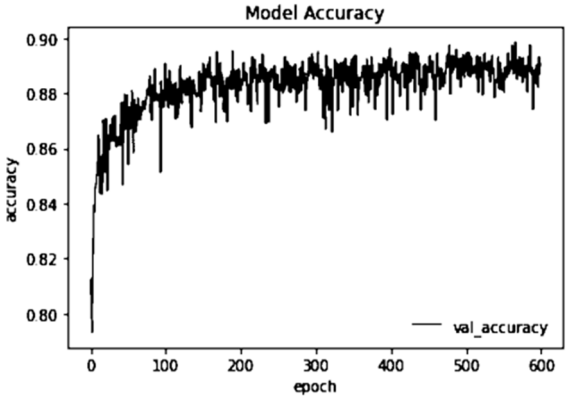


Fig. 4. Accuracy progress during validation phase of training

B. Model Performance

Model performances were evaluated using macro-average computation of accuracy, F1 score, Precision, Specificity, and Sensitivity. As loss value from loss function calculation is not an actual performance metrics, loss value has been excluded from performance metrics. Using the actual data provided by the dataset, and prediction by the neural network classifier, we created a confusion matrix to show general performance of the neural network classifier model. Confusion matrix from the data prepared for validation showing classification performance is shown in Fig. 4.

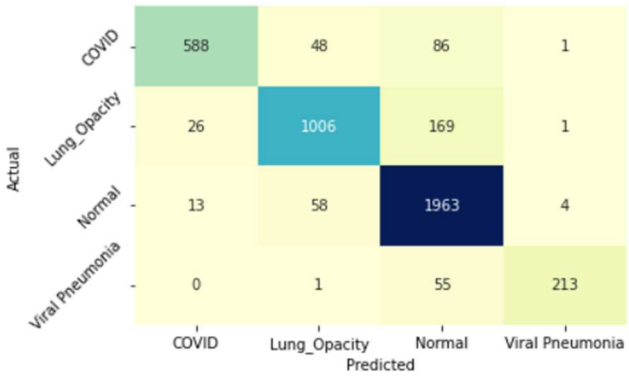


Fig. 5. Confusion matrix of model based on MobileNetV3Large

From the confusion matrix presented, it was shown that this neural network classifier model achieved a specificity rate of 85.87% for normal case. This rate is low considering that a low specificity for normal class means that a high number of cases were rather predicted as normal when it is not normal.

This false positive of normal prediction on actual case of COVID-19 might result in patient being discharged, infecting other people. Overall performance of this neural network classifier model is shown in Table II.

TABLE II. MODEL PERFORMANCE METRICS (%)

No	Metrics	Value
1.	Accuracy	89.08
2.	F1 score	88.10
3.	Precision	91.95
4.	Specificity	95.26
5.	Sensitivity	85.51

Performance metrics shows that although the accuracy performance of 89.08%, and macro-average specificity value of 95.26%, it was examined that its normal class specificity value is much lower at 85.87%. Further training using a larger set of data and fine tuning of the model is required before the model can be considered for deployment.

C. Comparisons With Other Models

Previous efforts [5][11] used a much more limited dataset to train their network models. These datasets were created in the beginning of the pandemic, and contains data of 1428 and 3886 sample sizes. These more limited datasets are order of magnitude smaller than currently used dataset. Therefore, comparison with previous effort has been made with emphasis on difference of training data amount. In comparing with those research, previous research did not provide all performance metrics which were used in this research. For those figures which were not included from previous research, we mark with 'n/a' (not available). Comparisons between transfer learning classifier model form this research and previous research, is elaborated in Table III.

TABLE III. COMPARISON OF DIFFERENT MODEL PERFORMANCE

Classifier	Data size	Acc. (%)	F1 (%)	Prec. (%)	Spec. (%)	Sens. (%)
MobileNetV3L	21165	89.08	88.10	91.95	95.26	85.51
MobileNetV2	1428	94.72	n/a	n/a	98.66	96.46
MobileNet	3886	98.32	97.01	95.29	98.25	98.78
VGG19	1428	93.48	n/a	n/a	92.85	98.75
VGG16	3886	98.71	97.59	96.43	98.67	98.78
InceptionV3	3886	98.39	96.97	96.39	98.69	97.56
Xception	3886	97.43	95.24	93.02	97.38	97.56

From the comparison, it is noticeable that the model based on MobileNetV3L have lower performance even when compared to its predecessor of MobileNet and MobileNetV2. However, we pointed out that MobileNetV3 is a successor model of MobileNetV2, which has been demonstrated for having better performance [21]. This indicates that those previous models being trained on smaller dataset might not generalize well, nor maintain its initial performance if trained and tested on the same dataset used on this research.

CONCLUSIONS

In this paper, we trained a model consisted of MobileNetV3L as feature extractor, and a custom layer for classification of COVID-19. Training was performed on dataset consisted of 21,165 images across four classes. Performance metrics of the model in terms of macro-average accuracy, F1 score, Precision, Sensitivity, and Specificity were 89.08%, 88.10%, 91.95%, 85.51%, and 95.26%. This result indicates that the model is capable in detecting correct disease from a Chest X-ray image data, 89.09% of the time, with 91.95% of the time it detect a disease class, turns out to be true. It is able to detect 85.51% of a disease case from total case. It is also able to detect in 95.26% of the case that people that didn't have a disease, actually didn't have that disease.

Comparisons with previous study indicated that due to training on limited data, previous research model performance metrics scored higher. However, those models weren't likely to hold its performance and generalization ability when faced with the new dataset with higher amount of image data available.

The model resulted from this study itself still has its shortcomings. Although trained on a large dataset, the current model is seen as still having potential to be improved. Further fine tuning, network design revision, and/or usage of different classifier should be considered to achieve a model with higher performance. Several techniques such as cost-sensitive training can be applied to tradeoff one performance metrics with the other, such as to increase specificity values at the cost of other performance metrics.

REFERENCES

[1] M. Chung, et al., "CT imaging features of 2019 novel coronavirus (2019-nCoV)," Radiology, vol. 295, no. 1, 2020, pp. 202–207.

[2] H. Shi, et al., "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study," The Lancet Infect. Dis., vol. 20, no. 4, pp. 425–434, 2020.

[3] H. Chen, et al., "Breath-borne VOC Biomarkers or COVID-19," medRxiv doi: 10.1101/2020.06.21.20136523, 2020.

[4] S. Bennett, R. S. Davidson, and R. N. Gunson, "Comparison of gargle samples and throat swab samples for the detection of respiratory pathogens," Jour. of Vir. Meth., vol. 248, 2017.

[5] M. M. Taresh, et al., "Transfer Learning to Detect COVID-19 Automatically from X-Ray Images Using Convolutional Neural Networks," Int. Jour. of Biomed. Imag., vol. 2021.

[6] N. K. Mishra, P. Singh, and S. D. Joshi, "Automated detection of COVID-19 from CT scan using convolutional neural network," Biocyb. and Biomed. Eng., vol. 41, pp. 572-588, 2021.

[7] M. Ghaderadeh, et al., "Deep Convolutional Neural Network–Based Computer-Aided Detection System for COVID-19 Using Multiple Lung Scans: Design and Implementation Study," Jour. of Med. Int. Res., vol. 23, no. 4, 2021.

[8] D. M. Hansell, A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Müller, and J. Remy, "Fleischner Society: glossary of terms for thoracic imaging," Radiology, vol. 246, no. 3, pp. 697–722, 2008.

[9] P. Sun, X. Lu, C. Xu, W. Sun, and B. Pan, "Understanding of COVID-19 based on current evidence," Jour. of Med. Vir., vol. 92, no. 6, pp. 548–551, 2020.

[10] A. Naris, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using X-ray images and deep convolutional neural networks," arXiv:2003.10849 [eess.IV], Mar. 2020.

[11] I. D. Apostolopoulos, and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," Phys. and Eng. Sci. in Med., vol. 43, no. 2, pp. 635-640, 2020.

[12] E. E. D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: a framework of deep learning classifier to diagnose covid-19 in x-ray images," arXiv:2003.11055 [eess.IV], Mar. 2020.

[13] P. K. Sethy, and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," unpublished.

[14] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with x-ray images," Comp. in Bio. and Med, vol. 121, 2021.

[15] S. Albali, "A deep neural network to distinguish COVID-19 from other chest diseases using X-ray images," Curr. Med. Img. Rev., vol. 17, pp. 109-119, 2021.

[16] D. Dansana, R. Kumar, and A. Bhattacharieetal, "Early diagnosis of COVID-19 affected patients based on X-ray and computed tomography images using deep learning algorithm," Soft Comp., pp. 1-9, 2020.

[17] J. Civit-Masot, D. Luna-Pereion, M. D. Morales, and A. Civit, “Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images,” App. Sci., vol. 10, no. 13, pp. 40-46, 2020.

[18] M. Farooq, and A. Hafeez, “Covid-resnet: a deep learning framework for screening of covid19 from radiographs,” arXiv:2003.14395 [eess.IV]

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: visual explanations from deep networks via gradient-based localization,” 2017 IEEE Int. Conf. on Comp. Vis. (ICCV), 2017.

[20] A. Howard, et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Visions Applications,” arXiv:1704.04861 [cs.CV], Apr. 2017.

[21] A. Howard, et al., “Searching for MobileNetV3”, arXiv:1905.02244 [cs.CV], May 2019.

[22] M.E.H. Chowdhury, et al., “Can AI help in screening Viral and COVID-19 pneumonia?” IEEE Access, Vol. 8, pp. 132665 – 132676, 2020.

[23] T. Rahman, et al., “Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images,” Comp. in Bio. and Med., vol. 132, May 2021.

[24] A. Mikolajczyk, and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” Proc. Of. Intl. Interdiscp. PhD Workshop, 2018.

[25] A. Gordon, et al., “MorphNet: fast and simple resource-constrained structure learning of deep networks,” Proc. of IEEE Conf. of CVPR, pp. 1586-1595, 2018.

[26] S. Bozinovski, “Reminder of the first paper on transfer learning for neural networks, 1976”, Informatica, vol. 44, pp. 291-302, 2020.

[27] D. M. Hawkins, “The problem of overfitting,” Jour. of Chem. Inf, and Comp. Sci., vol. 44, no. 1, pp. 1-12, 2004.

[28] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” arxiv.org/abs/1412.6980

The Impact of Large-Scale Social Restriction and Odd-Even Policies During COVID-19 Pandemic to Traffic Congestion and Air Pollution in Jakarta

Muhamad Fahriza Novriansyah
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
mfnovriansyah@gmail.com

Yudhistira Nugraha
Jakarta Smart City and School of Computing,
Telkom University
Jakarta - Bandung, Indonesia
yudhistira.nugraha@jakarta.go.id

Hansen Wiguna
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
hansenwgn@gmail.com

Andy Ernesto
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
andyernesto43@gmail.com

Andi Sulasikin
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
andisulasikin.jsc@gmail.com

Bahrul Ilmi Nasution
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
bahrulnst@gmail.com

Juan Intan Kanggrawan
Jakarta Smart City
Department of Communication, Informatics and Statistics
Jakarta, Indonesia
juan.tan.kang@gmail.com

Alex L. Suherman
Directorate of Research and Community Service Telkom University
Telkom University
Jakarta, Indonesia
alexsuherman@telkomuniversity.ac.id

***Abstract*—The COVID-19 pandemic has had a global impact on transportation mobility and air pollution, including Jakarta as the capital and busiest city in Indonesia. This paper reports the impact of two policies imposed by the Governor of Jakarta, namely the odd-even and the large-scale social restriction (PSBB) transitional phase-1, against the traffic congestion and air pollution quality in Jakarta during the COVID-19 pandemic. This paper investigates the odd-even and PSBB policy impact using paired T-Test. Moreover, this study assesses the relationship between traffic congestion and air pollution using the Pearson correlation. The result shows that the odd-even policy does affect significant only on MH Thamrin Street. Furthermore, the odd-even policy does not significantly affect air pollution reduction in Jakarta. This study also found that there is no meaningful relationship between traffic congestion and air pollution. These results can be used to reference future data-driven policy improvement on traffic congestion and air pollution management in Jakarta and other cities.**

Keywords— odd-even policy, large-scale social restriction, traffic congestion, air quality, covid-19, Jakarta

I. INTRODUCTION

The COVID-19 pandemic has influenced transportation mobility and air pollution in many countries, including Indonesia. This pandemic has affected the air quality in Jakarta, the largest city and capital city of Indonesia [1]. Lockdown or social restriction policy in many countries in Asia, Europe, South America, and North America led to significant air quality improvements [2], [3]. Traffic congestion is a severe problem in Jakarta, which has often occurred before the pandemic. Jakarta occupied the top ten congested cities globally, consecutively from 2017 to 2019 [4]. Jakarta has been deemed the second largest world pollution contributor, reported on June 26th, 2020 [5].

Traffic congestion and air pollution in Jakarta are significant problems that must be addressed in sustainable ways. Several cities in the world have implemented odd-even policies, the limitation of the motor vehicle - referring to the last two numbers on the license plate number of the vehicle, to reduce vehicle mobility and air pollution [6]. They are Rome, Paris, Mexico, New Delhi [7], and Beijing [8], [9]. In Beijing, during the 2008 Summer Olympic Games, the Odd-Even policy successfully reduced traffic congestion and increased travel speed [8]. Odd-Even's definition in the Jakarta context is a policy that limits the most rear police plate or number based on odd and even numbers [10]. The Provincial Government of Jakarta has implemented this policy on several roads since August 30th, 2016, to replace the ineffective *3-in-1 system* [11].

Literature studies have reported the effect of the Odd-Even policy [12]–[14]. The Odd-Even policy implementation shows a change in vehicles' volume on roads in Jakarta, with people tending to choose public transportation as an alternative option [12]. The Odd-Even policy application also increases the vehicle's average speed on several roads in Jakarta [13]. During the COVID-19 pandemic, the Provincial Government of Jakarta implemented an Odd-Even policy in conjunction with the PSBB Transitional Phase 1 [15]. This policy runs from August 3rd to September 13th, 2020 [16]. The Odd-Even policy is enforced on 25 roads with application times from 6-10 am and 4-9 pm at Jakarta time on weekdays.

This paper analyses the effectiveness of the odd-even policy during the COVID-19 pandemic on MH Thamrin, Neli Anggrek, and Merdeka Perintis streets. The results are obtained by conducting data exploration, testing them using the Paired-Test [17], [18], and analyzing the correlation between traffic congestion and air quality using the Pearson test [19].

The rest of this paper is organized as follows. Section 2 describes the methodology used to analyze the data. The following section shows the analysis and findings of the study.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

In Section 4, we discuss the odd-even effect on traffic congestion and air pollution and their correlation. Finally, we present the conclusion, recommendation, and future works.

II. Related Work

A. PSBB Transitional Phase Policy and Odd-Even Policy to Traffic Congestion

A recent study from Nissa et al. [20] about PSBB Transitional Phase Policy revealed an improved average vehicle speed in Jakarta due to the policy. The study compared the average vehicle speed before PSBB and during PSBB [20]. During PSBB Transitional Phase 1 Policy, the Provincial Government of Jakarta implemented the odd-even policy that could influence the average vehicle speed.

In the previous research from Supriana et al. [13], before the PSSB Policy was implemented, the Odd-Even policy increased the vehicle's average speed on several streets in Jakarta. The study was conducted using descriptive statistical analysis for average speed [13]. Other studies are about the impact of the odd-even policy on the vehicle's average speed in Beijing, China [8], [9]. The study from Ren et al. used a Nonparametric Wilcox test that shows that the odd-even policy can reduce traffic congestion [9].

B. PSBB Transitional Phase Policy and Odd-Even Policy to Air Quality

Recent studies from Pardamean et al. [21] about PSBB Transitional Phase Policy enhance the air quality of Jakarta. The study conducted by the statistical test used a chi-squared test to the AQI every location and PSBB Transitional Phase. It also used paired one-sided *t*-test for the seasonal trend [21]. Other related studies about the effect of a similar policy, which is a lockdown, on the air quality in several countries in the world [2]. The study from Ghosh and Ghosh [2] compares the mean of the air quality index between pre and post-lockdown period with the parameters PM_{10}, $PM_{2.5}$, CO, NO, NO_2, NH_3, NO_x, SO_2, and O_3.. This study concluded that the lockdown significantly improved the air quality of the world [2].

The study from Ren et al. [9] about the odd-even policy effect on air quality in Beijing, China. The study was conducted by a Nonparametric Wilcox test [9]. This study shows that the effect of the odd-even policy is little or limited to the air quality [9]. This study also shows that air quality is not very relevant to traffic congestion [9]. Other related studies in India from Goyan and Gabdhi [22] about air quality when the odd-even policy was implemented. The objective of this policy implementation is to reduce the air pollution problem [22]. This study uses the average Air Quality Index and meteorological conditions each day [22]. The result shows that the pollution levels were high when the odd-even policy was implemented [22]. This result can happen because of meteorological conditions like 'low daily temperatures' and 'low wind speeds' [22].

C. Research Group

Therefore, this study concludes that there has been no report or analysis of PSBB Policy and the odd-even policy effectiveness on traffic congestion and air pollution simultaneously in Jakarta during the COVID-19 pandemic.

III. Methodology

In this study, the analysis was performed using Data Exploration, Hypothesis Testing with Paired T-Test, and Correlation Testing with Pearson Correlation.

A. Datasets

Dataset was obtained from the Jakarta Environment Office for air quality and the *Waze Application* for traffic information. The Waze application is a navigation tool created for the local driver community to share traffic jams from the average speed of vehicles and traffic alerts from accident reports [23].

The data includes daily air quality data from three streets in Jakarta, namely Bundaran HI (DKI1), Kelapa Gading (DKI2), and Kebun Jeruk (DKI5). The air quality dataset can be seen in Table I. As shown in Table II, the transportation data includes traffic congestion data in daily average speeds on MH Thamrin, Anggrek Nelimurni, and Perintis Kemerdekaan streets. Dataset was collected during the PSBB Transitional Phase 1, which is June 5th to September 10th, 2020, on weekdays and when the Odd-Even policy was applied or not. On August 3rd, the Odd-Even policy was used until the end of PSBB Transitional Phase 1 on September 10th, 2020.

TABLE I. Air quality data in several streets in Jakarta

Date	Location	ISPU Max Value
05/03/20	Bundaran HI(DK1)	74
05/06/20	Kelapa Gading(DKI2)	91
…	…	…
10/09/20	Kebun Jeruk(DKI5)	97

TABLE II. Traffic congestion in several streets in Jakarta

Date	Street	Avg Speed
05/03/20	Jl. MH Thamrin	5.72
05/06/20	Jl. Perintis Kemerdekaan	6.72
…	…	…
10/09/20	Jl Anggrek Nelimurni	3.43

B. Data Collection and Analysis

The traffic congestion data contains the date, road, average speed, average congestion, and amount of traffic congestion. Air pollution data has start date, end date, *Indeks Standard Pencemar Udara* (hereafter referred to as the ISPU) value parameter ($PM_{2.5}$. PM_{10}, CO, SO_2, NO_2, and O_3), maximum ISPU value, location, and dominant ISPU parameter. These datasets were analyzed in several stages, as shown in Figure 1. First, in the pre-processed stage, the data were selected for specific variables. Next, data cleaning and data integration were performed. Data cleaning is part of the pre-processing stage to clean the data according to our particular study. Meanwhile, data integration is part of combining data after the data is cleaned. There are two focuses on data: the maximum ISPU value and average speed of vehicles per day.

The dataset was then filtered when the PSBB Transition Phase 1 policy came into effect due to an odd-even policy change. Subsequently, the dataset was filtered according to the time and day the odd-even policy was applied. The filter is

dedicated to the location of the road we want to analyze. The locations of the streets analyzed are MH Thamrin, Neil Anggrek, and Merdeka Perintis streets. These three roads were chosen because they are the closest to the ISPU sensors and implemented odd-even policies. Following this stage, data exploration, T-test, and correlation test were carried out.

In hypothesis testing, the null hypothesis [17], *H0*, is assumed that the Odd-Even policy has no significant effect on traffic congestion and air pollution. Meanwhile, the alternative hypothesis [17], *H1*, is considered that the odd-even policy significantly impacts traffic congestion and air quality. Subsequently, the T-test was used to test for significant differences by comparing the two means [17], [18]. We chose the *Paired T-Test* due to the normally distributed data, sample size <30, and the conditions of observation (pre-post) [18]. Paired T-Test was chosen by testing the same sample under one different condition [24], namely when the odd-even policy was applied and not. With the *Paired T-Test*, we can determine the Odd-Even policy's significant effect in those areas. Following this step, we conducted a correlation test between two variables: the vehicle's average speed (described traffic congestion) and the maximum value of the ISPU (defined air quality).

A correlation test was performed as a statistical method to determine a linear relationship between two continuous variables [25]. The correlation coefficient values are in the range of -1 to 1 [25]. If the coefficient value is "0", then there is no relationship. We use the *Pearson Correlation* to assess how strong the relationship is between these two continuous variables (numeric value).

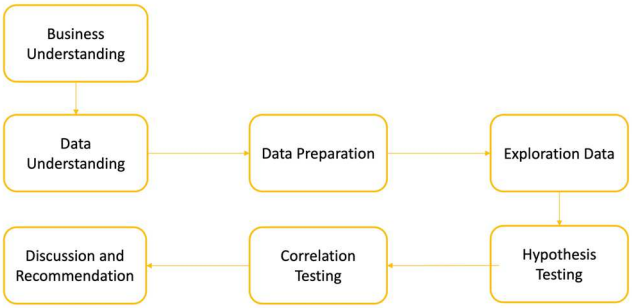


Fig. 1. Research Workflow

IV. Result

A. Data Exploration

We visualize and analyze the Odd-Even policy's application on traffic congestion and air pollution on MH Thamrin, Neil Anggrek, and Merdeka Perintis streets in data exploration. As shown in Figure 2, the lowest average speed occurs when the Odd-Even policy is not applied. Meanwhile, the fastest average speed only occurs when the odd-even policy is applied on MH Thamrin Street. It shows that the odd-even policy is only effective on MH Thamrin Street.

Figure 3 indicates that the three roads shown have the same pattern for air quality values, even when there are differences in sensor locations. The highest and the lowest air pollution value occurs when the Odd-Even policy is not enforced. The value of air pollution tends to decrease when the Odd-Even policy is started. However, if we look thoroughly, the Odd-Even policy's impact is not significant to air pollution on the road.

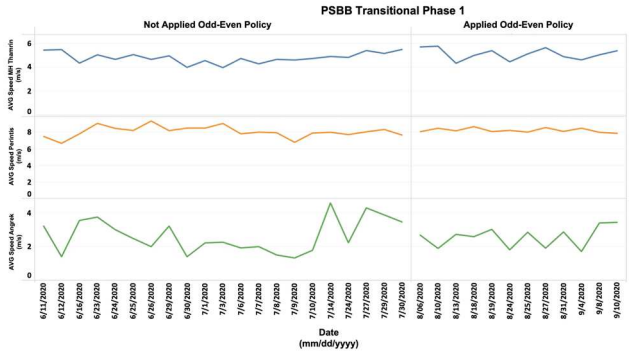


Fig. 2. Daily vehicles' average speed related to MH Thamrin (blue), Merdeka Perintis (Orange), and Neil Anggrek (Green) streets. The Odd-Even policy effect when applied (right) and not (left).

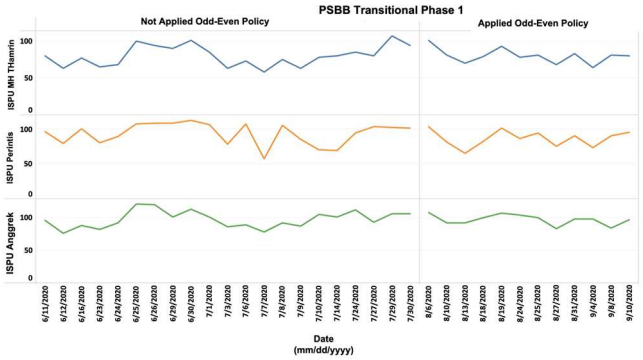


Fig. 3. Daily air pollution to MH Thamrin (blue), Merdeka Perintis (Orange), and Neil Anggrek (Green) streets. The Odd-Even policy effect when applied (right) and not (left).

B. Hypothesis Testing with Paired T-Test

The Paired T-Test was conducted to determine the significant differences when the Odd-Even policy was applied and not towards traffic congestion and air pollution during the COVID-19 pandemic. The Paired T-Test resulted in the p-value, as shown in Table 3 and Table 4. The *p-value* can determine the significant difference in the effect when the Odd-Even policy was applied rather than traffic congestion and air pollution. If the test results give a *p-value* <α, the odd-even application significantly impacts the street. If it provides a significant positive impact, the odd-even policy can be applied too on other roads.

TABLE III. The Paired T-test Results for Assessment of Odd-Even Policy in Terms of Vehicle's Average Speed

Street	Mean (Average Speed)		p-value	α
	Not Apply	*Apply*		
Jl MH Thamrin (DKI1)	4.834	5.398	0.000	0.05
Jl Perintis Kemerdekaan (DKI2)	8.153	8.188	0.794	0.05
Jl Anggrek Nelimurni (DKI5)	2.619	2.572	0.906	0.05

Table 3 shows that from the three streets, only one road has a *p-value* <α, namely MH Thamrin Street. The Odd-Even policy significantly affects only one road, where the average vehicle speed changes significantly. The other chosen streets that apply the Odd-Even policy show no significant effect.

TABLE IV. THE PAIRED T-TEST RESULTS FOR ASSESSMENT OF ODD-EVEN POLICY IN TERMS OF AIR QUALITY

Street	Mean (Air Quality)		p-value	α
	Not Apply	*Apply*		
Jl MH Thamrin (DKI1)	80.750	82.125	0.736	0.05
Jl Perintis Kemerdekaan (DKI2)	94.042	90.875	0.506	0.05
Jl Anggrek Nelimurni (DKI5)	96.333	96.917	0.890	0.05

As shown in Table 4, the *p-value*> α, suggesting that the mean value is the same. There is no significant effect of the Odd-Even policy on air pollution on MH Thamrin, Neil Anggrek, and Merdeka Perintis streets during the COVID-19 pandemic with the same mean value. However, when looking at the *p-value* for each sensor, the DKI2 sensor result is relatively low compared to the other sensors. It shows that the Odd-Even policy in DKI2 has a minor effect in reducing air pollution than other areas.

C. *Relationship between Traffic Congestion and Air Pollution*

This test aims to see the relationship between traffic congestion and air pollution using *Pearson Correlation*. Congestion indicator is obtained from the vehicle's average speed. For air pollution, the maximum ISPU value is taken. Average traffic congestion acts as the independent variable, while air pollution is the dependent variable. Variable values were obtained daily during working days, during odd-even implementation times, and within the PSBB Transitional Phase 1.

The results show that traffic congestion has a slight impact on air quality during the COVID19 pandemic. The MH Thamrin street shows the strongest relationship between the vehicle's average speed (congestion) and air pollution. However, the relationship is positive, which shows that the faster the vehicle's average speed, the greater the air pollution. Ideally, the relationship is inversely proportional, where the slower the vehicle's average speed, the higher the air pollution should be. This result indicates that traffic congestion is not the leading cause of air pollution during the COVID-19 pandemic in DKI1.

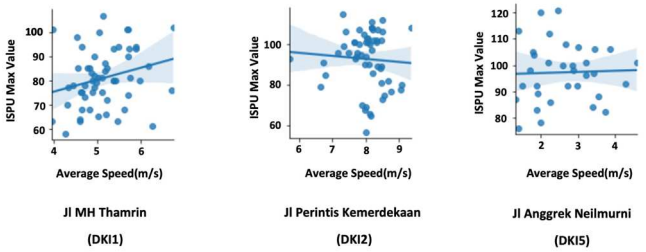


Fig. 4. Relationship between the vehicle's average speed and air quality.

As shown in Figure 4, the three streets show different results; only MH Thamrin (DKI1) has a stronger relationship. Perintis Kemerdekaan (DKI2) street has the right relationship, where the higher the vehicle's average speed, the less air pollution. On the other hand, the relationship on Anggrek Neilmurni (DKI5) street is the weakest. These three streets have a different relationship, which means that traffic congestion in Jakarta slightly impacts air quality during the COVID-19 pandemic. The three roads in Jakarta that have been analyzed demonstrate that Jakarta's primary source of air pollution did not come from traffic congestion during the COVID-19 pandemic.

V. DISCUSSION

In this study, we selected three streets since only three out of five air quality sensors in those streets apply the Odd-Even policy. The selected streets are closest to the air quality sensors that can represent Jakarta overall. The dataset was obtained from the Jakarta Environment Office and the Waze application. Data exploration was carried out to analyze the impact before and after implementing the Odd-Even policy on Jakarta's traffic congestion and air quality [26].

There was a similar air quality pattern on three roads in each sensor in the data exploration stage. It means that the air quality in several areas in Jakarta is the same. It also points to the same possible leading cause affecting air pollution in Jakarta. The highest and lowest ISPU value occurred when the Odd-Even policy was not applied. It means that the Odd-Even policy did not impact Jakarta's air quality during the COVID-19 pandemic. In terms of traffic congestion, it was found that the lowest vehicle's average speed occurs when the Odd-Even policy is not applied. Only one road had the highest vehicle average speed when the Odd-Even policy was applied, suggesting that the Odd-Even policy is not very useful on the other streets. Therefore, the Provincial Government of Jakarta needs to review the road choice to use this Odd-Even policy effectively.

We assessed the significance of the Odd-Even policy impact on traffic congestion and air quality with Paired T-Test. The result showed that only one of the three streets was significantly affected by the Odd-Even policy during the COVID-19 pandemic. The Odd-Even policy application is not very practical on many other streets in Jakarta. It means that the air quality in Jakarta is also not strongly affected by the number of vehicles in Jakarta, as in agreement with other reports [9], [27]. The nominal value of the correlation factor between traffic congestion and air quality in Jakarta also suggests that air quality in Jakarta can be affected by more significant factors, most likely energy and industry [28].

The Provincial Government of Jakarta needs to consider further study in improving air quality apart from restricting transportation mobility through the Odd-Even policy. Moreover, this policy can cause people to switch from private/personal to public transportations [12]. During this pandemic, the Odd-Even policy could increase the spread of COVID-19 in the public transportation cluster. However, the authors acknowledged some limitations in this study: the number of datasets due to a limited number of available sensors surrounding Jakarta.

This study is the first to analyze the impact of large-scale social restriction and the Odd-Even policies on traffic congestion and air pollution in Jakarta during the COVID-19

pandemic. We choose Jakarta because this city is the largest city and capital city in Indonesia with a huge problem in traffic congestion and air pollution. The result of this paper can be a reference for other cities in Indonesia to know the impact of large-scale social restriction and odd-even policies to solve air pollution and traffic congestion problems.

VI. Conclusion and Future Work

This paper explains the effectiveness of the Odd-Even policy in solving traffic congestion and air pollution in Jakarta, especially on MH Thamrin, Neil Anggrek, and Merdeka Perintis streets during the COVID-19 pandemic. The result shows that the traffic congestion on MH Thamrin Street is affected significantly. For air pollution, the Odd-Even Policy does not have a significant impact. The result indicates an insignificant correlation between traffic congestion and air pollution, where the leading cause of air pollution in Jakarta is not from the increase in vehicle volume. In conclusion, the Provincial Government of Jakarta needs to consider the implementation of the existing Odd-Even policy. Future works need to accommodate other sources of the dataset for transport and air pollution in Jakarta. For example, more streets can gain a deeper understanding of traffic congestion and air pollution. On the other hand, the Odd-Even policies can increase the number of people to switch from private/personal to public transportations and potentially increase the spread of COVID-19.

Acknowledgment

The authors thank the Jakarta Environment Office for providing the daily air quality in Jakarta. The contents are only the authors' views, not the Jakarta Provincial Government. There are no financial and conflicts of interest.

References

[1] "Daerah Khusus Ibukota Jakarta." https://indonesia.go.id/province/daerah-khusus-ibukota-jakarta (accessed Jan. 21, 2021).

[2] S. Ghosh and S. Ghosh, "Air quality during COVID-19 lockdown: Blessing in disguise," *Indian J. Biochem. Biophys.*, vol. 57, pp. 420--430, 2020.

[3] S. Muhammad, X. Long, and M. Salman, "COVID-19 pandemic and environmental pollution: A blessing in disguise?," *Sci. Total Environ.*, vol. 728, no. 4, p. 138820, 2020, doi: https://doi.org/10.1016/j.scitotenv.2020.138820.

[4] N. F. Shalihah, "Survei 2019, Jakarta Masuk Peringkat 10 Kota Termacet di Dunia," *Kompas.com*, 2020. https://www.kompas.com/tren/read/2020/01/31/052816565/survei-2019-jakarta-masuk-peringkat-10-kota-termacet-di-dunia (accessed Jan. 21, 2020).

[5] "Hari 11 PSBB Transisi, Senin Pagi Polusi Udara DKI Terburuk," *CNN Indonesia*, 2020. https://www.cnnindonesia.com/nasional/20200615105713-20-513363/hari-11-psbb-transisi-senin-pagi-polusi-udara-dki-terburuk (accessed Jan. 21, 2020).

[6] P. Singhal, G. Tan, K. Basak, and B. Marimuthu, "Visualization of Urban Traffic for the Management of Smart Cities," in *Proceedings of the 9th EAI International Conference on Simulation Tools and Techniques*, 2016, pp. 96–103.

[7] K. Singhania, G. P. Girish, and N. Emodi, "Impact of Odd-Even Rationing of Vehicular Movement in Delhi on Air Pollution Levels," *Low Carbon Econ.*, vol. 7, pp. 151–160, 2016, doi: 10.4236/lce.2016.74014.

[8] R. Li and M. Guo, "Effects of odd–even traffic restriction on travel speed and traffic volume: Evidence from Beijing Olympic Games," *J. Traffic Transp. Eng. (English Ed.*, vol. 3, no. 1, pp. 71–81, 2016, doi: https://doi.org/10.1016/j.jtte.2016.01.002.

[9] X. Ren, J. Liu, and J. Wen, "Effects of Odd-Even Traffic Restriction on Traffic Congestion and Air Quality," in *2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)*, Aug. 2018, pp. 1–6, doi: 10.1109/IEIS.2018.8597909.

[10] "Peraturan Gubernur (PERGUB) Provinsi Daerah Khusus Ibukota Jakarta Nomor 164 Tahun 2016," *Legal Documentation and Information Provincial Government of Jakarta*, 2016. https://jdih.jakarta.go.id/uploads/default/produkhukum/PERATURAN_GUBERNUR_NO.164_TAHUN_.2016__.pdf (accessed Feb. 15, 2020).

[11] R. A. Yori, A. Fuad, and A. Atto'ullah, "IMPLEMENTASI PERATURAN GUBERNUR DKI JAKARTA NO. 164 TAHUN 2016 TENTANG PEMBATASAN LALU LINTAS GANJIL-GENAP DI PROVINSI DKI JAKARTA," Universitas Sultan Ageng Tirtayasa, 2018.

[12] M. E. Fadhli and H. Widodo, "Analisis Pengurangan Kemacetan Berdasarkan Sistem Ganjil-Genap," *Planners Insight Urban Reg. Plan. J.*, vol. 2, no. 2, pp. 36–41, 2020.

[13] F. J. R. Supriana, M. L. Siregar, E. S. W. Tangkudung, and A. Kusuma, "Evaluation of Odd-Even Vehicle Registration Number Regulation Before and After Expansion of the Rule in Jakarta," in *Proceedings of the 2nd International Symposium on Transportation Studies in Developing Countries (ISTSDC 2019)*, 2020, pp. 151–156, doi: https://doi.org/10.2991/aer.k.200220.032.

[14] M. H. Yudhistira, R. Kusumaatmadja, M. F. Hidayat, and others, "Does Trafic Management Matter? Evaluating Congestion Effect of Odd-Even Policy in Jakarta," *Inst. Econ. Soc. Res.*, 2019.

[15] "Ganjil – Genap Diberlakukan kembali mulai 3 Agustus 2020," *Department of Transportation of DKI Jakarta Province*, 2020. https://dishub.jakarta.go.id/ganjil-genap-diberlakukan-kembali-mulai-3-agustus-2020/.

[16] "Pemprov DKI Jakarta resmi menerapkan kembali Pembatasan Sosial Berskala Besar (PSBB)," *Department of Transportation of DKI Jakarta Province*, 2020. https://dishub.jakarta.go.id/pemprov-dki-jakarta-resmi-menerapkan-kembali-pembatasan-sosial-berskala-besar-psbb/.

[17] K. R. Sundaram, S. N. Dwivedi, and V. Sreenivas, *Medical Statistics: Principles & Methods*. Anshan, 2010.

[18] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, and G. Pandey, "Application of student's t-test, analysis of variance, and covariance," *Ann. Card. Anaesth.*, vol. 22, no. 4, pp. 407–411, 2019, doi: 10.4103/aca.ACA_94_19.

[19] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, Aug. 2018, doi: 10.1016/j.tjem.2018.08.001.

[20] N. K. Nissa, Y. Nugraha, C. F. Finola, A. Ernesto, J. I. Kanggrawan, and A. L. Suherman, "Evaluasi Berbasis Data: Kebijakan Pembatasan Mobilitas Publik dalam Mitigasi Persebaran COVID-19 di Jakarta," *J. Sist. Cerdas*, vol. 3, no. 2, pp. 84–94, 2020.

[21] B. Pardamean, R. Rahutomo, T. W. Cenggoro, A. Budiarto, and A. S. Perbangsa, "The Impact of Large-Scale Social Restriction Phases on the Air Quality Index in Jakarta," *Atmosphere (Basel).*, vol. 12, no. 7, 2021, doi: 10.3390/atmos12070922.

[22] P. Goyal and G. Gandhi, "Assessment of Air Quality during the 'Odd-Even Scheme' of Vehicles in Delhi," *Indian J. Sci. Technol.*, vol. 9, Dec. 2016, doi: 10.17485/ijst/2016/v9i48/105801.

[23] Z. Lenkei, "Crowdsourced traffic information in traffic management : Evaluation of traffic information from Waze," KTH, Transport Planning, Economics and Engineering, 2018.

[24] E. Whitley and J. Ball, "Statistics review 5: Comparison of means," *Crit. Care*, vol. 6, no. 5, pp. 424–428, Oct. 2002, doi: 10.1186/cc1548.

[25] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, Sep. 2012, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23638278.

[26] "Pergub No 88 Tahun 2019 tentang Ganjil Genap," *Jakarta Transportation Office*, 2019. https://dishub.jakarta.go.id/download/pergub-no-88-tahun-2019-tentang-ganjil-genap/.

[27] S. Chowdhury, S. Dey, S. N. Tripathi, G. Beig, A. K. Mishra, and S. Sharma, "'Traffic intervention' policy fails to mitigate air pollution in megacity Delhi," *Environ. Sci. Policy*, vol. 74, pp. 8–13, 2017, doi: https://doi.org/10.1016/j.envsci.2017.04.018.

[28] R. H. T. Lauri Myllyvirta, Isabella Suarez, Erika Uusivuori, "Transboundary Air Pollution in the Jakarta, Banten, and West Java provinces," *Center for Research on Energy and Clean Air*, 2020. https://energyandcleanair.org/wp/wp-content/uploads/2020/08/Jakarta-Transboundary-Pollution_FINALEnglish.pdf (accessed Feb. 17, 2020).

Simulation of A Decision Support System Using Data Mining Method with C4.5 Algorithm: A Case Study

Faizal Sudrajat
Department of Information System
Telkom University
Bandung, Indonesia
faizalsudrajat@telkomuniversity.ac.id

Rachmadita Andreswari
Department of Information System
Telkom University
Bandung, Indonesia
andreswari@telkomuniversity.ac.id

Nia Ambarsari
Department of Information System
Telkom University
Bandung, Indonesia
niaambarsari@telkomuniversity.ac.id

***Abstract*—— Every university has a goal to produce quality and highly competitive students. To achieve this goal, Telkom University Information Systems Study Program has three learning methods, namely, class lectures, practicum, and expertise groups. To implement this learning method, stakeholders are needed, commonly referred to as Laboratory Assistants and Expert Group Members. However, in the stakeholder recruitment process there is still subjectivity and difficulty in determining applicants who fit the criteria. In order to improve the quality and excellent quality, every university is required to be able to maximize the processes that can have an impact on the quality and quality. This research was conducted to study the C4.5 Algorithm method which was implemented in the recruitment process for Laboratory Assistants and Expert Group Members. Data were collected through file collection and interviews conducted at the Expert Group Trustees and laboratories. From the simulation, 49 patterns were obtained for the old laboratory assistant recruitment data, 9 patterns for the new laboratory assistant recruitment data and 8 patterns for the Expertise Group recruitment data. From each of these data, the decision-making factors that most influence recruitment graduation are obtained, namely, the Interview attribute for the Laboratory Assistant recruitment data and the Motivation Letter attribute for the recruitment data of Expert Group Members.**

Keywords— Decision Tree; C4.5 algorithm; data mining; decision support system

I. INTRODUCTION

Telkom University, especially the Information Systems study program, has several learning methods, the first is class lectures, the second is practicum lectures and the third is learning through a skill group or commonly called the Expertise Group. Each learning method has advantages and disadvantages, but these three things are important to realize the formation of the ideals of higher education. Learning in the classroom requires lecturers to carry out learning activities, in the laboratory a laboratory assistant is needed for the implementation of practicum activities, and in the laboratory, collaboration between students is needed to create a learning atmosphere regarding the materials in the related Expertise Group and supervised by the supervisor.

Laboratory assistant is an element whose job is to carry out practical guidance to students according to the schedule and objectives. Member of the Expertise Group is an element whose duty is to carry out learning that specifically discusses related Expertise Group material, responds to fellow batches below him, holds public lectures, and holds industrial visits. Each assistant and member of the Expert Group has a great responsibility in carrying out their duties. Thus, to get qualified candidates for assistants and members of the Expertise Group, a good selection is needed.

Constraints that are often found in the process of recruiting prospective Laboratory Assistants and prospective Experts Group Members in the Information Systems Study Program are the difficulty of determining applicants who meet the criteria to become Laboratory Assistants and prospective Experts Group Members based on the criteria that have been set by each Laboratory and Expertise Group. To avoid making a decision to accept assistants and members of the Expertise Group that is influenced by subjectivity factors, a decision support system is needed that can help to decide applicants who can become Laboratory Assistants and Expert Group Members. This system is expected to assist in the recruitment of prospective Laboratory Assistants and prospective Experts Group Members based on predetermined criteria

Based on several studies, the decision tree classification method with the C4.5 algorithm can be used in grouping data effectively according to the desired results. Therefore, this study discusses the prediction simulation of graduation for laboratory assistant candidates for the Information Systems study program at Telkom University based on existing attributes by choosing the decision tree classification method using the C4.5 algorithm.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

II. LITERATURE REVIEW

A. Simulation

Simulation is known as a modeling technique that describes the causal relationship of a system to produce system behavior that is almost the same as the actual system behavior. Simulations can be used to generate an actual historical record and statistical conclusions of all activities that occur [1].

B. Data Mining

Data mining is the process of finding patterns in large data sets and building relationships to solve problems through data analysis. The main goal of data mining is to identify patterns and convert them into more understandable structures for further analysis. The data mining process divides the overall task of finding patterns from the data into a well-defined set of subtasks. Data mining is used to find patterns and evaluate possible future events [2].

1) Data Mining Process

Data mining can be described as the process of obtaining interesting new knowledge, such as important patterns, rules and structures from large amounts of data stored in databases and other information storage areas [3].

In general, the Data Mining process consists of the following steps:

- Data Cleaning: is the process of removing irregular data, inconsistent data, deleting data that is wrong, incomplete, inaccurate, or has the wrong format in the database in order to produce quality data.
- Data Integration: is the process of merging data into a new database from several data sources.
- Data Selections: is the process of analyzing relevant data from an integrated database.
- Data Transformation: is the process of transforming data into a format suitable for processing.
- Data Mining: is a process in which methods are applied to find valuable and hidden knowledge from data.
- Pattern Evaluation: is the process of identifying patterns to be represented in knowledge based.

2) Classification

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts that aim to be used to predict the class of objects whose class labels are unknown [3]. Classification can also be called a method for organizing data systematically or according to some rules. Data classification consists of two process steps, namely the learning stage where in this step will build a classification model, and the second step is the classification stage where the model that has been built will be used to predict the class label for the new data given [4].

3) Decision Tree

Decision tree is a method that exists in classification techniques in data mining. The decision tree method converts very large data into a decision tree that represents the rules. Decision trees are also useful for exploring data, finding hidden relationships between the number of candidate input variables and a target variable [3]. The concept of a decision tree is to turn data into a decision tree with decision rules that have been made. The benefit is its ability to break down complex decision-making processes into simple ones.

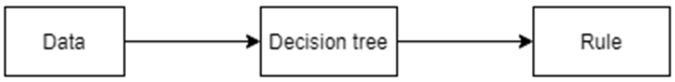


Fig 1. Decision tree concept

A decision tree is also referred to as a flow diagram that is shaped like a tree structure where each internal node represents a test of an attribute, and the branch is the output of the tests carried out and then the leaf node represents the class distribution. The top node is called the root node. The test path is that first all data must pass through the root node, and lastly, pass through the leaf node which will conclude the class prediction for the data.

Decision tree is one of the algorithms to solve classification problems on a data. The reason for using Decision tree as a method for data classification problems is because this method generalizes all data, therefore all data classified will produce a high level of accuracy. In addition, the level of prediction precision generated from the Decision Tree method is better than other classification methods, this can be proven in journal references that use case studies [6].

TABLE I. CLASSIFIER COMPARISON

Method	Level of accuracy
Decision Tree	98.89 %
Naive Bayes	96.67 %
Random Forest	95.56 %

4) C4.5 Algorithm

C4.5 algorithm and decision tree are two inseparable models, because to build a decision tree, C4.5 algorithm is needed. In the late 1970s to early 1980s. J. Ross Quinlan a researcher in the field of machine learning developed a decision tree model called ID3 (Iterative Dichotomiser), although this project had been previously created by E.B. Hunt, J. Marin, and P. T. Stones. Then Quinlan made an algorithm from ID3 development called C4.5 which is based on Supervised learning [7].

The C4.5 algorithm is an algorithm used to build a decision tree from the data. The C4.5 algorithm is a development of the ID3 algorithm which is also an algorithm for building a decision tree. The C4.5 algorithm recursively visits each decision node, choosing the optimal branch, until no more branches are possible. The C4.5 algorithm uses the concept of information gain or entropy reduction to choose branching [8].

In general, the C4.5 algorithm for constructing a decision tree is as follows [9]:

1. Prepare training data
2. Calculate the value of Entropy using the formula:

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log 2 p_i \quad (1)$$

Where,

S : Case Collection.

m : Number of partitions S.

p_i : the probability obtained from the number of cases Si divided by S

3. Look for the gain value using formula:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (2)$$

Where:
S : Space (data) sample used for training.
A : Attributes.
N : number of attribute partitions A
$|S_i|$: Number of samples for the value of *i*.
$|S|$: The total number of data samples.
$Entropy(Si)$: Entropy for samples that have a value of *i* after the Gain value is obtained

4. Look for the Split Information value with the formula

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log 2 \frac{|S_i|}{|S|} \quad (3)$$

where:
S : The sample space (data) used for training.
A : Attribute.
N : number of attribute partitions A
$|S_i|$: The number of samples for the value of *i*.
$|S|$: Total sample data

5. Look for the Gain Ratio value with the formula

$$Gainratio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \quad (4)$$

6. The highest gain ratio value will be used as the root node and Node 1 will be formed
7. Repeat the 2nd process until all branches have been counted
8. The results of these calculations will produce a rule which will later form a decision tree

5) Classification Performance Evaluation

Evaluation of classification performance is a very important process. Evaluation of classification performance aims to find out how the performance of a classification is whether the resulting classification results are in accordance with existing rules or vice versa. The performance evaluation of the classification model is based on the number of tests that were predicted correctly and incorrectly. These calculations are tabulated in a matrix table commonly known as the Confusion matrix [10]. The table provides details regarding the correct and incorrect classifications. The Confusion Matrix table is divided into 2 classes, namely the prediction class and the actual class as shown in the following table.

TABLE II. CLASSIFIER COMPARISON

		Predicted	
		Yes	no
Actual	yes	TP	FN
	no	FP	TN

- True Positive (TP): the positive proportion contained in the data set that is classified as positive.
- False Negative (FN): the proportion of negatives contained in the data set that are classified as negative
- False Positive (FP): the proportion of negatives in the data set that are classified as positive.
- True Negative (TN): the proportion of positives in the data set that are classified as negative

The confusion matrix is very useful for analyzing how well the classifications used from various classes are [4]. TP and TN provide information on prediction results according to actual conditions, while FP and FN provide information on prediction results that do not match actual conditions.

There are several steps that can be taken to conduct an evaluation [9], as follows:

- Precision
  Precision aims to measure the proportion of cases that are predicted to be positive which are also true positives in the actual data. precision can be calculated using the formula

$$\Pr ecision = \frac{TP}{TP + FP} \quad (5)$$

- Recall
  Recall aims to measure the proportion of true positives (TP) to positive tuples that are correctly identified. Precision can be calculated using the formula

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (6)$$

- Accuracy
  Accuracy aims to find out the correct prediction. Accuracy can be calculated using the formula

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

III. METHODS

A. Systematic Research

1) Review stage

At this stage the identification of the problem is carried out. Problem identification is done through case studies and literature studies. At this stage, solutions and objectives are determined within the limitations of this study.

2) Collection stage

The data preparation stage will be carried out using the data of the prospective laboratory assistant selection participants that have been previously processed, so that the authors can directly distribute training data and testing data.

3) Analyzing stage

The data processing stage will be carried out with the input data which has been divided into the categories of training data and testing data. The training data is processed to get the gain value. The data that is processed is selected the data with the highest gain value which is then used as a determination of the rules for classification. After that, the process of implementing the rules that have been determined and finally performing the performance to determine the accuracy of this decision-making system

4) Interpretation Stage

At this stage an analysis will be carried out regarding the classification process that has taken place, what factors affect the acceptance of laboratory assistants and provide suggestions for improving performance for further research

IV. Results

A. Modelling Implementation

In the modeling process using the C4.5 algorithm, the input data generated from the data preprocessing stage was obtained as many as 947 records for the old Laboratory Assistant recruitment data, 95 records for the new laboratory assistant recruitment data and 105 records for the Expert Group recruitment data. The amount of data used for training is 70%. The implementation of the C4.5 algorithms is carried out using the RapidMiner tool. The process of forming the prediction model of the C4.5 algorithm begins by performing calculations on the Training Set including the calculation of entropy, information gain, split information, and gain ratio. The attribute that has the greatest gain ratio value is the Interview attribute with a gain ratio value of 0.323 as the root node in the Laboratory Assistant Recruitment Training Set. Then do the same thing repeatedly to find the next node by calculating the gain ratio value for each remaining attribute until all attributes are used up and all cases have been classified according to their class. The final result of the decision tree from all calculations on the Laboratory Assistant and Expertise Group data is shown in the figure below:

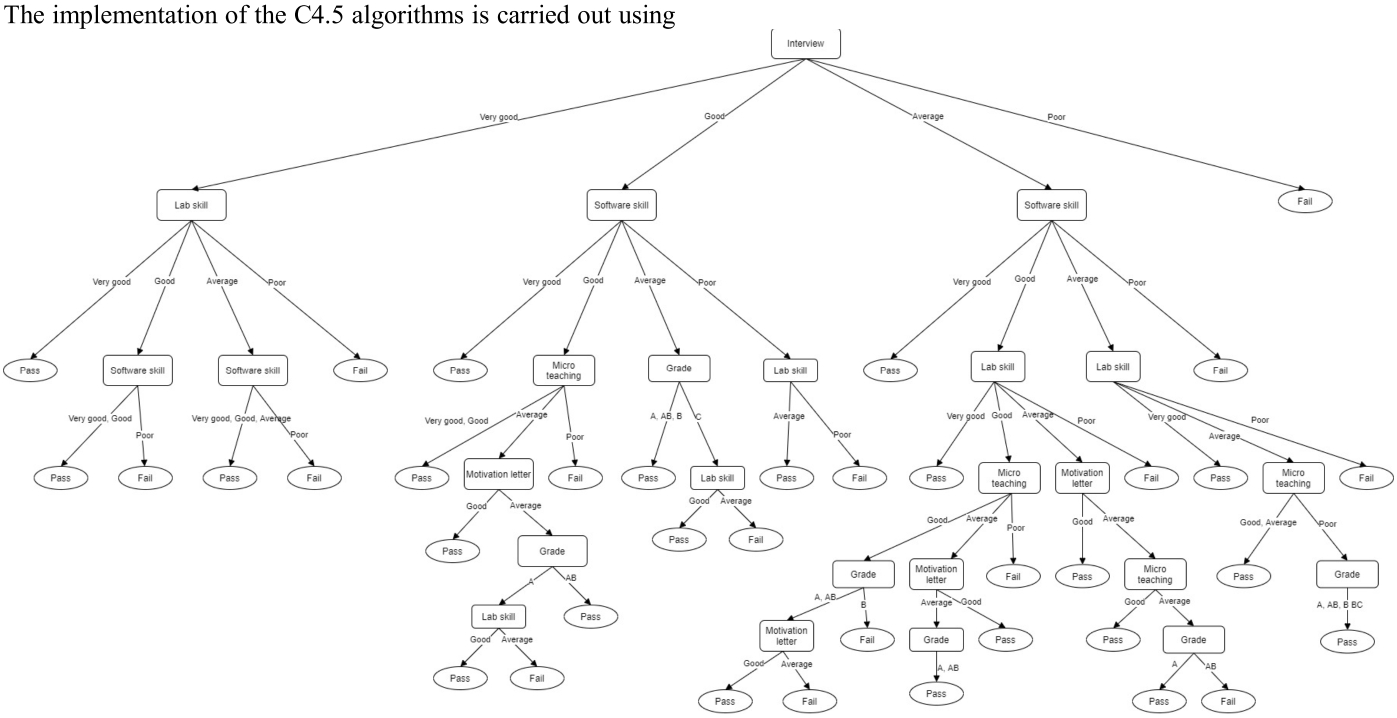


Figure 2. Overall Decision Tree Transformation Laboratory Assistant Recruitment

B. Prediction Model Implementation

In the Prediction Model implementation stage, the percentage of data used is 70% training set and 30% testing set, both laboratory assistant recruitment data and Expert Group recruitment data.

The number of patterns generated from the simulations carried out using data mining techniques are 49 patterns for old laboratory assistant recruitment data, 9 patterns for new laboratory assistant recruitment data and 8 patterns for Expertise Group recruitment data.

All attributes are very influential on the prediction of graduation. For Laboratory Assistant recruitment data, the most influential attribute is Interview. In the Expert Group Member recruitment data, the most influential attribute is the Motivation Letter. The Interview and Expertise Group attributes are the most influencing factors for graduation because these attributes are the root node or the root of the resulting decision tree.

TABLE III. Results of the Implementation of the Old Data for the Recruitment of Laboratory Assistants

Accuracy: 97.52%	**True Fail**	**True Pass**	**Class Precision**
Pred. Fail	134	3	97.81%
Pred. Pass	4	141	97.24%
Class Recall	97.10%	97.92%	

The table above shows the results of the implementation of the prediction model on the laboratory assistant recruitment data. For the accuracy of the prediction model, the results obtained are 97.52%. for class recall, 97.10% for class recall true Failed and 97.92% for class recall true Passed. The results of the precision class, the results obtained are 97.41% for the Fail prediction and 97.24% for the Pass prediction.

TABLE IV. Implementation Results of the Expert Group Recruitment Data

Accuracy: 93.55%	**True Fail**	**True Pass**	**Class Precision**
Pred. Fail	5	2	71.43%
Pred. Pass	0	24	100%
Class Recall	100%	92.31%	

The table above shows the results of the implementation of the prediction model on the Expert Group recruitment data. The results of the prediction model accuracy are 93.55%, 100% for the recall class true Fail and 92.31% for the recall class true Pass. In the precision class, the results obtained are 71.43% for the Fail prediction and 100% for the Pass prediction.

C. Implementation of Rules on the Web

The final result of the decision tree with the c4.5 algorithm is to produce a rule that was previously used in the implementation of the prediction model for data testing. These rules will determine whether a data record belongs to a category, in this study the categories are "Passed" and "Fail". These rules will be implemented in a web-based decision-making system for the selection of laboratory assistant candidates and the selection of members of the expertise group.

D. Classification Result Analysis

The classification process produces several outputs. In the laboratory assistant data classification process, the interview attribute is the main factor to be accepted as a laboratory assistant. Meanwhile, in the process of classifying Expertise Group Member data, the motivation letter attribute becomes the main factor to be accepted as an Expert Group Member.

1) Laboratory Assistant Recruitment Decision Making Pattern

In the classification process on Laboratory Assistant data, the interview attribute is the main factor in passing to become a laboratory assistant. The following is an example of some of the patterns that have been transformed by researchers to make it easier to understand the patterns resulting from the classification process of old and new data.

TABLE V. OLD DATA LABORATORY ASSISTANT RECRUITMENT DECISION MAKING PATTERN

No	I	SS	MT	ML	G	LS	Decision	Sum	
								Pass	Fail
1	Good	Good	Good				Pass	74	0
2	Good	Good	Average	Good			Pass	15	0
3	Good	Good	Very Good				Pass	16	0
4	Very Good	Very Good				Good	Pass	14	0
5	Very Good	Good				Average	Pass	3	0
6	Good	Very Good					Pass	14	0
7	Good	Good	Average	Average	A	Good	Fail	0	2
8	Good	Average			B	Average	Fail	0	2
9	Good	Poor				Poor	Fail	0	1
10	Average	Good	Good	Good	AB	Good	Fail	2	2
11	Average	Average	Average	Average	A	Average	Fail	3	5
12	Average	Good	Average	Average	A	Average	Fail	0	1

Information:
MT : Micro Teaching
ML : Motivation Letter
SS : Software Skill
LS : Lab Skill
G : Grade
I : Interview
CV : Curriculum Vitae

2) Decision Making Patterns for Recruitment of Expert Group Members

In the classification process on the Expertise Group data, the attribute the motivation letter is the main factor in passing to become a member of the expertise. The following is an example of some patterns that have been transformed by researchers to make it easier to understand the patterns resulting from the classification process.

TABLE VI. DECISION MAKING PATTERNS FOR RECRUITMENT OF EXPERT GROUP MEMBERS

No	ML	GPA	CV	I	Decision	Sum	
						Pass	Fail
1	Good	AB			Pass	13	0
2	Good	B			Pass	22	0
3	Good	BC		Good	Pass	1	0
4	Good	BC		Very Good	Pass	2	0
5	Very Good				Pass	21	0
6	Good			Average	Fail	0	1
7	Average		Average	Average	Fail	0	1

Information:
ML : Motivation Letter
GPA : Grade Point Average
I : Interview
CV : Curriculum Vitae

3) Pattern and Performance Analysis based on Recruitment Data

Furthermore, an analysis of the three data will be carried out by comparing the state of the data and the performance results along with the pattern.

TABLE VII. DECISION MAKING PATTERNS FOR RECRUITMENT OF EXPERT GROUP MEMBERS

Data	Data State	Pattern	Performance				
			Accuracy	*Precision*		*Recall*	
				P	F	P	F
Old Laboratory Recruitment	A total of 36.42% of the scores for the 2015 and 2016 laboratory data were taken from the results of the assessment and recapitulation of each laboratory, and the remaining 63.57% was taken from the results of interviews and value adjustments with the rubric determined by the research team.	49 patterns	97.52%	97.41%	97.24%	97.10%	97.92%
New Laboratory Recruitment	A total of 85.13% of the scores for the 2017 class of laboratory data were taken from the results of the assessment and recapitulation of each laboratory, and the remaining 14.38% was taken from the results of interviews and value adjustments with the rubric determined by the research team.	9 patterns	100%	100%	100%	100%	100%
Expertise Group Recruitment	A total of 31.17% of the scores for the Expertise Group data were taken from the results of the assessment and recapitulation of each laboratory, and the remaining 68.82% were taken from the results of interviews and value adjustments with the rubric determined by the research team.	8 patterns	93.55%	71.43%	100%	100%	92.31%

From the table above, it can be seen that the classification performed on the three data has good accuracy. However, the number of patterns generated from the old data Laboratory Recruitment data is more than the pattern generated from the new data Laboratory Recruitment data and from the Expert Group Recruitment data. This is due to several factors, including the amount of training data, variance of attribute values, and the number of attributes. These three factors are interrelated. In the three data, it can be seen that the recruitment of new data laboratory assistant recruitment data shows very good results which produce a value of 100% for all aspects of performance. However, the patterns generated by the data are few.

Based on the simulations carried out, the researcher assumes that among the three data, good and effective performance is the old data laboratory assistant recruitment data pattern, because the resulting patterns are many, although the performance is smaller than the new data laboratory assistant recruitment data.

The new data laboratory assistant recruitment data pattern results in a small performance against the old laboratory data set testing because the resulting pattern is not able to classify all conditions in the data contained in the testing set.

V. CONCLUSION

The number of patterns generated from the simulations carried out using data mining techniques are 49 patterns for old laboratory assistant recruitment data, 9 patterns for new laboratory assistant recruitment data and 8 patterns for Expertise Group recruitment data.

All attributes are very influential on the prediction of passing. For Laboratory Assistant recruitment data, the most influential attribute is Interview. For membership recruitment Expertise Group, the most influential attribute is Motivation Letters for passing because these attributes are the root node or the root of the resulting decision tree.

In the development of further research, it is expected that the data used are real data, so that the pattern formed can be implemented in the relevant agency. Then the amount of training data, the variance of attribute values and the number of attributes must be considered, so that the pattern formed can see the overall condition of the data record.

REFERENCES

[1] O. Veza, "Simulasi Pengendalian Persedian Gas Menggunakan Metode Monte Carlo dan Pola LCM (Studi Kasus di PT.PKM Group Cabang Batam)," *Jt-Ibsi*, vol. 01, no. 01, pp. 1–15, 2016.

[2] K. . Sanjana, "A Survey of Data Mining Tasks," *Int. J. New Technol. Res.*, vol. 4, no. 3, p. 263117, 2018.

[3] D. Yunita, "Perbandingan Algoritma K-Nearest Neighbor dan Decision Tree untuk Penentuan Risiko Kredit Kepemilikan Mobil," *J. Inform. Univ. Pamulang*, vol. 2, no. 2, p. 103, 2017, doi: 10.32493/informatika.v2i2.1512.

[4] J. Han, M. Kamber, and J. Pei, "Data Mining," in *Data Mining (Third Edition)*, Third Edit., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 1–38.

[5] R. Kumara and C. Supriyanto, "Klasifikasi Data Mining Untuk Penerimaan Seleksi Calon Pegawai Negeri Sipil 2014 Menggunakan Algoritma Decision Tree C4.5," *UDiNus Repos.*, pp. 1–10, 2014.

[6] N. Frastian, S. Hendrian, and V. H. Valentino, "Komparasi Algoritma Klasifikasi Menentukan Kelulusan Mata Kuliah Pada Universitas," *Fakt. Exacta*, vol. 11, no. 1, p. 66, 2018, doi: 10.30998/faktorexacta.v11i1.1826.

[7] M. Tabrani, "Klasifikasi Penerima Beasiswa Kopertis Dengan Menggunakan Algoritma C.45," *Pilar Nusa Mandiri J. Comput. Inf. Syst.*, vol. 12, no. 1, pp. 72–80, 2016, [Online]. Available: http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/261.

[8] I. Rahmayuni, "Perbandingan Performansi Algoritma C4.5 dan Cart Dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang," *Teknoif*, vol. 2, no. 1, pp. 40–46, 2014.

[9] D. melina Agustina and Wijanarto, "Analisis Perbandingan Algoritma ID3 Dan C4 . 5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum pada PDAM Kabupaten Kendal," *J. Appl. Intell. Syst.*, vol. 1, no. 3, pp. 234–244, 2016.

[10] Gorunescu, *Data Mining. Concepts, Model, and Techniques*, 12th ed. Berlin: Springer, 2011.

Search System for Translation of Al-Qur'an Verses in Indonesian using BM25 and Semantic Query Expansion

Arif Rhizky Gilang Purnama[1], Intan Nurma Yulita[2], Afrida Helen[3]
Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran
Department of Computer Science, Universitas Padjadjaran
Padjadjaran University
Sumedang, Indonesia
e-mail:[1] arif17003@mail.unpad.ac.id, [2]intan.nurma@unpad.ac.id , [3]helen@unpad.ac.id

***Abstract*—Al-Qur'an is a source of life guidance for Muslims. The digital version of the Qur'an is already available on the android platform. The Indonesian Ministry of Religion provides an official application called the Ministry of Religion's Qur'an which has a translation search system in it. However, adjustments need to be made so that Indonesian people can look for translations of verses that have a definite order based on their level of relevance to the keywords entered. This study aims to develop a search system for translations of Indonesian-language Qur'an verses that already exist in the Ministry of Religion's Qur'an application.**

This study uses the BM25 algorithm with Word2vec as the Semantic Query Expansion method. Data as many as 6236 translated documents in the application are used to create a model of the search system. Tests on hyperparameters are carried out to get the most optimal model. The research results obtained several hyperparameter values in SQE including a window of 7, and a query expansion term of 1. In the BM25 hyperparameter, the optimal condition is obtained when the k1 variable is 1.8 and the b variable is 0.85. The search system was evaluated using the Mean Average Precision and compared with the search system that was previously available in the Ministry of Religion's Qur'an. The MAP score increased with the proposed method, from 0.53718 to 0.66556.

Keywords— Al-Qur'an, BM25, Information Retrieval, PySastrawi, Semantic Query Expansion, Word2vec

I. INTRODUCTION

Technology that is increasingly developing can provide the Qur'an in digital form. The Ministry of Religion's Qur'an application is a form of digital Qur'an that is officially provided by the Ministry of Religion of the Republic of Indonesia. This application contains verses of the Qur'an and their translations in Indonesian along with their interpretations.

Indonesia is one of the countries with the highest Muslim population in the world. Based on population census data from the Central Statistics Agency (BPS) in 2010, the Muslim population in Indonesia reached 87.18% of the total population of Indonesia [19]. The Qur'an plays a very important role as a guide for Muslims. Muslims can look for laws that cover all aspects of human life in the Qur'an. With increasingly developed technology, searching for a document can be done quickly using machines that are effective in data computing processes. The specific technology for finding the desired information from a data set is a search engine[1]. This search engine technology can be used to search for translations of verses of the Qur'an in Indonesian. This search system is found in the Ministry of Religion's Qur'an application version 2.1.4, but cannot yet calculate and sort documents based on their relevance level ranking.

Several methods or algorithms can be implemented for searching a document. One of them is BM25. BM25 is a search algorithm that uses a probability approach [2]. This process is done by searching for each keyword in the query entered by the user. However, BM25 cannot find relevant documents for terms that have a semantic relationship with the initial query. So, there will be relevant documents that may not appear even though the document is relevant to what the user wants to search for. To overcome this problem, we need semantic query expansion, a method to get terms that have a semantic relationship with the initial query. The term obtained will be combined with the initial query. In previous research on information retrieval microblogs, it is proved that the score performance is increased with query expansion [3].

To answer the existing problems, this research is intended to improve the search system that already exists in the Ministry of Religion's Qur'an application, namely a search system for translations of Indonesian-language Al-Qur'an verses using the BM25 algorithm. This study also uses Word2vec as a semantic query expansion to find terms that have a semantic relationship with the initial query.

II. LITERATURE REVIEW

A. Text Mining

Text mining is a process to get a pattern or meaningful knowledge from an unstructured collection of texts. This process is carried out automatically using machine learning techniques [4]. Text mining can identify facts to relationships in words hidden in large collections of text (big data). After the text is extracted, the information obtained can be used for further analysis or displayed using graphs that can be easier for humans to see.

The information obtained from the text mining process can be used for the optimization process of the search system. Data in the corpus can be entered into the Word2vec model, a model that can map each term in the corpus in vector space.

B. Text Preprocessing

Text preprocessing is a technique for cleaning and preparing data in the form of text. This process is important to do before training a machine that aims to complete a particular task. Text that has gone through preprocessing will be easier to predict and analyze by machines. It can also improve the

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

results of the machine learning process. The following are the types of text preprocessing.

1) Case folding

The simplest process in text preprocessing is case folding. The purpose of this case folding is to convert the text in the document into lower case, also to remove characters other than 'a-z' such as numbers and punctuation marks in sentences.

2) Tokenization

Tokenization is a technique for dividing the sentence into parts that smaller. These little parts form a word, or which can be called a token.

3) Stop-word removal

Stop-word is a common word that usually appears in large numbers and is considered to not have meaning, so it must be cleaned from the corpus. Examples of stop words in the Indonesian language are "*dan*", "*yang*", "*di*", etc. This stop-word removal is intended to allow the engine to focus on words that are more important to learn.

4) Stemming

Stemming is the process of removing the inflection of a word to its basic form. For example, the words "*memakan*" dan "*dimakan*" will be transformed into "*makan*" [5].

C. Semantic Query Expansion

Query expansion is a technique for changing or transforming the initial query so that the engine can better understand the actual intent and context that the user wants to search for [6]. One type of method in query expansion is corpus-specific terms. This approach is done by analyzing terms that are different but have a similar context (semantic relationship) in the corpus [7]. This can then be called semantic query expansion. The semantic query expansion method is part of the automatic query expansion method. This method can generate a term that has a semantic relation to the term in the initial query.

Word2vec is a model that can produce word embeddings, which is a vector representation of words with numeric values [8]. The model training process is carried out by receiving input in the form of a text corpus which then produces a vector space, which usually has hundreds of dimensions. Each word in the corpus is positioned in vector space based on the training process carried out by the Word2vec model. The model can map each term accurately on the vector space by looking at the context of the closest term in a sentence. After each term is mapped on vector space in the form of word embeddings, the level of similarity between words can be calculated using cosine similarity.

Word2vec has 2 types of architecture, namely Continuous Bag-of-words (CBOW) and Skip-gram. In CBOW, the context representation of the closest words in a sentence is combined to predict the target word in the sentence. On the other hand, Skip-gram only accepts input of 1-word representation to predict the context adjacent to the word [9]. In its use for query expansion, the CBOW architecture is better than the skip-gram architecture [10]. Initial queries can be combined with terms generated by the Word2vec model, i.e. terms that have semantic relationships. These merged queries will be used as input to the search system with the following illustration:

$$q_0 = t_1 + t_2 + \cdots + t_n \quad (1)$$

$$q_e = t_1 + \mathrm{sim}(t_1) + t_2 + \mathrm{sim}(t_1) + \cdots + t_n + \mathrm{sim}(t_n) \quad (2)$$

Information:

q_0 : Initial query
q_e : Query that has been expanded
t_i : Term *i* in query
$sim(t_i)$: Term QE from the calculation of *cosine similarity*

D. Information Retrieval

Information Retrieval is a branch of science in text mining that focuses on getting information as desired. This includes how to represent information, information storage techniques, and methods for accessing that information such as web pages, online catalogs, structured documents, and multimedia objects.

At present, research on IR includes web page search, text classification, system architecture design, interface display, data visualization, to information filtering. From a computer point of view, IR focuses on how to compile an efficient document index, how to optimally process queries from users, and how to build algorithms to determine relevant documents based on the ranking of these documents [11].

E. BM25

Okapi BM25 or commonly referred to as BM25 is a ranking algorithm that is used to sort the match results against documents, based on the terms they are looking for. This BM25 is the best formula in the best match class because this formula is effective and has accuracy in sorting documents based on the searched query [12]. The BM25 method equation has a value of k1, k3, and a value of b as a parameter or constant value.

$$BM25 = \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right).\left(\frac{(k_1+1)tf_{td}}{K + tf_{td}}\right).\left(\frac{(k_3+1)tf_{tq}}{k_3 + tf_{tq}}\right) \quad (3)$$

$$K = k_1.((1-b) + (b(\frac{L_d}{L_{ave}})) \quad (4)$$

Information:

N : Number of terms in the corpus
df_t : Number of terms containing query
tf_{td} : Number of term frequency
tf_{tq} : Number of query frequency
k_1 : Hyperparameter for TF calibration on document
k_3 : Hyperparameter for TF calibration on query
b : Hyperparameter for document length calibration
L_d : Document length by word count
L_{ave} : Average document length by word count

F. Evaluation Method

The evaluation process in the information retrieval system is used to assess how well the information is obtained based on the queries entered by the user. Mean Average Precision is a method of evaluating the performance of the search system in the field of information retrieval. In information retrieval, precision is the ratio of the relevant documents to the entered query with all retrieved documents [13]. The equation for calculating precision is as follows:

$$precision = \frac{|(relevant\ doc) \cap (retrieved\ doc)|}{|(retrieved\ doc)|} \quad (5)$$

By default, precision counts all relevant documents to be included in the calculation. This constraint can be changed so that the calculation is only for the top k search results, where

k is the maximum value of retrieved documents. This type of calculation is called precision at k or P@k. The calculation of precision at k is used in determining the average precision atk (AP@k), where k is the maximum limit of documents to be retrieved in one query. The equation for calculating average precision for one query is as follows:

$$AP@k = \frac{1}{GTP}\sum_{k}^{n} P@k.rel@k \quad (6)$$

Information:
k : Maximum limit of documents taken
GTP : Number of relevant documents at k limit
$P@k$: Precision at k
$rel@k$: Relevance function at k

The relevance function is an indicator function that will be worth 1 if the document at k is relevant, and 0 if the document at k is irrelevant. Calculation of precision for queries that are more than 1 can be calculated using the mean average precision. The mean average precision equation is as follows:

$$mAP = \frac{1}{Q}\sum_{i=1}^{Q} AP_i \quad (7)$$

The calculation of the mean average precision can be calculated independently by determining the variable k (top rank limit) and variable Q (number of queries).

III. METHODOLOGY

The object under study is the translation of the Indonesian-language Qur'anic verses in the Ministry of Religion's Qur'an, an application launched by the Ministry of Religion of the Republic of Indonesia. The data used in the study is data sourced from the Ministry of Religion's Qur'an website.

A. Data Collection

The data collected is data that is already available in the open-source repository created by Rio Astamal on the GitHub platform. The data in the repository is sourced from the Ministry of Religion's official Qur'an application. The data consists of 6236 text files which are divided into 114 *surah* folders. The documents are converted into CSV that has 2 columns, namely *ayah_number* which contains a three-digit letter number and a three-digit verse number connected by a colon (:), and *ayah_content* which contains a translation of the Qur'anic verse in Indonesian.

B. Data Preprocessing

The text preprocessing stage is an important stage to do. The Al-Qur'an dataset and user queries will be processed before entering the next stage. Each document will go through the process of case folding, noise removal, stop-word removal, stemming, and tokenization.

C. Model Testing Data

In the testing process using Mean Average Precision method, we need a set of queries with a fixed number. In this study, the number of queries is limited to only 5, but with 2 different types of queries based on the number of keywords. Test queries that match these criteria are listed in Table I.

TABLE I. TEST QUERY LIST

No	Query Type	Query
1	One keyword	puasa
2		haji
3		berkurban
4		zina
5		zakat
6	Two keywords	kafarat berpuasa
7		berbuat maksiat
8		ibadah salat
9		hari kiamat
10		maskawin menikahi

This limitation is chosen as a form of adjustments to the Ministry of Religion's Qur'an search system which does not have the capability for stemming initial queries. Therefore, the query chosen for this test uses queries that can generate documents in the Ministry of Religion's Qur'an search system. This is because the comparison process cannot be carried out if the results on one of the search systems being compared cannot produce results.

D. BM25 Formulation Process

The formation process needs to be adjusted to the program to be made. In this study, because there is no relevant feedback or supporting information about which documents were previously relevant, the IDF formula that suitable for these conditions is the formula of equation 8 in the research of Beaulieu et al [14].

$$IDF = \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \quad (8)$$

The next limitation is that the query entered by the user is relatively short, so it does not require weight normalization in the query. Equation 9 is a formula that has removed the query weight normalization process where the IDF value is in the equation 8 and the K value is in equation 4.

$$BM25 = IDF.\left(\frac{(k_1 + 1)tf_{td}}{K + tf_{td}}\right) \quad (9)$$

E. Word2vec Model Calculation Process

The Word2vec model training process will generate word embedding for each word entered into the model. To calculate the degree of similarity of a word in Word2vec, the cosine similarity formula is used. The value of cosine similarity is a number between 0 and 1. The equation is as follows:

$$similarity = \cos(\theta) = \frac{A \bullet B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (10)$$

IV. RESULT

A. Hyperparameter Testing Analysis

To build the most optimal search system model, hyperparameter testing is required. Some of these hyperparameters include variables *k1* and *b* in BM25, *window size* in Word2vec, and *term count* which will be added to the semantic query expansion method.

1) Window size Word2vec

Window size has a default value of 5. The window sizes used in this experiment are 3, 5, 7, and 10. Other hyperparameter variables are taken from the default values, namely k1 of 1.2 and b of 0.75.

TABLE II. RESULT OF TESTING HYPERPARAMETER WINDOW SIZE WORD2VEC

Query Term	Window	k1	b	Total QE Term	MAP
1	3	1.2	0.75	1	0.53542
1	5	1.2	0.75	1	0.60465
1	**7**	**1.2**	**0.75**	**1**	**0.62924**
1	10	1.2	0.75	1	0.60063

In Table II. it can be concluded that a larger window size will improve the model's performance in finding terms that have semantic relationships. However, empirically, a window size that is too large does not always result in a good MAP score because the model will map word embeddings differently. This causes the output when looking for similar terms with the to be different. For example, the closest term to "*zina*" in window size 7 is "*piara*", while in window size 10 it is "*menikah*". will affect the search results for documents that cause different MAP scores.

2) Number of Terms in Semantic Query Expansion

The number of QE terms used in this experiment were 1, 2, and 3. In this experiment, other hyperparameter variables were taken from the most optimal value in the previous experiment. The experimental results can be seen in Table III. The number of terms of 1 gives the best MAP score at 0.62924.

TABLE III. THE RESULTS OF HYPERPARAMETER TEST RESULTS ON THE NUMBER OF TERMS ON SQE

Query Term	Window	k1	b	Total QE Term	MAP
1	**7**	**1.2**	**0.75**	**1**	**0.62924**
1	7	1.2	0.75	2	0.51758
1	7	1.2	0.75	3	0.44555

From this MAP score, it can be explored further for the analysis of each unit of average precision of the search results. Table IV is the result of the average precision calculation for the query unit. The AP@10 score in bold is the highest score among the other QE term hyperparameters.

TABLE IV. THE RESULTS OF THE HYPERPARAMETER TEST ON THE NUMBER OF TERMS ON SQE

Query	AP@10		
	QE Term 1	*QE Term 2*	*QE Term 3*
puasa	0.41667	**0.62500**	0.61111
haji	**0.63252**	0.25000	0.20000
berkurban	0.80694	**0.91071**	0.63333
zina	**0.95937**	0.70218	0.45000
zakat	**0.34167**	0.10000	0.33333

The AP@10 score for QE Term 1 tends to be better than the others. For the number of QE term 2, the query "*puasa*" and "berkurban" have increased scores. As for the QE term worth 3, there is no increase in the score at all.

A more detailed picture can be obtained from the results of the analysis on one of the queries. In this case, the query "haji" will be analyzed to see the cause of the drastic decrease in the MAP score from the number of QE term 1 to the number of QE term 2. Figure 1 is one example of three terms that have similarities to the term "haji", which is obtained with the cosine similarity formula in the equation 10 formula.

When the number of QE terms to be added is 1, only 1 term that has the highest cosine similarity score will be added to the initial query, namely "*umrah*". The term "*umrah*" has a high similarity in the context of worship performed at the Grand Mosque, Mecca. However, when the number of QE terms to be added is 2, then there are two terms with the highest cosine similarity scores that will be added to the initial query, namely "*umrah"* and "*ibadah*".

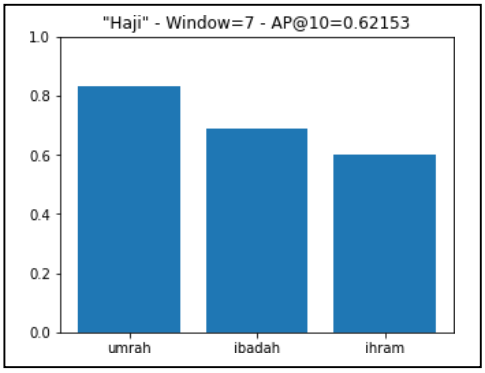


Fig. 1. Cosine Similarity Score By Entering the Term "*Haji*"

TABLE V. TERMS FOUND IN SEARCH "HAJI" (QE TERM 1)

Document Rank	Total QE Term	Term in Result	
		haji	*umrah*
1	1	3	0
2		2	0
3		1	0
4		1	1
5		1	0
6		1	0
7		1	0
8		1	0
9		1	0
10		1	1

Table V shows the number of terms that appear in each document in the search results. The search results displayed are documents with the top 10 rankings for QE Term 1. There are two documents containing the query expansion term "*umrah"*, namely documents with rank 4 and rank 10.

TABLE VI. TERMS FOUND IN SEARCH "HAJI" (QE TERM 2)

Document Rank	Total QE Term	Term in Result		
		haji	*umrah*	*ibadah*
1	2	2	0	0
2		1	0	0
3		1	0	0
4		1	0	0
5		0	0	1
6		0	0	1
7		0	0	1
8		0	0	2
9		0	0	1
10		1	0	0

Table VI displays the search results for documents with the top 10 ranking for QE Term 2. There are 5 documents with the term "*ibadah*" but without the term in the

initial query. This is called alteration focus [15], where search results with terms obtained from query expansion may have a more dominant ranking than search results using terms in the initial query. This alteration focus causes the average precision score to decrease drastically and can affect the overall MAP scope.

From the results of the analysis above, it can be concluded that the more terms added will cause a change in focus. However, this alteration focus is not universally applicable and depends on each query. This is because there are still queries that still have an increase in the average precision score when the number of QE terms is 2.

3) Variable k1 BM25

The limit for testing from the value of k1 is 0 to 3 with a default value of k1 of 1.2. The k1 sizes used in this experiment are 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. This is due to the need for research for smaller and larger numbers compared to the default value of k1 size, which is 1.2. This experiment uses a hyperparameter variable from the most optimal value in the previous experiment. The results of k1 hyperparameter tuning can be seen in Table VII.

TABLE VII. HYPERPARAMETER TEST RESULTS K1 BM25

Query Term	Window	k1	b	Total QE Term	MAP
1	7	1.0	0.75	1	0.59114
1	7	1.2	0.75	1	0.62924
1	7	1.4	0.75	1	0.61014
1	7	1.6	0.75	1	0.64617
1	**7**	**1.8**	**0.75**	**1**	**0.65903**
1	7	2.0	0.75	1	0.65317

For each change in the k1 variable, it will give a different BM25 score for each document unit. Different BM25 scores will result in a shift in the ranking order of documents. In previous studies, it was concluded that the most optimal value was obtained at a distance between 0.5 to 2.0, with an increase of 0.1 to 0.2 [16]. With this specific corpus, the most optimal k1 value is 1.8. At 1.8, the value of k1 is relative to a larger number. This means that the greater the value of k1, the increase in the addition of scores for TF will be relatively slower.

4) Variable b BM25

The size limit for hyperparameter b testing is 0 to 1 with b defaulting to 0.75. The b sizes used in this experiment were 0.65, 0.75, 0.85, and 0.95. In this experiment, other hyperparameter variables were taken from the most optimal value in the previous experiment. The experimental results can be seen in Table VIII.

TABLE VIII. HYPERPARAMETER TEST RESULTS B BM25

Query Term	Window	k1	b	Total QE Term	MAP
1	7	1.8	0.65	1	0.58353
1	7	1.8	0.75	1	0.65903
1	**7**	**1.8**	**0.85**	**1**	**0.66556**
1	7	1.8	0.95	1	0.65857

This b value controls how much influence the length of the document has on the calculation of the BM25 score. The position of the variable b in the formula for equation 4 is as a divisor in the overall formula for BM25, namely in equation 9. This results in the greater the value of b, the lower the score. This change in the normalization results will slightly change the ranking order in the search results which affects the MAP scope for each different b value.

In previous studies, it was concluded that the best value was obtained at a distance between 0.3 to 0.9, with an increaseof 0.1. And based on empirical data, it was concluded that for the specific corpus of translations of the Qur'an in the Indonesian language version of the Ministry of Religion, the most optimal b value was 0.85. This indicates that only 0.85 or a weight of 85% of the length of the document is used in calculating the BM25 score.

B. Comparative Analysis of Search System Performance

After obtaining the most optimal model performance by tuning hyperparameters, in this section, a comparison test will be conducted to evaluate the proposed search system. In a previous study, it was concluded that the performance of BM25 can be decreased by automatic query expansion or semantic query expansion using word embeddings. One of the factors is that when users use a search system, search results that are different from the initial query can cause users to be confused. This automatic query expansion technique is considered less stable and can be overcome with Interactive Query Refinement [11] so that users can choose to use a search system with automatic query expansion or not.

To compare system performance, a comparison of the MAP score with the previous search system found in the Ministry of Religion's Qur'an will be carried out. Testing is done by varying the number of terms in the query. Table IX is the result of the comparison test of the search system.

TABLE IX. SEARCH SYSTEM MODEL PERFORMANCE COMPARISON RESULTS

Total Query Term	Searching System	MAP
1	Qur'an Kemenag	0.53718
	BM25 + SQE	0.66556
2	Qur'an Kemenag	0.98319
	BM25 + SQE	0.75185

TABLE X. SEARCH RESULTS ON THE PROPOSED METHOD FOR 1 QUERY TERM

Document Rank	Ayah Number	Detected Term	Relevance
1	108:002	berkurbanlah	Yes
2	005:027	kurban	Yes
3	**005:090**	**berkurban**	**No**
4	022:037	kurban	Yes
5	022:034	kurban	Yes
6	005:002	kurban	No
7	003:183	kurban	No
8	048:025	kurban	Yes
9	**002:196**	**berkurban**	**No**

TABLE XI. SEARCH RESULTS ON THE MINISTRY OF RELIGION'SQUR'AN FOR 1 QUERY TERM

Document Rank	Ayah Number	Detected Term	Relevance
1	**002:196**	**berkurban**	**No**
2	**005:090**	**berkurban**	**No**

In the query set that has only 1 term, the proposed search system method has increased the MAP score. This is because

the search system in the Ministry of Religion's Qur'an has no capabilities of stemming its query, resulting in not showing other relevant documents with different form of term. Table X is the search results on the proposed search system. Table XI is the search results in the Ministry of Religion's Qur'an search system. The line in bold is the intersection of the paragraphs found in the two search results. The proposed method still finds documents that have other forms of words from "*berkurban*", such as "*kurban*", and "*berkurbanlah*". While the Ministry of Religion's search engine ignores other documents that have the word "*berkurban*" in another form.

On the other hand, the comparison score of the search system performance for query sets that have 2 terms indicates that the Ministry of Religion's Qur'an search system is superior. This can be analyzed from the search results of one of the queries, namely "*kafarat berpuasa*". Table XII. is the search results on the proposed search system. Table XIII. is the search results in the Ministry of Religion's Qur'an search system. The line in bold is the intersection of the paragraphs found in the two search results.

TABLE XII. THE SEARCH RESULTS ON PROPOSED METHODS FOR 2 QUERY TERMS

Document Rank	Ayah Number	Detected Term	Relevance
1	**005:095**	**kafarat, berpuasa**	**Yes**
2	002:183	berpuasa	No
3	002:196	berfidyah, berpuasa	Yes
4	019:026	berpuasa	No
5	066:005	berpuasa	No
6	002:185	berpuasalah, berpuasa, menggantinya	Yes
7	002:187	puasa	No
8	058:004	tidak mampu, berpuasa	Yes
9	033:035	berpuasa	No
10	**005:089**	**kafarat, kafaratnya, berpuasa**	**Yes**

TABLE XIII. THE SEARCH RESULTS ON THE MINISTRY OF RELIGION'S QUR'AN FOR 2 QUERY TERMS

Ranking dokumen	Nomor Ayat	Detected Term	Relevan
1	**005:095**	**kafarat,berpuasa**	**Yes**
2	**005:089**	**kafarat, kafaratnya, berpuasa**	**Yes**

In Table XII. several relevant paragraphs are not listed in Table XIII. These paragraphs are on rank 3, 6, and 8. This is because the detected term in the document does not have a word that is similar to the initial query, namely "*kafarat*" and "*berpuasa*". However, the relevant paragraphs in Table XII. contain terms that have a semantic relationship with the initial query. These terms appear in other forms. The data is in Table XII. shows that in the 3rd ranking document there is the word "*berfidyah*", the 6th ranking document has the word "*menggantinya*", and the 8th ranking document has the words "*tidak mampu*" and "*berpuasa*". In this case, the proposed search system can detect relevant clauses that have other terms that still have a semantic relationship with the initial query.

The data in Table XIII. or the Ministry of Religion's Qur'an search system is only able to retrieve the same document as the initial query. In this case, the MAP score in the Ministry of Religion's Qur'an search system is relatively superior. However, in terms of the number of relevant documents, the search system with the proposed method obtains more relevant data, which is 5 relevant data compared to 2 relevant data.

V. CONCLUSION AND FUTURE WORK

Our proposed method could identify and retrieve verses that contains similar context and more relevant document in quantity than the previous method used in the original Ministry of Religion's app. Based on existing empirical data, the search system proposed in this study is relatively better to the search system in the Ministry of Religion's Qur'an if the query length is only 1 term and with query limits that have been determined at the beginning of the study. More term in a query means more term added from the query expansion, and that will cause alteration focus to the search result. To improve the accuracy of the search system performance comparison score, further research is needed with more and more complex query constraints and different approach on the query expansion calculation method to prevent alteration focus.

ACKNOWLEDGMENT

The Authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, and A. Burgun A, "Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse" Journal of the American Medical Informatics Association, 24(3), 607-613, 2017.

[2] K. S. Jones, S. Walker and S. Robertson, "A probabilistic model of information retrieval: development and comparative experiments," Information Processing and Management 36, pp. 809-840, 2000.

[3] S. Patel, P. N. Bhatt and P. C. Shah, "Query Expansion for Effective Retrieval from Microblog," in IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017.

[4] Z. Zhu, "A Step-by-step Tutorial for Conducting Sentiment Analysis," 2020. [Online]. Available: https://towardsdatascience.com/a-step-by- step-tutorial-for-conducting-sentiment-analysis-a7190a444366.

[5] M. Anandarajan, C. Hill, and T. Nolan, T, Text preprocessing, Practical Text Analytics (pp. 45-59). Springer, 2019.

[6] S. Sharma, "Understanding the Search Query - Part I," 2019. [Online]. Available: https://towardsdatascience.com/understanding-the-search- query-part-i-632d1b323b50.

[7] Y. Qiu and H. Frei, "Concept Based Query Expansion," in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993.

[8] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.

[9] R. Kulshrestha, "NLP 101: Word2Vec — Skip-gram and CBOW," 2019. [Online]. Available: https://towardsdatascience.com/nlp-101-word2vec- skip-gram-and-cbow-93512ee24314.

[10] N. Yusuf, M. A. M. Yunus, N. Wahid, N. Wahid, N. M. Nawi and N. A. Samsudin, "Enhancing Query Expansion Method Using Word Embedding," IEEE 9th International Conference on System Engineering and Technology, pp. 232-235, 2019.

[11] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval - The concepts and technology behind search. Second Edition, New York: Pearson Education Limited, 2011.

[12] A. I. Kadhim, "Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-ID", 2019 international conference on advanced science and engineering (ICOASE) (pp. 124-

128), 2019.

[13] A. Hast, "Consensus Ranking for Increasing Mean Average Precision in Keyword Spotting", 2nd International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, Vol. 2602, pp. 46-57, 2020.

[14] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker and P. Williams, "Okapi at TREC-5," 1997.

[17] D. Yeke, "Improving Document Ranking With Query Expansion Based On BERT Word Embeddings," 2020.

[18] C. Carpineto and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval," ACM Computer Survey, p. 50, 2012.

[19] B. P. Statistik, "Penduduk Menurut Wilayah dan Agama yang Dianut," 2010. [Online]. Available: https://sp2010.bps.go.id/index.php/site/tabel?tid=321&wid=0.

[15] M. Mitra, A. Singhal and C. Buckley, "Improving Automatic Query Expansion," SIGIR, pp. 206-214, 1998.

[16] C. Macdonald, N. Tonellotto, S. MacAvaney, and I. Ounis, "PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval", In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 4526-4533), 2021.

Multi-Label Classification for Scientific Conference Activities Information Text Using Extreme Gradient Boost (XGBoost) Method

Cryssa A. E. Piter[a1], Setiwan Hadi[a2], Intan Nurma Yulita[ab3]
[a]*Department of Computer Science, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
Bandung, Indonesia
e-mail: [1] cryssa17001@mail.unpad.ac.id, [2]setiawan.hadi@unpad.ac.id, [3]intan.nurma@unpad.ac.id

Abstract— **One of the government's aims is to improve the quality of education that requires lecturers to conduct scientific research. This scientific research must be disseminated later, forexample through scientific conferences. One important aspect ofsubmitting scientific articles is field suitability. A large number of scientific conferences is accompanied by a large number of fields. This of course will make it difficult for researchers to find suitable scientific conferences. Therefore, to assist in the selection, an automatic grouping of information on scientific conference activities was carried out. This study aims to classifythe information data on scientific conferences in Indonesia witha total of 1005 data. The data that has been collected will be classified into three labels, namely Economics, Science and Engineering, and Social Studies. This classification is multi-label, meaning that each data can have more than one label. Thisstudy uses the Extreme Gradient Boost (XGBoost) method. Testing of the hyperparameters is carried out to get the optimal model. This test produces a Word2Vec hyperparameter with a dimension of 100 and a window size of 15. For the XGBoost hyperparameter, optimal conditions are obtained when estimators 100, learning rate 0,1, maximum depth 6, minimum child weight 10, and gamma 3,6. This model was then evaluatedusing the K-fold Cross-Validation which resulted in an average hamming score of 79.52% and an f1 score of 85.88%.**

Keywords— Multi-Label Classification, Extreme Gradient Boost, Word2Vec, Scientific Conference

I. INTRODUCTION

In the current era of technology, science is developing rapidly. Many studies have been carried out that lead to new discoveries in almost every field of education. The challenge faced in this era is how to keep up with technological developments to improve the quality of education. One of the government's efforts to improve the quality of education in Indonesia is to require lecturers to conduct scientific research. This is regulated in Permenristek Dikti No. 44 of 2015 concerning National Standards for Higher Education. The aim is to produce quality human resources, especially graduates of higher education, who have high knowledge and skills, also master technology. The results of the research must later be disseminated to the public, one of the ways is through scientific conferences. Scientific conferences are held as a place for researchers to share their scientific findings. In this activity, presentations were made to get criticism and input, as well as scientific development through collaboration with other researchers.

Submitting a scientific article takes into account several aspects, one of which is the suitability of the topic or field. A large number of scientific conferences is certainly balanced by the number of fields that exist. These fields should be grouped to assist researchers in finding conferences with appropriate fields. One way to group data is to use tags. Tagging can be defined as the process of assigning short textual descriptions or keywords (tags) to information objects [1]. Multi-label classification is a classification in which several labels can be assigned to each data. In a multilabel problem, there is no limit on how many classes can be assigned [1]. In the past, multi-label classification was mainly used for text grouping and medical diagnosis. For example in a medical diagnosis, a patient may have diabetes and prostate cancer at the same time.

In performing multilabel classification there are several methods that can be used, one of which is Extreme Gradient Boost (XGBoost). Since its introduction in 2014, XGBoost has become one of the most popular methods used in Kaggle competitions. Besides being known for its good performance, XGBoost is also known for its flexibility and speed [2]. Therefore, this study uses the XGBoost method. With the grouping of conference fields, it is hoped that it can assist researchers in finding conference activities according to the desired field. One of the purposes of using multi-label classification is to study and find relationships between these fields so that they can be grouped automatically. To realize this solution, the author proposes making a thesis to create a multilabel classification system for informational texts on scientific conference activities using the Extreme Gradient Boost (XGBoost) method.

II. LITERATURE REVIEW

In section II, the results of the literature study that have been carried out for the theoretical foundations used in the research will be explained. These theories include machine learning, text mining, multi-label classification, XGBoost, cross-validation, confusion matrix, hamming score, and unified modeling language.

A. Machine Learning

Machine learning is an implementation of artificial intelligence that provides the ability for the system to learn and improve its capabilities automatically without having to be programmed directly. Its main purpose is to allow computers to learn automatically without human intervention or other assistance [3].

B. Text Mining

Text mining can be defined as a process to find and extract information from unstructured text data [4]. The goal is to find a useful pattern. Information obtained based on these patterns and relationships is used to reveal facts, trends, or models.

1. Text Pre-processing

One of the stages in classifying text is pre-processing. Pre-processing is one of the important components in text mining

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

algorithms. The pre-processing step usually consists of case folding, tokenizing, stop-word removal, and stemming [5].

a. Case Folding

In text pre-processing, the case folding process aims to convert all letters in a text document into lowercase letters. In addition, other characters will be omitted and considered as delimiters.

b. Tokenization

The tokenization stage or also known as parsing is the stage to break the sequence of characters into several partsof words or phrases called tokens. At the same time also remove punctuation. This token helps in understanding the text by analyzing the word order.

c. Stop-word Removal

Stop-word removal is usually performed on documents todelete some words. Stop-words are words that appear in the text but are considered unimportant, such as prepositions, conjunctions, and others. The goal is not to increase the number of features so as to reduce the performance of the model in performing text.

d. Stemming

Stemming is the process of mapping and breaking shapeswords into their root form. This method aims to get the roots of derived words.

2. Term Frequency – Inverse Document Frequency (TFIDF)

One of the stages in classifying text is pre-processing. Pre-processing is one of the important components in text mining algorithms. The pre-processing step usually consists of case folding, tokenizing, stop-word removal, and stemming [6].

3. Word Vector Representation

To process data in the form of text in machine learning, a text extraction process is needed into a numeric vector to represent each word. This process is known as word embedding. One model of word embedding is Word2Vec. Word2Vec is a neural network that processes text data. In Word2Vec, initially, words are represented in the form of a matrix, then converted into an N-dimensional vector [7]. Word2Vec will first build vocabulary from the corpus of training texts and learn the vector representation of each word. In addition, Word2Vec has the ability to calculate the distance between each word. Then, the words will be grouped by distance.

C. Multi-Label Classification

Multi-label classification is a grouping where multiple labels can be assigned to instances in the dataset. That is, an instance can belong to more than one class. Examples of multilabel classification applications include article and website classification, music categorization, genomic functionality discovery, and others [8]. The multilabel classification method can be divided into two groups, namely the Transformation Problem (TP) method, and the Adaptation Approach (AA) method. The TP method group does not depend on the algorithm. This method converts a multilabel classification task into one or more single-label classification, regression, or rating labels. The second group of methods, the Adaptation Approach, extends algorithm-specific learning to handle multilabel data directly.

D. Extreme Gradient Boosting

Extreme Gradient Boosting, also known as XGBoost, is adecision-tree-based algorithm that uses a gradient boosting framework to solve machine learning problems. Gradient boosting is a technique that creates a new model to predict the error of the previous model. This causes the error in the new model to be smaller. XGBoost and gradient Boosting share the same principle. The main difference between the two lies in the implementation details. XGBoost achieves better performance by controlling tree complexity using regularization techniques [2].

Hyperparameters have an important role in helping the model learning process. Hyperparameters are important because they directly affect the training algorithm and have asignificant impact on the performance of the trained model. The following are XGBoost hyperparameters used in this study:

a. Learning_rate

The learning rate, also known as eta, is used to help the model prevent overfitting. After each boosting step, a newfeature weight is obtained, and eta shrinks the feature weight to make the boosting process more conservative. Eta has a range of values from 0 to 1.

b. Max_depth

Maximum depth determines the maximum depth of the tree. Increasing this value will make the model more complex and allow for overfitting. Max depth has a rangeof values from 0 to infinity.

c. Min_child_weight

Minimum child weight is the minimum weight that a nodehas. If tree partitioning results in nodes with a total weightless than min_child_weight, the build process stops further partitioning. This hyperparameter has a value ranging from 0 to infinity.

d. Gamma

Gamma determines the pruning of nodes from the constructed tree. The larger the gamma value, the more conservative the algorithm. This hyperparameter has a value ranging from 0 to infinity.

E. Cross-Validation

Cross-validation is a data sampling method to assess the generalizability of predictive models and prevent overfitting [9]. The purpose of this method is to provide an estimate of the model's performance on the new data. This method has a single parameter, k, which refers to the number of data sets to be broken down into certain data samples. Therefore, this method is often referred to as k-fold cross-validation. 10-fold cross-validation means the data set is randomly divided into ten separate subsets. The model is trained in the training set and then applied to the validation set.

The confusion matrix is a way to evaluate errors in classification problems [10]. The confusion matrix provides a summary of the comparison of the predicted results given by the model and the actual results. For multilabel cases, a confusion matrix will be presented for each label sequentially. The confusion matrix has several performance metrics to measure the performance of the model that has been created. Metrics that are often used are:

a. Precision is the ratio of correct predictions for positivedata compared to overall positive predicted results

b. Recall is the ratio of correct predictions for positive data compared to all data that are predicted to be positive.

c. Accuracy is the ratio of correct prediction, either positive or negative, to the overall data.

For multi-label cases, the metrics use sample-based evaluation. Where labels are evaluated first and then averaged for all labels.

F. Hamming Score

In single-label and multi-class classification problems, accuracy is one of the most commonly used evaluation methods. However, in multi-label classification, the prediction for an instance is a set of labels. Thus, the prediction can be completely correct, partially correct, or completely false. To find out which instances are partially correct, one way is to evaluate the average difference between the predicted label and the actual label for each test instance, then the average for all examples in the test set [11]. For each document d_j , accuracy is measured using a hamming score which symmetrically measures how close T is to S [12].

III. METHODS

A. Data Collection

The data collected is the Indonesian scientific conferences activities on the internet. Data were collected from easychair.org, wikicfp.com, call4paper.com, and allconferencealert.com sites. The data used in building the research model amounted to 1,005 scientific conference data. Fig. 1 shows the amount of data per label that was collected.

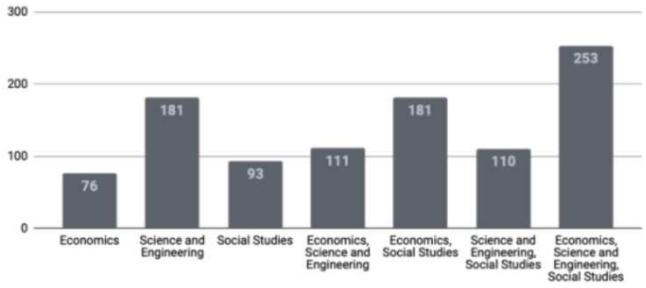


Fig. 2. Number of Data per Label

B. Data Preprocessing

The data that has been collected will go through the pre-processing stage. The goal is to make the data more structured because the existing data still have affixes, abbreviations, or symbols. This stage is very important because it will affect thequality of the data to be classified. The pre-processing stage carried out is shown in Fig. 2.

C. TF-IDF Weighting

TF-IDF is a technique that measures how important a word is in a document. TF (Term Frequency) measures the frequency of words in a document. IDF (Inverse Document Frequency) measures the ranking of certain words for their relevance in the text. This research uses the TfidfVectorizer library.

The main concept of TF-IDF is if a word or phrase appears in a text document with high-frequency TF and is rarely found in other documents. Thus, the word or phrase is considered to have good class differentiation abilities in classifying [6].

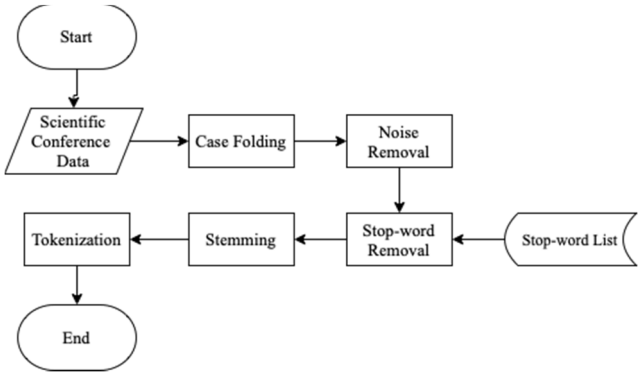


Fig. 2. Data Preprocessing Stage

B. Word Vector Representation

The words that have gone through the preprocessing stage will then be converted into vector form. Each word has a vector that represents its meaning using word2vec. The formation of the word2vec model using the gensim library in python. This word2vec model uses the Skip-Gram architecture, which predicts the output around the input. This input will be converted into one-hot encoded vectors.

D. Multi-Label Transformation

At this stage, the data will be separated for each class as a single-label as many as the number of classes. The data is labeled according to the target. The data will be assigned a value of true if it matches the target and false otherwise. The multilabel transformation in this study uses the Scikit-learn library, namely the MultiLabel Binarizer.

E. XGBoost Model

At this stage, a model will be created using the XGBoost method. For the multilabel case, the model will be defined using the MultiOutputClassifier class from the Scikit-learn API. This class helps classify multi-output data.

IV. RESULT

A. Hyperparameter Test Analysis

To get the optimal model, hyperparameter has an important role. For this reason, several tests were carried out on the hyperparameters that would be used. Tests were carried out on the dimensions and window size of Word2Vec, as well as min_child_weight and gamma on XGBoost. The goal is to find the best precision, recall, f1 score, and hamming score.

1. Word2Vec

In building the Word2Vec model, there are two hyperparameters tested, namely dimension size and window size. The test is carried out using other hyperparameters that have constant values, namely, n_estimators of 100, max_depth of 6, and learning_rate of 0.1.

a. Dimension Size

The dimensions of the vectors used in this test are 50, 100, 150, and 200. The results show a significant difference in hamming scores. The smaller the dimensions used, the more information is discarded. The use of vector dimensions that are too large can also cause the appearance of additional vectors that are not representative. The test results can be seen in Table 1. When the experiment was carried out with a dimension of 50, the Hamming score had a low value. Meanwhile, when the dimensions are 150 and 200, the Hamming score has decreased. Thus, the dimension of the vector with a size of 100 gives optimal results in representing words.

b. Window Size

The window sizes used in this test are 2, 5, 10, and 15. Large window sizes tend to capture more topic information, while smaller window sizes tend to capture more about the word itself. When the window size is 20, the hamming score has decreased compared to window sizes 5 and 10. For window size 2, it has the lowest hamming score. Based on the test results in Table 2, optimal results are obtained when the window size is 15.

TABLE I. DIMENSION SIZE HYPERPARAMETER TEST RESULTS

Dimension Size	Precision (%)	Recall (%)	F1 Score (%)	Hamming Score (%)
50	87,11	88,54	84,38	76,90
100	88,31	88,44	85,22	78,25
150	88,10	87,86	84,68	77,26
200	87,86	87,73	84,40	77,02

TABLE II. WINDOW SIZE HYPERPARAMETER TEST RESULTS

Window Size	Precision (%)	Recall (%)	F1 Score (%)	Hamming Score (%)
2	87,64	87,27	83,97	76,47
5	88,37	88,28	85,17	78,02
10	88,42	88,83	85,50	78,59
15	88,37	89,84	85,99	79,33
20	88,32	88,23	85,03	77,93

2. XGBoost

In building the XGBoost model, there are two hyperparameters to be tested, namely min child weight and gamma. The test is carried out with other hyperparameters with constant values, namely the dimension size of 100, window size of 15, n_estimators of 100, max_depth of 6, and learning_rate of 0.1.

a. Minimum Child Weight

The min child weight sizes used in this test are 6, 8, 10, and 12. These sizes were chosen because a min child weight value that is too low causes the model to be overfitting, while a value that is too high causes the model to be under fitted. In this hyperparameter experiment, the min child weight of 6 has a fairly good hamming score, but the model is overfitting. Meanwhile, the value of min child weight 12 experienced a decrease in the hamming score. Thus, the optimal model is obtained when min child weight is 10. This can be seen in Table 3.

b. Gamma

The sizes of the gamma used in this study are 3,2, 3,4, 3,6, and 3,8. These sizes were chosen because after experimenting with other sizes, the gamma value was too low, causing the model to be overfitting and too high to be causing the model to be underfitting. A gamma size below 3 causes an overfitting model, while a gamma size above 4 causes an underfitting model. Gamma is worth 3.2 and 3.4 has a low Hamming score. Gamma value of 3.6 produces the most optimal model. This model has the highest Hamming score and does not experience overfitting or underfitting. Meanwhile, when gamma is 3.8, the Hamming score has decreased. The results of this experiment can be seen in Table 4.

TABLE III. MINIMUM CHILD WEIGHT HYPERPARAMETER TEST RESULTS

Minimum Child Weight	Precision (%)	Recall (%)	F1 Score (%)	Hamming Score (%)
6	88,27	89,05	85,37	78,34
8	88,51	88,56	85,31	78,23
10	88,56	88,96	85,51	78,47
12	88,20	88,84	85,24	78,19

TABLE IV. GAMMA HYPERPARAMETER TEST RESULTS

Gamma	Precision (%)	Recall (%)	F1 Score (%)	Hamming Score (%)
3,2	87,93	89,21	85,35	78,33
3,4	88,36	89,25	85,62	78,76
3,6	88,71	89,72	86,14	79,52
3,8	88,62	89,37	85,88	79,12

B. Evaluation

Evaluation is done by using the confusion matrix. The evaluation was carried out for each label, namely "Economics", "Science and Engineering", and "Social Studies". The evaluation results are shown in Table 5.

TABLE V. CONFUSION MATRIX RESULTS

Real Result	Prediction Result (%)					
	Economics		Science and Engineering		Social Studies	
	(-)	(+)	(-)	(+)	(-)	(+)
(-)	29,1	9,3	27,8	7,2	25,9	10,9
(+)	8,6	53,5	8,1	57,4	8,4	55,3

The result of this confusion matrix is the average for each fold in the 10-fold cross-validation. For the Economics label, the model managed to correctly predict negative data by 29.1% and positive data by 53.5%. In addition, the model also predicts incorrectly negative data by 9.3% and positive data by 8.6%. For the Science and Engineering label, the model managed to correctly predict negative data by 27.8% and

positive data by 57.4%. The model also predicts wrong negative data by 7.2% and positive data by 8.1%. As for the Social Studies label, the model managed to correctly predict 25.9% of negative data and 55.3% of positive data. Furthermore, the model predicts incorrectly negative data by 10.9% and positive data by 8.4%.

V. CONCLUSION AND FUTURE WORK

Based on the research that has been carried out, the implementation of XGBoost in conducting multi-label classification of conference activity information texts gives the best result with the optimal Word2Vec model. However, ourresearch is still lack data and label used. By adding new sources to obtain information on scientific conference activities and new labels, the application could be made more widely used. This research resulted in a hamming score of 79.52% and an f1 score of 86.14%. 2. To improve its accuracy, further research could use feature selection to eliminate features that do not affect the model.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] I. Katakis, G. Tsoumakas, and I. Vlahavas, Multilabel text classification for automated tag suggestion. Proceedings of the ECML/PKDD, 2008.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[3] S. Raschka, Python Machine Learning. Packt publishing ltd, 2015.

[4] M. Allahyari *et al.*, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv [cs.CL]*, 2017.

[5] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving text preprocessing for student complaint documentclassification using sastrawi," *IOP* Conference Series: Materials Science and Engineering, vol. 874, p. 012017, 2020.

[6] H. Fan and Y. Qin, "Research on text classification based on improved TF-IDF algorithm," in Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering *(NCCE 2018)*, 2018.

[7] A. H. Ombabi, O. Lazzez, W. Ouarda, and A. M. Alimi, "Deep learning framework based on Word2Vec and CNNfor users interests classification," in 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT*)*, 2017, pp. 1–7.

[8] C. Tawiah and V. Sheng, "Empirical comparison of multi-label classification algorithms," Proc. Conf. AAAI Artif. Intell., vol. 27, no. 1, pp. 1645–1646, 2013.

[9] D. Berrar, "Cross-Validation," 2019.

[10] E. Beauxis-Aussalet and L. Hardman, Visualization of confusion matrix for non-expert users. IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings, 2014.

[11] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Oregon State University, Corvallis, vol. 18, pp. 1–25, 2010.

[12] S. Godbole and S. Sarawagi, "Discriminative Methods for Multi-labeled Classification," in Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 22–30.

Prediction of Stock Price Data of PT. Ramayana Lestari Sentosa Tbk. using Long Short Term Memory Model

Luki Setiawan[ab1], Nurul Fathanah Muntasir [ac2], Syafa Fahreza [ad3], Asep Sholahuddin [ae4]
[a]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[b]*Department of Mathematics*, *Universitas Padjadjaran*
Sumedang, Indonesia
[c]*Department of Electrical Engineering, Hasanuddin University*
Makassar, Indonesia
[d]*Department of Computer Science, Telkom University*
Bandung, Indonesia
[e]*Department of Computer Science, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]luki18001@mail.unpad.ac.id, [2]nurulmuntasirr@gmail.com , [3]syafafahreza7@gmail.com , [4]asep.sholahuddin@unpad.ac.id

***Abstract*— The investment that Indonesian people are interested in during the pandemic is investing in stocks. However, stock investment will not only generate profits but also risk losses so that to minimize the risk of loss, forecasting is required. One of the prediction methods that can be used is Long Short Term Memory (LSTM) for stock data modeling, this is because LSTM is suitable for predicting and processing important events with relatively long intervals and delays in the time series data. The data used in this research is time series data of PT. Ramayana Lestari Sentosa Tbk. in the period 2 March 2020 to 15 September 2021 with LSTM model using the same number of epochs but different number of nodes, activation functions and optimizers which show varying results and accuracy levels. The selection of the number of epochs 5 is the optimal number of epochs selected. In the experiment, the researchers use 2 types of nodes (4 and 16), 4 types of activation functions (Linear, Sigmoid, Relu, and Tanh), and 3 types of optimizers (Adam, SGD, and Adagrad). The results of the experiment show that the best LSTM model based on the smallest RMSE test value is 15.37 on the use of LSTM model with 4 nodes, Linear activation function, and Adam optimizer. Meanwhile, the stock price prediction for the next day (16 September 2021) obtained based on LSTM model is IDR 639.05. Based on this, it can be concluded that in predicting RALS stock price data using LSTM model with the number of nodes 4 is better than the number of nodes 16. Linear activation function is also better than the activation function of Sigmoid, Relu and Tanh. Then, Adam optimizer is better than SGD and Adagrad optimizer.**

Keywords—LSTM, time series, Prediction, Stocks, RALS.

I. INTRODUCTION

The Covid-19 pandemic in Indonesia was first discovered around early or mid-March 2020 and is not expected to affect the stock market, but seeing the increasing number of confirmed victims, the stock market reacted negatively [5]. In fact, the Covid-19 outbreak has affected the business and investment environment [1]. Shares are a sign of ownership of individual investors, institutional investors or traders of a number of funds invested in a company [2]. As for the closing price of stock data can be found on the website yahoo.finance.com or on registered securities. Of course, investing in stocks can only lead to profits but also losses. For this reason, stock price predictions are urgently needed in minimizing the risk of loss and maximizing the profits for stock investors.

One of the stock markets affected by Covid-19 is PT Ramayana Lestari Sentosa Tbk (RALS). Based on the financial reports submitted to the Indonesia Stock Exchange (IDX), RALS was recorded with an income of IDR 2.52 trillion in 2020. Then, the realization of RALS' income has decreased by 54.8 percent or IDR 5.59 trillion since the 2019 period. Therefore, it is very necessary to predict the price of RALS shares.

Stock prices are included in the category of periodic data (time series). In solving this time series data problem, the basic steps taken are to collect and select the right variables and then choose the best model that provides the best accuracy [3]. There are several ways to model time series data such as multivariate models, linear models and nonlinear models. In stock market analysis there are very large and non-linear data, in dealing with this diversity of data, an efficient model is needed that can identify complex relationships and hidden patterns in this large data set [8]. From several test results in research that has been carried out in the range of 2017-2019, LSTM can predict stock prices with good performance and relatively small error rates. LSTM introduces the memory cell, gate structure, which has been proved to be able to effectively associate memories and input remote in time [11]. LSTM has a selective memory feature on a long-term scale, which is suitable for predicting and processing important events with relatively long intervals and delays in the time series data. [4]. In this study, LSTM method will be applied in forecasting RALS stock prices in the initial period of the pandemic around 2 March 2020 to 15 September 2021.

II. LITERATURE REVIEW

A. Long Short Time Memory (LSTM)

Long Short Term Memory (LSTM) is designed to solve problems in Recurrent Neural Network (RNN). This problem occurs because the RNN may lose important information

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

obtained at the beginning if the sequence is long enough (forward propagation). In addition, Long Short-Term Memory (LSTM) can also be used to overcome vanishing gradients or, in other words, gradients have very small values and are close to 0, even in extreme cases, this can cause a stop condition when running (backward propagation) [7]. Long Short-Term Memory (LSTM) algorithm was discovered in 1997 and used for the development of the Recurrent Neural Network (RNN) [7], Long Short-Term Memory.

LSTM has a more complex structure than RNN. LSTM consists of cell state and 3 gates. Cell State has a function to store information on the network. Forget gate determines which information will be forwarded / used because it is important or will be forgotten. A value close to 0 means it will be forgotten while a value close to 1 will be used. The input gate will determine the information that will be used in this step by how it works is to update the state column replaced with the hidden state and the current input. The output gate has a function to determine the hidden state in the next step. The hidden state contains information from the previous input. The hidden state is also used to make predictions. In the first step, Long Short-Term Memory (LSTM) that will be used is "forget gate". Forget gate is used to determine the information in "cell state" (C_{t-1}). Inside the forget gate, it will process S_{t-1} and x_t. In the f_t formula, "sigmoid" activation is used, meaning that it will produce a number from 0 to 1. If the value is close to 1, the information will continue to be used/stored while if it is close to 0, the information will be deleted/forgotten. The input gate is to terminate the value and update the existing data. It consists of two main operations, to calculate the values of i_t and $\tilde{C}_t$. i_t operation will use a sigmoid, this means that it will produce a value from 0 to 1 while $\tilde{C}_t$ uses tanh which means that it will produce a value from -1 to 1. Then, cell state is additional memory owned by Long Short Term Memory (LSTM) and this is not owned by Recurrent Neural Network (RNN). In the formulas C_t , f_t dan i_t, this acts as a weight to determine the old information to be stored (C_{t-1}) and new information to be added ($\tilde{C}_t$). In this output gate step, the value of s_t will be determined. The output gate itself functions to control how many states pass through it and determine the hidden state in the next step.

B. Optimizer

Optimizer is an algorithm that is used to maximize the function of attributes such as the weight of the neural network and minimize the error rate. In the training phase of the algorithm, it will work continuously so that it can minimize the loss function and maximize the accuracy of the model. In this research, we used 3 optimizers, Adam, Stochastic gradient descent (SGD), and Adaptive gradient (AdaGrad).

1. *Adam*

   Adam is an optimizer that is very easy to implement because it only requires a small amount of memory and is efficient. The most important thing is that Adam is very suitable for processing parameterized data [6].

2. *Stochastic gradient descent (SGD)*

   SGD stands for stochastic gradient descent which is also an optimizer. The way SGD works is by finding a new weight from all training data, then SGD will analyze each data [9]. The advantage of using SGD itself is that there is a terminalizer needed when looking for a new weight.

3. *Adaptive gradient (AdaGrad)*

   AdaGrad is an algorithm that is usually used to adjust the learning rate, so it can improve performance by using sparse gradients [12].

C. Evaluation

Root Mean Square Error (RMSE) is a method used to determine the level of accuracy of a model. RMSE itself calculates the root of the average number of squares of error values in the predicted model [13]. RMSE is a method that is easy to implement and has often been used in many cases related to prediction or forecasting [10]. Meanwhile, Mean Square Error (MSE) is the average number of squares of error values in the model which is used to calculate the error of the data sample and is not used for model estimation.

III. METHOD

A. FLOWCHART

This study will apply the Long Short-term Memory method with stages such as the flow chart in Fig. 1. The stages are data preprocessing, data split, training using the LSTM method, and evaluation of prediction results.

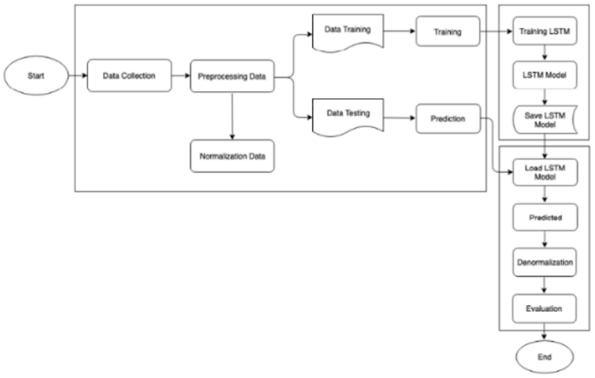


Fig. 1. Flowchart

B. Dataset

In this research, we obtained stock data that will be used from the *yahoo finance* website by taking stock data of PT. Ramayana Lestari Sentosa Tbk. (RALS.JK) in the period 2 March 2020 to 15 September 2021. From the data obtained, there are 372 rows of data with 4 columns that will be used in this study such as open, high, low, close columns. As for the explanation of these terms: open is the opening price of a stock in one trading day while; close is the last price or closing price before the stock market closes; then low and high is the daily price rate of the stock in a rational buy or sell position. In this study, the data used is from the "close" column which will then be used as time series data. The application used is Google Colab and the device specifications use a laptop with the Windows 10 operating system.

The stock price data was plotted based on 3 labels, OHLC avg, HLC avg and closing price. OHLC avg is the average of

the opening price, high price, low price, and closing price of the stock. HLC is the average of the highest price, lowest price, and closing price of the stock. Fig. 2 is the result of the plot of the stock price data of PT. Ramayana Lestari Sentosa Tbk. The x and y axes show the period and stock value.

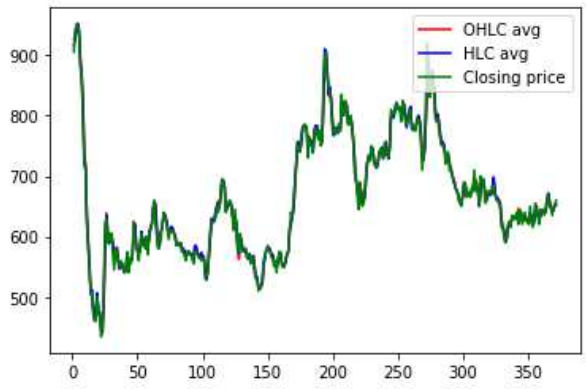


Fig. 2. Plot dataset of RALS

C. Research Stages

1. Data Preparation Stage

The first stage is collecting data sourced from yahoo finance, and obtained as many as 372 rows and 4 columns. Normalization of the data is performed so that the data is in the range [0,1]. The advantages of normalizing the data itself can change the numbers to be smaller but do not forget to maintain the pattern. Then the data will be further divided into 75% and 25%. 75% of the data will be used as training data and the other 25% will be used as testing data.

2. Long Short Term Memory (LSTM) Model Design Stage

LSTM model used in this study consists of 2 hidden layers with different combinations of nodes and activations. There are 2 types of nodes that will be used by researchers, nodes 4 and 16. On the other hand, the activation functions used by researchers are Linear, Sigmoid, Relu and Tanh.

3. Determination of Epoch Number

The epoch selection is performed in order to get a good accuracy value and the program can run in a faster time. To determine the most optimal epoch, an epoch plot of 1 to 20 is formed in one of the models with a certain node, activation function and optimizer. After that, the best 1 epoch will be selected.

4. Data Training and Data Prediction

We conduct training data using the number of epochs selected, as much as 5, and also use 3 different optimizers including Adam, SGD, and Adagrad. To visually measure the accuracy of the training data and predictive data against the actual data, a comparison plot is formed between the actual data and the predicted model data.

5. Denormalization

The results obtained are in the form of interval values, therefore we must return the data by denormalization. Denormalization itself is useful to make it easier for readers to know the output issued.

6. Evaluation (RMSE AND MSE)

RMSE in this study is a method used to determine the level of accuracy of a model as a whole. Meanwhile, MSE is used to calculate the data sample error at each epoch and is not used to estimate the overall model.

IV. Result and Discussion

A. Normalization Stock Price Data

Table 1 shows the differences between the stock price data before normalization and after normalization.

TABLE I. Table of Stock Price Data Before and After Normalization

Before Normalization	After Normalization
905	0,9127907
940	0,98062016
950	1
945	0,99031008
945	0,99031008
…	.
650	0,41860465
635	0,38953488
650	0,41860465
650	0,41860465
660	0,4379845

B. Determination of Epoch Number

In determining the number of epochs to be used in the training data, the researcher tried to iterate running the training data program with the LSTM model (4 nodes and linear activation function), with a different number of epochs to the loss value (MSE) obtained. The iteration is represented in Fig. 3. The number of epochs 5, as the most optimal epoch.

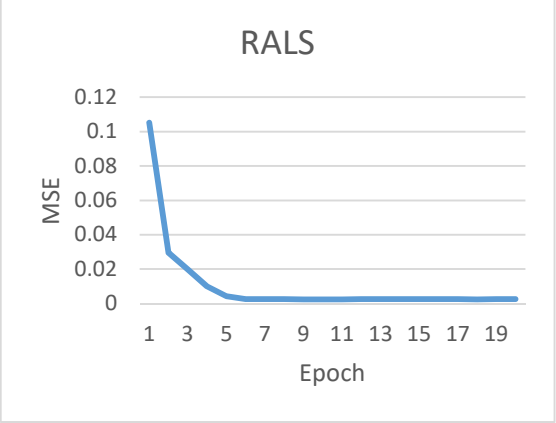
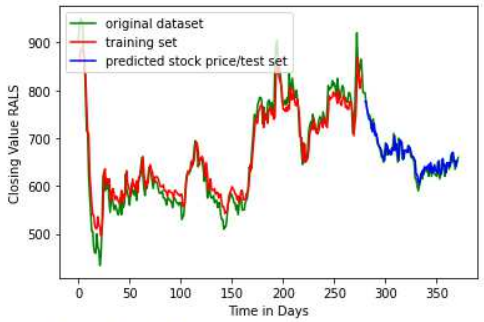


Fig. 3. Chart of *epoch* vs MSE

C. Data Training and Data Prediction

Fig. 4 shows an example of the denormalization and plot results of training data and test data using the LSTM node 4 model, linear activation function, and the Adam optimizer.




```
Last Day Value: 653.0597534179688
Next Day Value: 573.5641479492188
```

Fig. 4. Example of data training results and prediction

D. Table of Experimental Results

Table 2 is overall experimental results using 2 types of nodes, 4 types of activation functions and 3 types of optimizers.

TABLE II. TABLE OF EXPERIMENT

Node	Activation	Optimizer	Train RMSE	Test RMSE	Prediction
4	Linear	Adam	27,36	15,37	639,05
4	Linear	SGD	112,83	42,12	271,95
4	Linear	Adagrad	201,86	172,97	60,32
4	Sigmoid	Adam	27,45	17,32	551,74
4	Sigmoid	SGD	111,58	40,15	302,56
4	Sigmoid	Adagrad	113,79	48,08	347,51
4	ReLU	Adam	255,57	233,02	0
4	ReLU	SGD	108,68	40,19	287,35
4	ReLU	Adagrad	210.75	183,5	50,31
4	Tanh	Adam	35,14	15,47	509,66
4	Tanh	SGD	112,42	40,55	302,09
4	Tanh	Adagrad	222,69	194,81	28,14
16	Linear	Adam	26,35	15,91	674,62
16	Linear	SGD	112,47	40,28	290,91
16	Linear	Adagrad	192,34	162,06	74,14
16	Sigmoid	Adam	26,24	18,4	572,28
16	Sigmoid	SGD	111,69	40,2	302,88
16	Sigmoid	Adagrad	114,55	47,21	338,86
16	ReLU	Adam	255,57	233,02	0
16	ReLU	SGD	110	39,75	290,6
16	ReLU	Adagrad	255,57	233,02	0
16	Tanh	Adam	25,47	17,71	579,84
16	Tanh	SGD	111,26	39,93	291,22
16	Tanh	Adagrad	192,57	161,79	72,2

Based on Table 2, we obtain the lowest RMSE train and test RMSE values. The lowest RMSE train value is obtained when using the LSTM method with 16 nodes, Tanh activation function, and Adam optimizer. The lowest RMSE test value was obtained when using the LSTM method with 4 nodes, linear activation function, and Adam optimizer. To obtain the best prediction results, the researchers chose the method with the smallest RMSE test value (LSTM method with 4 nodes, linear activation function, and Adam optimizer). This is also supported visually in Fig. 6 which shows that the actual/original data plots almost coincide with the train and test data plots. This method produces a train RMSE value of 27.36, a test RMSE of 15.37 with a price prediction result per share one day ahead (16 September 2021) of IDR 639.05.

V. CONCLUSIONS AND SUGGESTIONS

The results of the prediction of the time series data for the stock price of PT. Ramayana Lestari Sentosa Tbk. in the period 2 March 2020 to 15 September 2021 using the LSTM model with the same number of epochs but different number of nodes, activation functions and optimizers show varying results and accuracy levels. The selection of the number of epochs 5 is the optimal number of epochs chosen by the researcher in order to get the best accuracy value and the program can run in a faster time. In the experiment, the researcher used 2 types of nodes (nodes 4 and 16), 4 types of activation functions (Linear, Sigmoid, Relu, and Tanh), and 3 types of optimizers (Adam, SGD, and Adagrad). This was performed by researchers in order to get results with the best accuracy.

After conducting the experiment, the best LSTM model was obtained based on the smallest RMSE test value of 15.37 on the use of the LSTM model with 4 nodes, Linear activation function, and Adam optimizer. The result of stock price prediction one day ahead (16 September 2021) obtained based on the LSTM model is IDR 639.05. This shows that to predict the stock price data of PT. Ramayana Lestari Sentosa Tbk. using the LSTM model with the number of nodes 4 is better than the number of nodes 16. Linear activation function is also better than the activation function of Sigmoid, Relu and Tanh. Then, Adam optimizer is better than SGD and Adagrad optimizer.

Suggestions for further research is to conduct comparative experiments with various different datasets so that more datasets are used. In addition, predictions can also use other deep learning algorithms such as GRU (Gated Recurrent Unit) or RNN (Recurrent Neural Network) according to the needs and performance effectiveness.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCE

[1] A. Al-Awadhi, K. Alsaifi, A. Al-Awadhi and S. Alhammadi, "Death and contagious infectious diseases: Impact of the COVID-19 virus on stock market returns", Journal of Behavioral and Experimental Finance, vol. 27, p. 100326, 2020.

[2] J. Clapp, "The rise of financial investment and common ownership in global agrifood firms". Review of International Political Economy, 26(4), 604-629, 2019.

[3] Z. Berradi, M. Lazaar, "Integration of Principal Component Analysis and Recurrent Neural Network to Forecast the Stock Price of Casablanca Stock Exchange," Procedia Computer Science, 148, 55–61.

[4] J. Du, Q. Liu, K. Chen and J. Wang, "Forecasting stock prices in two ways based on LSTM neural network", 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019.

[5] K. Khan, et al., "The Impact of COVID-19 Pandemic on Stock Markets : An Empirical Analysis of World Major Stock Indices," Journal of Asian Finance, Economics and Business, 7(7), pp. 463–474.

[6] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv: 1412.6980, 2017.

[7] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[8] M. Vijh, D. Chandola, V. Tikkiwal, and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", Procedia Computer Science, vol. 167, pp. 599-606, 2020.

[9] I. N Yulita, S. Purwani, R. Rosadi, and R. M Awangga, "A quantization of deep belief networks for long short-term memory in sleep stage detection", 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA) (pp. 1-5), 2017.

[10] N. R. Sari, W. F. Mahmudy, A. P. Wibawa, and E. Sonalitha, "Enabling external factors for inflation rate forecasting using fuzzy neural system," Int. J. Electr. Comput. Eng., vol. 7, no. 5, pp. 2746–2756, 2017.

[11] S. Yao, L. Luo, and H. Peng, "High-frequency stock trend forecast using lstm model," 2018 13th International Conference on Computer Science & Education (ICCSE), 2018.

[12] T. Tieleman, G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural networks for machine learning, *4*(2), 26-31, 2012.

[13] Y. Abdillah, Suharjito, "Failure prediction of e-banking application system using Adaptive Neuro Fuzzy Inference System (ANFIS)," Int. J. Electr. Comput. Eng., vol. 9, no. 1, pp. 667–675, 2019.

Design of Museum Historical Heritage Management System Using Blockchain Digital Certificate and Hyperledger Composer

Sachi Hongo[ab1], R. Sudrajat[ab2], Intan Nurma Yulita[ab3]
[a]*Department of Computer Science*
[b]*Research Center for Artificial Intelligence and Big Data*
Universitas Padjadjaran
Sumedang, Indonesia
e-mail: [1]sachi16001@mail.unpad.ac.id, [2]r.sudrajat@unpad.ac.id, [3]intan.nurma@unpad.ac.id

Abstract—Computerized technology in the management system of digital museum objects is still limited in its application in centralized storage. This action provides vulnerability to manipulating original data from a collection. In this research, a historical object management system in Dapps was designed using a digital blockchain certificate based on Hyperledger Composer. Within the architecture, the blockchain-based web app or Dapps will manage the collection transparently, secure the access control, and store every transaction into its log. The Source of data used in the form of asset collection was taken from the Prabu Geusan Ulun Museum in Sumedang Regency. We evaluate the system by dividing the test into black boxes and white boxes. The functionality of the system is carried out in the black box using human interaction and measurements of security and performance are carried out in the white box using some testing tools.

Keywords—Blockchain, Data Manipulation, Museum

I. INTRODUCTION

Indonesia's cultural heritage is indeed an added value so that it is admired by foreign nations. Most of these heritages such as historical and antique objects are treated, stored, and secured in museums scattered throughout the archipelago. The Prabu Geusan Ulun Museum which is located in Sumedang district leaves various historical evidence that has occurred since the time of the Sumedang Larang Kingdom [1]. This museum contains various heirlooms, artifacts, manuscripts inherited from ancient Sumedang Kings as well as grants from the community and other cultural bodies whose total value can reach tens of billions [2]. To make it easier for visitors, the manager of the Prabu Geusan Ulun Museum provides a manuscript catalog that contains a summary of all the collections contained in the museum manager. In addition, transliteration and translation of the catalog from the Pegon script were also carried out into a language that is easily understood by visitors. The reality that occurs in the field is that there is still a lot of collection information that is processed and handled poorly [3]. The media used is still using the old method, a book consisting of 108 pages. Also, the development of centralized web technology in the distribution of information has been discontinued since 2014. Lack of awareness and understanding of the importance of this information has also resulted in some old collections not being recorded in the book collection. The possibility of misuse and falsification of data is quite high. The previous development using a centralized system also will cause the vulnerability of data leakage by third parties because the storage is only in one centralized place. Therefore, research was carried out using blockchain technology to support the collection management by creating the Dapps system. The system created will make the historical tracking of inventory more secure, transparent for its track record, and also has good performance for its implementation in the real world. To validate the results of the development system, we conducted several experiments involving functionality and system security, latency, and bandwidth consumption.

II. PRELIMINARY

A. Blockchain

Blockchain is a system of recording transactions in various places that are widely spread through many computers, each of which contains identical records [4]. Blockchain was originally introduced via Bitcoin by Satosi Nakamoto in 2008 in the field of cryptocurrency. At that time, blockchain was designed only to avoid double spending. However, now blockchain has been widely implemented in various ways, including digital identity, digital voting, or decentralized notaries.

Digital assets (such as units of credit, bonds, holdings, or fundamental rights) are managed as block lists of ordered transactions in blockchain. Each block in the blockchain will be linked to the previous block via a hash. Thus, the transaction history in the blockchain cannot be changed or deleted without changing the entire contents of the blockchain [5]. This makes the blockchain safe from hacker attacks.

B. Smart Contract

A smart contract or digital certificate is an agreement between two people in the form of computer code. Smart contracts run on a blockchain network, so they are stored on a public database and cannot be changed. Transactions that occur in smart contracts are processed by the blockchain, which means smart contracts can be sent automatically without third parties (banks, governments, brokers, etc.). Transactions only occur when the conditions in the agreement are met. In the absence of a third party, there is no entity to trust in running a smart contract [6]. Smart contracts have endless possibilities of use. Today smart contracts are used for financial trading, insurance, even for crowdfunding.

C. Hyperledger

Hyperledger contains collaborations across banking, finance, Internet of Things, manufacturing, supply chain, and technology. This project was introduced in 2015 where many companies were interested in its existence and awareness of

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

the benefits of working with collaborations it offered [7]. These companies decided to pool resources to create an open-source blockchain technology that anyone can use. Hyperledger is under the Linux Foundation which has grown rapidly in recent years.

D. Hyperledger Composer

Hyperledger Composer is a set of collaboration tools for building blockchain business networks that make it simple and fast for business owners and developers to create smart contracts and blockchain applications to solve business problems [8]. Built with JavaScript, leveraging modern tools including node.js,npm, CLI, and popular editors, Composer offers business-centric abstractions as well as sample apps with easy-to-test devops processes to create robust blockchain solutions that drive alignment across business requirements with technical development.

E. Decentralized Application

Decentralized Application or Dapps is a term for applications or systems that users use to interact with the blockchain network. As the name implies, the Dapps application has peer-to-peer characteristics, does not have a middleman/intermediary so that we and other users can make transactions directly [9]. The development and success of Dapps will determine whether blockchain technology can be widely adopted in society.

III. System Architecture

A. Data Requirement

In this study, the authors used data taken directly in the field from the Prabu Geusan Ulun Museum which is located in Sumedang Regency, West Java, Indonesia. All heirlooms from the Prabu Geusan Ulun Museum collection that have been inventoried are arranged and exhibited in several buildings consisting of a protected room. Classification is also carried out with the aim of creating groupings and facilitating the management and research of collections so that they can be used optimally for the purposes of education, study, and recreation. The entire collection can be classified as shown in Table 1.

TABLE 1. PRABU GEUSAN ULUN COLLECTION IN 2020

Classification	Collections Quantity
History	246
Ethnography	931
Graphics	13
Art	32
Numismatics	461
Biology	32
Philology	35
Archeology	37
Keramology	189
geography	1
Total	1977

From the total 1977 collection, only ± 700 of them are in adequate description and image. So in the implementation and testing, only 700 data are used for the entire operational system.

B. System Feature

After getting the data and discussing it with the internal team, we decided to divide the two main actors who play an important role in accessing the features in this Dapps system.

- Public users: Acted as a general user, this actor will only be able to access information such as asset collections, news, contacts, and historical articles. All these features are only given access to read data from the system.
- Administrator: This actor has special access to all existing features such as data collection including the holder, news, and others. Some access that is given directly to the system is create, read, update, delete, move ownership.

C. System Flowchart

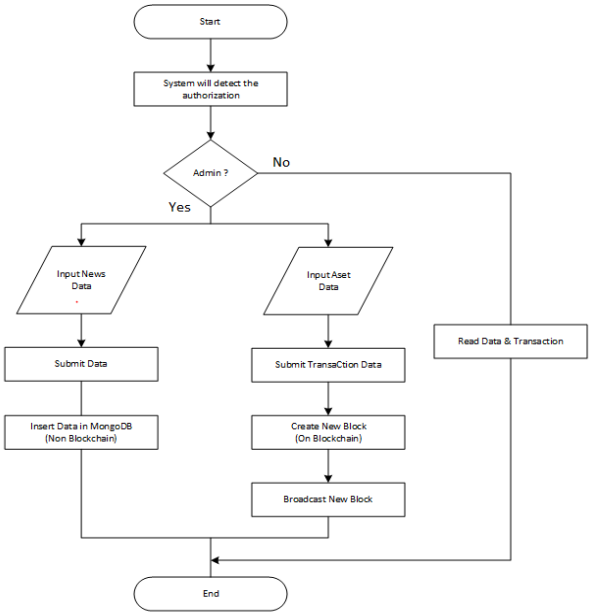


Fig. 1. Historical object management flowchart

The initial flow of the system began when the system will recognize the type of actor who will access the system, as shown in Fig. 1. Here, the role of the Access Control List (ACL) that we have already written in the system will strictly control the permissions of access to the system. This is the sample of ACL:

```
   Rule PublicOnlyCanRead {
    description: "Allow all public users read
access to all asset"
    participant: "model.Pengguna"
    operation: READ
    resource: "model.Koleksi"
    action: ALLOW
}
```

If the detected actor is administrator, then special access to some feature can be done. When the asset collection data is inputted by the admin, the blockchain system will

form a new block from the previous block containing the incoming collection data. The data block will be distributed to all nodes or peer in the system. Blocks in the blockchain chain have different encryption values. On the other hand, if the detected actor is not an admin, then the system will automatically only provide access to read data.

When admin submits the news data feature, the data will be inserted into MongoDB database. Therefore this data will not be processed into blocks like asset collections.

D. System Implementation

To implement the system design we discussed, we used Amazon Web Service (AWS) as the public host for the Dapps we built. In this architecture, two primary servers powered by AWS EC2 Instances are used. The first server uses linux virtualization and hyperledger framework which will be used as the operational core of Dapps (Fig. 2). The use of virtualization using the Linux operating system is intentionally used because the blockchain ecosystem developed by Hyperledger has fully supported the development of this operating system [10]. Container is used to accommodate 3 nodes which will share information and verify the suitability of museum collection data stored there. The virtual network environment setup, including resource deployment, connectivity, and security from inside and outside the server will be performed by AWS VPC.

The second server uses EC2 instances and linux virtualization as the host of the MongoDb database. This database is used to store additional data from application features such as news, publications, and contact data displayed to public users

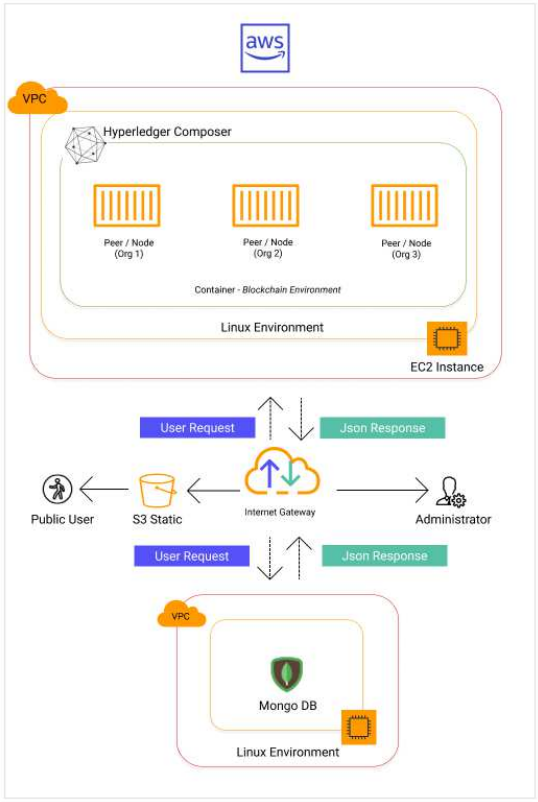


Fig. 2. The system design using Hyperledger Composer

For access by public users, the AWS S3 Static service is used to store web pages. This page will later interact with the data using the Application Programming Interfaces (APIs) that we have created from both servers. Admin will access the web page directly by using a special interface which will access data via API too.

IV. Result and Discussion

A. Implementation of interface functionality

The following is the result of the implementation of the interface design that has been made in the form of an museum management information system for general users and administrators. The general user interface design is devoted to the display when people first open a website page. It includes the homepage, history, collections, and contacts menu (Fig. 3 and 4).

Fig. 3. Home page for public users

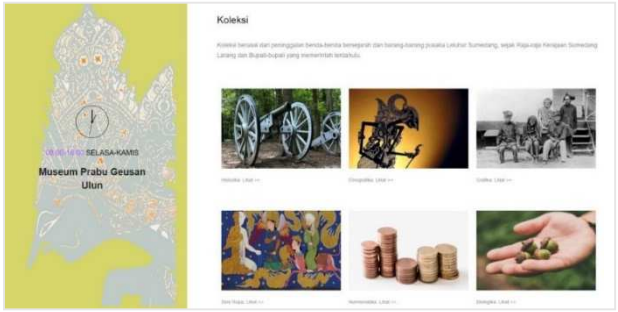


Fig. 4. Display of collection for public users

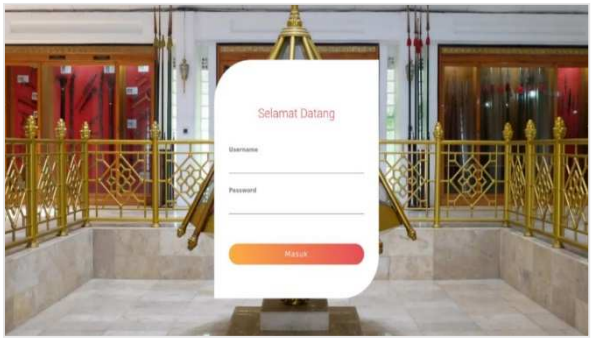


Fig. 5. Login page for admin

The design of the administrator interface is specifically for key parties such as administrators who have access rights to carry out an inventory of historical object data, as well as to publish news. The login page in Figure 5 contains an authentication system where only the administrator has access to enter the main system

The dashboard page in Figure 6 contains key information such as number of collections, news, collectors, heirs, and organization. Especially for collections, later the number of collections will be divided into 10 parts according to the distribution of collection types at the Prabu Geusan Ulun Museum. The data is displayed in the form of bar and circle charts.

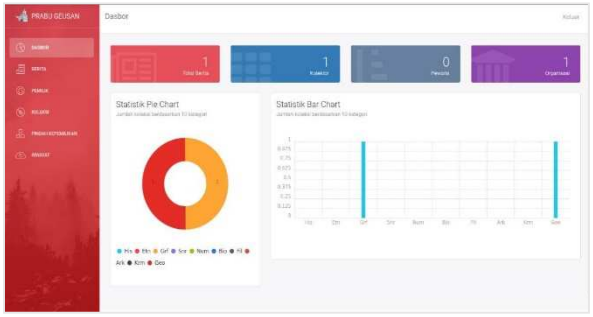


Fig. 6. Dashboard for admin

B. *Quality Control Testing*

1) Functional feature – Black box testing: This test involves human interaction to test all the features contained in Dapps. We tested it on several general respondents and the head curator of the Prabu Geusan Ulun Museum. These results also test which activities are prohibited and not prohibited to public users and admins. It shows in Table 2.

TABLE 2. ACCESS ACTIVITIES TESTING BASED ON ROLE

Access Activities	Black Box Testing	
	Admin	Public
Execute the 'Create Holder' function on the system	✓	×
Execute the 'Read Holder' function on the system	✓	×
Execute the 'Update Holder' function on the system	✓	×
Execute the 'Create Asset' function to the system	✓	×
Execute the 'Read Asset' function on the system	✓	✓
Execute the 'Update Asset' function on the system	✓	×
Execute the 'Move Ownership' transaction to the collection	✓	×
Execute the 'Create News' function to the system	✓	×
Running the 'Read News' function on the system	✓	✓
Execute the 'Update News' function on the system	✓	×
Execute the 'Delete News' function on the system	✓	×
Execute the 'Read Transaction' function on the system	✓	×

Note: ✓ means can access × means can not access

2) Rest server security - White box testing: The REST server security can be configured to authenticate clients, as shown in Fig. 7 and 8. When this option is enabled, clients' systems must authenticate to the REST server using an access token before they are permitted to call the REST API. The token is generated automatically by the hyperledger composer rest server. This security authentication can be simulated using Postman tools.

Fig. 7. Generated token from Hyperledger Rest Server

Fig. 8. Simulation of invalid token auth using Postman

3) Business network archive (BNA) security - White box testing: Each transaction in the system will be marked with its own identity. Each transaction identity will be converted to a hash into a series of identities consisting of letters and numbers. In this case, the transaction with a unique id known as a digital fingerprint/certificate or smart contract (Fig. 9).

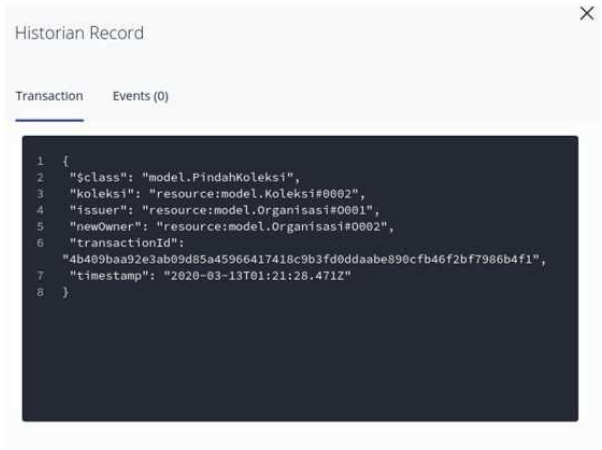


Fig. 9. Sample transaction with unique Id

The fingerprint has advantages over traditional logging systems. The use of it can make it easier to verify the source of claims against museum data collections. This system will also make it difficult for third parties to modify or defraud data information because all transactions will be transparently recorded along with the changed information.

4) API latency and bandwidth – White box testing: Latency testing is carried out to measure the time it takes for data to get to its destination across the network and API. It is usually measured as a round trip delay, the time taken for information to get to its destination and back again. The type of data being tested has a JSON (Javascript Object Notation)

format. We converted the previous 700 data into that form so that it can be processed by an automated performance tool called Apache Jmeter. The data size is shown in Table 3.

TABLE 3. DATA SIZE PER STAGE

Total Data	Data Format	Data Size (Kilobyte)
100	JSON	38
200	JSON	76
300	JSON	114
400	JSON	151
500	JSON	189
600	JSON	227
700	JSON	265

In this test, there will be seven experimental stages of incoming data transactions. In the first stage, 100 data will be entered first. The next stage will be added 100 more data until the total data entered is 700 data until the last stage. We simulate it with 3 users based on number of admin officers in the museum. Figure 10 shows the results of system latency or delay on seven trials.

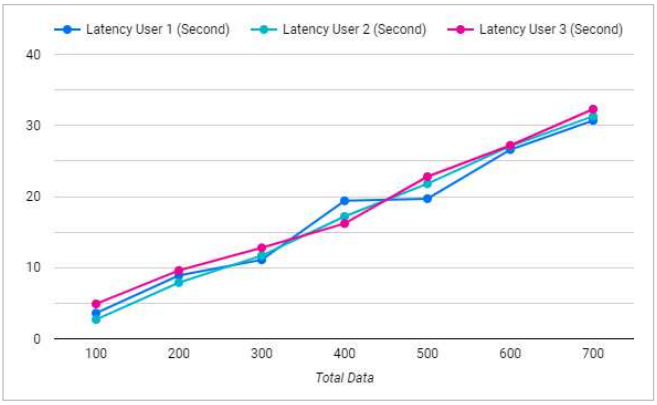


Fig. 10. Latency result of each trial

Bandwidth testing is carried out to measure the consumption of data transfer used when data is sent and received by the client systems. In seven experimental stages of data transactions that have been carried out, the results show that the more data that we sent, the larger the bandwidth will be consumed (Fig. 11).

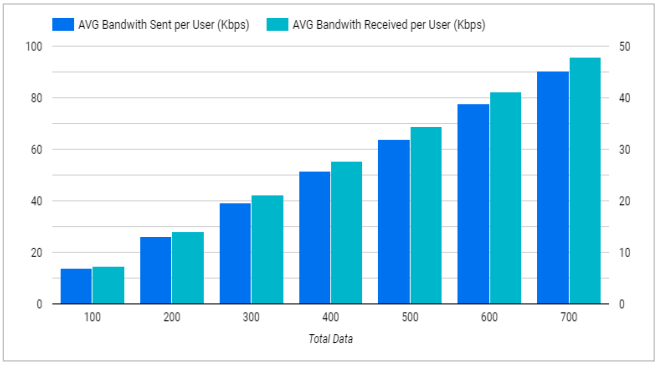


Fig. 11. Bandwith result for each trial

V. CONCLUSION AND SUGGESTION

Based on the development research that has been done, the entire Dapps system built with Hyperledger Composer relies on two main actors that have been arranged in ACL to access the feature. We have also added authentication tokens as an additional layer of security for accessing data in the system. All transaction activities that occur will be recorded in the BNA and automatically hashed with unique id in a digital certificate. With 700 data, the system throughout the API can process all data entities with an average latency of ± 32 seconds. The whole implementation shows that our system had met the criteria for building an information recording and management of historical object information that is safe, secure, and transparent. For future development and research, we suggest testing the system with various techniques and a larger data scale to achieve better performance and security.

ACKNOWLEDGMENT

The Author thanks the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] R. Arazak, Kajian Wisata Pusaka Museum Prabu Geusan Ulun Di Kabupaten Sumedang (dissertation). STPB, Bandung. (2017).

[2] Kosasih, A., & Mahdi, S., PELESTARIAN NASKAH-NASKAH KUNO DI MUSEUM PRABU GEUSAN ULUN SUMEDANG. Jurnal Universitas Padjadjaran, 91(5), (2012), 1689–1699.

[3] Hermawan, D., Sofian, M., & Kuswara, K., Improving the function of the prabu geusan ulun museum in Sumedang regency as a tourist attraction for historical and cultural education. Panggung, 27(4). https://doi.org/10.26742/panggung.v27i4.288. (2017)

[4] Champagne, P. The book of Satoshi: the collected writings of bitcoin creator Satoshi Nakamoto.: Phil Champagne. E53 Publishing LLC. (2014).

[5] Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., Pautasso, C., & Rimba, P., A Taxonomy of Blockchain-Based Systems for Architecture Design. Proceedings - 2017 IEEE International Conference on Software Architecture, ICSA 2017, (2017). 243–252.

[6] Ma, R., Gorzny, J., Zulkoski, E., Bak, K., & Mack, O. V. Fundamentals ofSmart Contract Security. Momentum Press. (2019).

[7] Blummer, T., Bohan, S., Bowman, M., Cachin, C., & Gaski, N. An Introduction to Hyperledger. Hyperledger White Paper. (2018).

[8] Gaur, N., Desrosiers, L., Ramakrishna, V., Novotný Petr, Baset, S. A., & O'Dowd, A. Hands-on blockchain with Hyperledger: Building decentralized applications with Hyperledger fabric and composer. Packt Publishing. (2018).

[9] Tatar, J. How Blockchain Could Change How We Vote. Inc_ The Balance. (2020).

Affiliates, A. and. (n.d.). Managed blockchain hyperledger fabric. Retrieved April 6, 2020, from https://docs .aws.amazon.com/managed-blockchain/latest/hyperledger-fabric-dev/amazon-managed-lockchain-hyperledger-fabric-dev.pdf.

Electronic News Sentiment Analysis Application to New Normal Policy During The Covid-19 Pandemic Using Fasttext And Machine Learning

Rividya Permata Aluna[ab1], Intan Nurma Yulita[ab2], R. Sudrajat[ab3]
[a]*Research Center for Artificial Intelligence and Big Data*
[b]*Department of Statistics, Faculty of Mathematics and Natural Science*
Universitas Padjadjaran
Bandung, Indonesia
e-mail: [1]ivy.aluna@gmail.com, [2]intan.nurma@unpad.ac.id, [3]r.sudrajat@unpad.ac.id

***Abstract*—The new phase in handling COVID-19 in Indonesia, called New Normal, gives various public perspectives regarding this policy. This study aims to analyze public sentiment towards the New Normal policy through an electronic news comment column. This study uses text data in the form of comments were collected from electronic news media sites, namely www.detik.com and www.kompas.com, and taken from the comments column on Instagram social media, namely the @detikcom account. Also, use FastText method to extract features by converting data into vector values and using three classification methods, Naive Bayes (NB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). This study conducted a hyperparameter test to obtain the most optimal model. Testing the hyperparameters from FastText produces an optimal model with dimensions of 250, window size 8, epoch 1.000, and a learning rate of 0,0025. Hyperparameter testing was also carried out on the SVM and MLP classifiers. Hyperparameter testing of the SVM and MLP classifiers produces the most optimal model with the SVM method using the RBF kernel, C of 1.000, gamma of 10. In contrast, the MLP method uses the relu activation function, hidden size layer (250,250), adam optimizer, alpha 0,0001, and adaptive learning rate. The classification model was evaluated using K-fold cross-validation to produce an average f1score. The result is for the NB method 72,25% f1score, for the SVM method 92,21% f1score, and for the MLP method 90,75% f1score.**

Keywords— FastText, Multilayer Perceptron, Naive Bayes, New Normal, Sentiment Analysis, Support Vector Machine

I. INTRODUCTION

According to WHO, COVID-19 (Corona Virus Disease 2019) is an outbreak that is spreading because the latest type of coronavirus was revealed and was previously unidentified before the virus occurred in Wuhan, China in December 2019. The COVID-19 handling phase has entered a new phase called "New Normal" [1]. The various kinds of government policies make people confused and assess the discrepancy between policies. The public's anxiety is reflected in their opinions in the electronic news commentary column. This discrepancy shows that political communication is not smooth between government elites. With data published by Hootsuite in the 2020 report, it explains that internet use in Indonesia is 760 million Indonesians. The internet is widely used by users, one of which is the use of the internet in the socio-political field. These internet users are commonly referred to as netizens. Netizens or online communities accessing this field are around 50.46% of the total internet users in Indonesia, which is 143.26 million people. Netizens access matters related to socio-politics and provide opinions or comments on these political issues. To find out the reactions given by the community regarding the policies that the government makes, sentiment analysis can be carried out.

Sentiment analysis is computational research of opinions, sentiments, and emotions expressed textually [2]. Sentiment analysis is one of the fields of natural language processing (NLP), which forms a system for recognizing and extracting opinions in text form. Currently, there is a lot of text-based information provided on the internet in the format of forums, blogs, social media, and websites with comments. Many approaches can be made to classify one using machine learning models.

Machine Learning is generally used for binary classification and prediction as positive or negative sentiment. Machine Learning Algorithms are further classified as supervised, unsupervised, and semi-supervised. Examples of machine learning algorithms include Naive Bayes, Decision Tree, Support Vector Machine, and Multilayer Perceptron. One of the most popular methods used in document classification today is the Naïve Bayes classifier method. The Naïve Bayes classifier method is highly efficient when applied to large databases and diverse data [3].

Furthermore, the classification method is the Support Vector Machine. This method can generalize in classifying a pattern, excluding the data used in the learning phase of the process [4]. In addition to using the Naive Bayes method and the Support Vector Machine, there is a Multilayer Perceptron classification method. The multilayer perceptron is an artificial neural network algorithm that adopts the workings of neural networks in living things. This algorithm is reliable because the learning process can be carried out in a directed manner [3].

When modeling using machine learning, data must be assigned a value to be calculated. One way is to use word embedding FastText. FastText (Bojanowski et al., 2017) is a word embedding method that develops word2vec. This method studies word representation by considering subword information. Each word is represented as a set of n-gram characters. Thus, it can help capture the meaning of shorter words and allow embedding to understand the suffixes and prefixes of the word [5].

Regarding the New Normal policy taken by the government in dealing with COVID-19, this study intends to

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

examine public sentiment towards the New Normal policy in dealing with COVID-19 in Indonesia. This study builds software that can assist research when conducting sentiment analysis. This study also compares three machine learning classification methods: Naive Bayes, Support Vector Machine, and Multilayer Perceptron.

II. METHODOLOGY

A. Machine Learning

Machine Learning or machine learning is an approach in AI that is widely used to replace or imitate human behaviour to solve problems or perform automation. As the name implies, Machine Learning tries to mimic how humans or intelligent creatures learn and generalize [3]. Machine Learning is generally used for binary classification and prediction as positive or negative sentiment.

B. Text Pre-processing

Pre-processing of the data used to extract interesting and non-trivial and knowledge of unstructured text data. Pre-processing techniques such as Tokenization, Stopword removal, lowercase conversion, and Stemming are used for text documents [6].

- The Stopwords removal technique removes stopwords such as prepositions, articles, pronouns, etc. (which does not give the document's meaning, i.e., in, a, an, with). Removing stopwords will reduce the dimensions of the term space.
- Tokenization is to identify meaningful keywords. Inconsistencies can be different numbers and time formats. The challenge in tokenization depends on the type of language. Languages like English and French are referred to as space-delimited because white spaces separate most words. Tokenization is also affected by the writing system and typographic structure of the word. Another problem is abbreviations and acronyms that must be converted into standard forms.
- Lowercase conversion is for words that appear the same each time they appear. This method is used for all text content is all text in lowercase, and it is straightforward to analyze the content.
- Finally, stemming is used to identify the root/stem of a word. This method aims to eliminate various suffixes, reduce the number of words, have stems that match accurately, and save time and memory space.

C. FastText

This is an extension of Mikolov's embedding. The FastText approach is based on the skip-gram model, where each word is represented as a bag n-gram character. The vector representation is associated with each n-gram character; the words are defined as the sum of these representations. Representation word studied taking into account the left and right large windows context words. Unlike Mikolov embeddings, FastText can provide embedding for misspelled words, rare words, or those not in the training corpus because FastText uses n-gram character word tokenization [5].

D. Naive Bayes

The Naïve Bayes Classifier originally provided by Thomas Bayes is easy to implement and computationally efficient compared to other machine learning algorithms. This is a supervised classifier used to calculate the probability of data being positive or negative.

The most efficient and effective inductive learning algorithm for machine learning and data mining is Naïve Bayes. It is based on Bayes' Theorem assuming independence among predictors. For real-world applications, its competitive performance in the classification is very rarely correct. In simple words, the Naïve Bayes classifier assumes that a predetermined property is not related to the presence of any other features present. The Naïve Bayes model is especially useful for large data sets [7].

E. Bernoulli Naive Bayes

In the Bernoulli classifier Naïve Bayes algorithm, a feature is an independent binary variable which that states whether a term that exists in the document is considered or not. Similar to the multinomial model in the classification process, this algorithm is also a popular approach for text classification but differs from the multinomial system in calculating frequency terms. Bernoulli is only interested in designing whether a term is present in the document under consideration [8].

F. Support Vector Machine

The Support Vector Machine (SVM) has been selected for classification in the experiment. A support-vector engine is a learning machine for the two-group classification problem introduced by [9]. It is used to classify text as positive or negative. SVM works well for text classification because of its advantages such as its potential to handle large features. Another advantage is that SVM is robust when there is a sparse set of instances and also because most problems can be linearly separated [4].

G. Multilayer Perceptron

MLP operates as a universal function approximator having at least one hidden layer and multiple non-linear units, making it efficient to learn any relation between input variable sets. Multilayer Perceptron (MLP) has a uni-directional data flow, just like data flowing from the input layer to the output layer. For example, the Neural Network multilayer perceptron (MLP) starts with the input layer having every node as a predictor variable. Then, neurons (input nodes) are interconnected with the neurons in the forward flowing and the next layer (the hidden layer).

Multilayer perceptron has two phases. The activation is propagated from the input layer to the output layer in the first phase, called the forward step. In the second phase, the errors between the actual & operational values and the requested nominal values are replicated in the reverse direction. Due to its application as a universal function approximator, MLP is a famous algorithm, having at least one hidden layer with multiple non-linear objects that can learn almost all functions or relationships within a given input and output variable set because of its “backpropagation” [7].

H. Cross-Validation

Cross-Validation is a statistical method for evaluating and comparing learning algorithms by dividing the data into two segments. One is used to study or train the model, and the other is used to validate the model. In typical cross-validation, the training and validation sets must cross in successive loops so that each data point has a chance to be validated. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-

validation or involve repeated rounds of k-fold cross-validation. K-fold cross-validation was used to eliminate bias in the data. Repeat this process k times so that each partition is only used for testing once [10].

I. Confusion Matrix

Confusion matrices are a way to evaluate errors in classification problems [11]. The confusion matrix summarizes the comparison of the prediction results shown with actual models and results. For the multilabel case, a confusion matrix will be presented for each label sequentially. In addition, the confusion matrix has some performance metrics to measure the performance of models that have already been made. Frequently used metrics are:

a. The accuracy value is the proportion of the number of correct predictions.

b. Recall is used to compare the proportion of TP to positive tuples.

c. Precision is the proportion of cases with a positive diagnosis.

d. The F1-score is the weighted ratio of the average precision and recall.

III. METHODS

This research was conducted with research stages which include data collection, data labeling, data preprocessing, word embedding, the process of designing and building a classification model, and evaluation. Fig. 1 is the flow or stages of the research carried out.

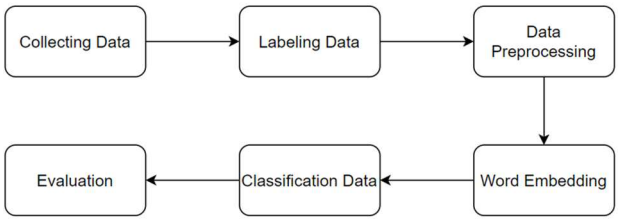


Fig. 1 Research Procedure

A. Data Collection

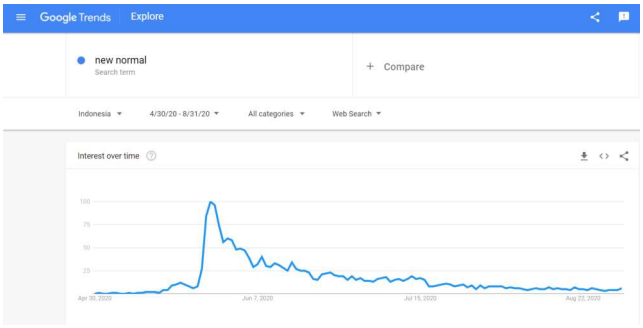


Fig. 2 Trend New Normal in Indonesia

Text data in the form of comments were collected from electronic news media sites, namely www.detik.com and www.kompas.com, and taken from the comments column on Instagram social media, namely the @detikcom account. Data collection was only taken on news published from May 2020 to July 2020 because media sites that reported news about the New Normal had decreased after July 2020.

Fig. 2 shows the trend of the New Normal in Indonesia, the decline since July 2020. Data were taken manually to the comments in the comments field of electronic news and the technique for scrapping it online by using an extension of Google Chrome that is Web Scraper for data commentary on social media Instagram.

The data used in building the model in this study are 1,601 comments in the form of a *.xlsx file format. Of the 1,601 data, labelling was carried out into positive sentiment class and negative sentiment class. The number of positive class data used in this study was 515 data, while for the negative class there were 1,086 data so that the dataset in this study experienced class imbalance. Fig. 3 is the percentage of the class label distribution, it can be seen that the data has an imbalance class because positive data has a percentage of 32% and negative data is 68%.

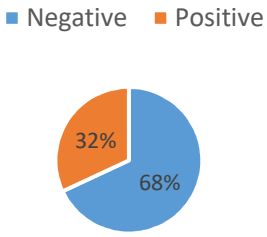


Fig. 3 Percentage of class label sharing

B. Data Labelling

The comment data collected in a file with the extension *xlsx is then labeled by assigning a meaningful tag to each data. Comment data are classified into positive and negative sentiment classes.

C. Data Pre-processing

At this stage, the data will be prepared so that it can be analyzed. The data to be analyzed in this study is commentary data. The pre-processing step is carried out to clean data from elements that are not needed. The pre-processing stage consists of case folding, noise removal, stopword removal, and tokenization. Fig. 4 flowchart of pre-processing data stages.

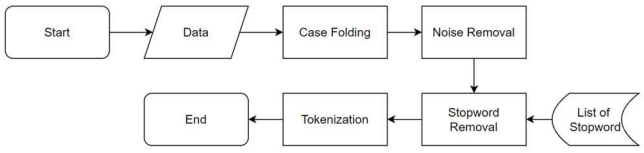


Fig. 4 Data Pre-processing

D. Word Embedding

After the data has gone through the pre-processing stage, the data will be processed by word vector representation, otherwise known as word embeddings. At this stage, the words will be converted into vectors used in the classification stage. This research uses FastText word embedding with skip-gram architecture.

IV. RESULTS

A. New Normal Policy Sentiment Analysis Results

Based on data taken from electronic news comments discussing the New Normal policy during the COVID-19 pandemic, 1,601 data were obtained. From the 1,601 data obtained, 1,086 data were labelled negative, and 515 data were labelled positive. Data collection was taken from May 2020 to July 2020.

B. FastText Test Result

Dimension size is a vector size that represents a word in the data. The dimensions of the experiment were carried out using dimension sizes of 10, 50, 100, 200, 250, and 350. Tests carried out with other hyperparameters were made to have a constant value, namely the window size of 8, the learning rate of 0.0025, and the epoch of 1,000.

In hyperparameter testing to determine the optimal dimension size, the results show dimensions with 250 giving the best F1score value. The results of these experiments can be seen in Table I; The number of dimensions with a size of 250 is the optimal vector size representing a word in the data. Dimensions that are too large can cause additional vectors to appear that do not represent the word's sentiment.

TABLE I RESULTS OF TESTING HYPERPARAMETER SIZE DIMENSIONS FASTTEXT

No.	Size Dimension	Window	Learning Rate	Epoch	F1-score
1	10	8	0,0025	1.000	76,32%
2	50	8	0,0025	1.000	85,53%
3	100	8	0,0025	1.000	87,28%
4	200	8	0,0025	1.000	87,54%
5	**250**	**8**	**0,0025**	**1.000**	**88,39%**
6	350	8	0,0025	1.000	87,32%

C. Classification Test Results

After getting the optimal FastText model, then testing is carried out to get the optimal classification model. The classification model uses three machine learning algorithms, namely Naive Bayes, SVM, and MLP. Tests are carried out to obtain the optimal classification model by testing the hyperparameters of the algorithm. This study also uses the SMOTE algorithm because the data has an imbalanced class problem.

a. Classification with Naive Bayes

In the classification model using Naive Bayes, the model is made using Bernoulli Naive Bayes. The alpha used in this study is an alpha of 1. The classification results with Naive Bayes can be seen in **Error! Reference source not found.** by looking at the average F1-score of 10 fold cross-validation; an average F1-score of 72.25% is obtained.

b. Classification with Support Vector Machine

The next test is performed on the hyperparameter classifier support vector machine. The hyperparameters tested are the number of C and gamma. During the test, the FastText hyperparameter uses the results from the previous test.

1) Number of C

In testing, the amount of C gamma is made constant, namely 1. Testing the amount of C is carried out using the number of C 0.1, 1, 10,100, 1000, and 10000. The result shows that C of 1000 gives the best f1score value. The results of these experiments can be seen in Table 2. The amount of C of 1000 is the optimal amount of C because if the amount of C is too large, it can cause overfitting, while the amount of C that is too small can cause underfitting.

TABLE II TEST RESULTS AMOUNT OF C

No.	C	F1-score
1	0.1	81.53%
2	1	83.89%
3	10	87.52%
4	100	90.10%
5	**1000**	**90.56%**
6	10000	90.16%

2) Gamma

The gamma test uses the amount of C based on the previous test results, which is the most optimal amount of C, which is 1000. The test is carried out using a gamma of 100, 10, 1, 0.1, 0.01, 0.001. The result shows that a gamma of 10 gives the best f1score value. The results of these experiments can be seen in Table . A gamma of 10 is the optimal gamma because if it is too large, it can cause overfitting, while a too-small gamma can cause underfitting.

TABLE III TEST RESULTS OF HYPERPARAMETER GAMMA

No.	Gamma	F1-score
1	100	83.27%
2	**10**	**92.07%**
3	1	90.56%
4	0.1	87.86%
5	0.01	83.23%
6	0.001	81.13%

c. Classification with Multilayer Perceptron

The test was carried out on the hyperparameter classifier multilayer perceptron. The tested hyperparameters are hidden layer sizes, activation function, and solver. During the test, the FastText hyperparameter uses the results from the previous test. The other hyperparameter is made constant, namely the adaptive learning rate.

1)Hidden Layer Sizes

When testing the number of hidden layer sizes, other hyperparameters are made constant, using the activation function relu and adam's solver. Tests were carried out using the number of hidden layer sizes of (10,10), (50,50), (100,100), (200,200), (250,250), and (300,300). The result shows that the hidden layer sizes of (250,250) give the best f1score value. The results of these experiments can be seen in Table IVII. The number of hidden layer sizes of (250,250) is the optimal number of hidden layer sizes.

TABLE IVII TEST RESULTS FOR HIDDEN LAYER SIZES

No.	Hidden Size Layer	F1-score
1	10,10	81.85%
2	50,50	87.76%
3	100,100	89.80%
4	200,200	90.19%
5	**250,250**	**90.76%**
6	300,300	90.11%

2) Activation Function

The activation function test uses the number of hidden layer sizes based on the previous test results, which is the most optimal number of hidden layer sizes (250,250). In addition, the test was carried out with other hyperparameters made of constant value, namely using the adam solver. The test is carried out using the activation functions, namely relu and tanh. The result shows that the relu activation function gives the best f1score value. The results of these experiments can be seen in Table V. The relu activation function is the optimal activation function.

TABLE V TEST RESULTS FOR ACTIVATION FUNCTION HYPERPARAMETER

No.	Activation	F1-score
1	**relu**	**90.76%**
2	tanh	82.73%

3) Solver

Solver testing uses the number of hidden layer sizes based on the previous test results, which is the most optimal number of hidden layer sizes (250,250) and the activation function is relu. The test was carried out using the solver, namely adam and sgd. The result shows that Adam's solver gives the best f1 score. The results of these experiments can be seen in VI. The adam is an optimal solver.

TABLE VI TEST RESULTS FOR SOLVER HYPERPARAMETER

No.	Solver	Learning Rate	F1-score
1	**adam**	**adaptive**	**90.76%**
2	sgd	adaptive	64.72%

D. Evaluation Result

The training process looks at the FastText model using hyperparameter dimensions of 250. The classification model using Naïve Bayes with an alpha of 1, the SVM model uses hyperparameters C 1,000 and gamma 10, and the MLP model uses hidden layer sizes (250,250), the relay activation function, and Adam's solver. The optimal model is then evaluated using a confusion matrix and K fold cross-validation. The results of the evaluation using the confusion matrix can be seen in Table III. The confusion matrix is the average result of the confusion matrix when performing 10 fold cross-validation. Evaluation using a confusion matrix resulted in the Naive Bayes classification model getting the average data predicted to be correct with a negative class of 89.90 data and 36.90 incorrectly predicting a negative class. While the average data predicted correctly with a positive class amounted to 69.50 data, and incorrectly predicted a positive class amounted to 16.50 data. The SVM classification model predicted the average data correctly, with negative classes totaling 97.60 data and incorrectly predicting negative classes totaling 7.90. While the average data predicted correctly with positive class amounted to 98.50 data and incorrectly predicted 8.80 data in positive class. In the MLP classification model, the average data correctly predicted the negative class amounting to 95.10 data and incorrectly predicted the negative class amounting to 8.60. While the average data predicted correctly with positive class amounted to 97.80 data and incorrectly predicted 11.30 data in positive class.

TABLE III CONFUSION MATRIX EVALUATION RESULTS

Actual Results	Prediction Results					
	Naive Bayes		*SVM*		*MLP*	
	Negative	*Positive*	*Negative*	*Positive*	*Negative*	*Positive*
Negative	89,90	16,50	97,60	8,80	95,10	11,30
Positive	36,90	69,50	7,90	98,50	8,60	97,80

Table IV shows the results of the evaluation using k-fold cross-validation. From 10 folds, the average F1-score for each classification model is obtained for the Naive Bayes model, the average F1-score is 72.25%, for the SVM model, the average F1-score is 92.21%, and for the MLP model, it is obtained the average F1-score is 90.75%. Evaluation with k-fold cross-validation resulted in the SVM classification model being the most optimal classification model, which can be seen from the average F1-score, which has the highest value compared to the Naive Bayes and MLP classification models.

TABLE IV K-FOLD CROSS VALIDATION EVALUATION RESULTS

Fold	Naïve Bayes	SVM	MLP
	F1-score	*F1-score*	*F1-score*
1	75,13%	93,66%	89,76%
2	74,23%	91,74%	91,30%
3	74,42%	92,47%	91,89%
4	72,64%	89,81%	86,41%
5	70,65%	94,29%	90,24%
6	72,92%	94,34%	93,07%
7	72,46%	93,62%	90,83%
8	68,34%	92,86%	89,42%
9	70,39%	91,63%	90,91%
10	71,36%	87,72%	93,64%
Average	**72,25%**	**92,21%**	**90,75%**

The K-fold cross-validation and confusion matrix evaluation results found that the SVM model has the highest F1-score value. This value indicates that the FastText model can work well using the SVM classification model. SVM is the best algorithm because Naive Bayes only looks at the probability of the available data, so it is not suitable if it is with FastText that can capture words outside of the already available data. In addition, because MLP in the case of classification performs fast forward based on activation

without storing feedback so that feedback is not stored in each neuron cell.

V. Conclusion

Based on the research that the author has done in analyzing electronic news sentiment on the New Normal policy during the COVID-19 pandemic using FastText, the following conclusions can be drawn:

- In research conducted on the data taken, it was found that public sentiment towards the New Normal policy during the COVID-19 pandemic in Indonesia was negative. It can be seen from the percentage of data that is 68% of data with negative sentiment and 32% of data with positive sentiment.
- The FastText model built in this study results from hyperparameter testing obtained from the best FastText model with dimensions of 250, window size 8, learning rate 0.0025, and epoch 1000, which produces an average F1-score of 88.39%.
- The classification model has gone through hyperparameter testing. A classification model is obtained with the Naive Bayes model, SVM with RBF kernel, C 1000, and gamma 10, MLP with hidden size layer (250,250), relu activation function, and Adam's solver. The Naive Bayes model produces an average F1-score of 72.25%, the SVM model produces an average F1-score of 92.21%, and the MLP model produces an average F1-score of 90.75%.

Acknowledgment

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

References

[1] H. LIPI, 26 June 2020. [Online]. Available: http://www.politik.lipi.go.id/kegiatan/1406-sinergi-pemerintah-dalam-masa-new-normal-covid-19.

[2] B. Liu, "Sentiment Analysis and Subjectivity," 2009.

[3] M. Ahmad, S. Aftab, S. S. Muhammad and S. Ahmad, "Machine Learning Techniques for Sentiment Analysis: A Review," *International Journal Of Multidisciplinary Sciences And Engineering, VOL. 8 , NO. 3,* 2017.

[4] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," 1998.

[5] A. Geet D'Sa, I. Illina and D. Fohr, "BERT and fastText Embeddings for Automatic Detection of Toxic Speech," *IEEE,* 2020.

[6] B. K. Poornima, D. Deenadayalan and A. Kangaiammal, "Text Preprocessing on Extracted Text from Audio/Video using R," *International Journal of Computational Intelligence and Informatics,* vol. 6, 2017.

[7] M. Ahmad, S. Aftab, S. Shah Muhammad and S. Ahmad, "Machine Learning Techniques for Sentiment Analysis: A Review," *International Journal Of Multidisciplinary Sciences And EngineerinG,* vol. VIII, 2017.

[8] G. Singh, B. Kumar and L. Gaur, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 International Conference on Automation, Computational and Technology Management (ICACTM),* 2019.

[9] C. CORTES and V. VAPNIK, "Support-Vector Networks," in *Machine Leaming, 20*, Boston, Kluwer Academic, 1995, pp. 273-297.

[10] P. Refaeilzadeh, L. Tang and H. Liu, "Cross Validation," 2009.

[11] E. Beauxis-Aussalet and L. Hardman, "Visualization of Confusion Matrix for Non-Expert Users (Poster)," 2014.

SK-MOEFS Multi-Objective Evolutionary Fuzzy System Library effectiveness as User-Friendly Cryptocurrency Prediction Tool

Dio Satyaloka
Research Center for AI and Big Data
Dept. of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
dio.satyaloka@gmail.com

Stacyana Giamiko
Research Center for AI and Big Data
Dept. of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
stacyana15001@mail.unpad.ac.id

Akik Hidayat
Research Center for AI and Big Data
Dept. of Computer Science
Universitas Padjadjaran
Sumedang, Indonesia
akik@unpad.ac.id

***Abstract*—The emergence of Cryptocurrency has long foreshadowed a more accessible exchange market. Cryptocurrency is easy to use and trade, and this ease of access into the market brought newcomers into the Crypto-trading scene. This surge of inexperienced newcomers causes market instability and major loss amongst themselves. AI models, algorithms, and systems have been long used as an important aspect of prediction. However, the use of AI systems is complex. AI tools and systems often use complicated mathematical formulas and are not easily understood. Amongst these AI systems, Fuzzy Rule-Based Systems (FRBSs) has one of the most easily understood displays. With accuracy that rivals of other less-understood methods, such as Neural Network, FRBSs present us a choice that is easily used by users while keeping the interface as basic and simple as possible. This paper aims to study the use of FRBSs using SK-MOEFS (SciKit-Multi Objective Evolutionary Fuzzy System) Python Library in predicting a bull signal or a bear signal in the Cryptocurrency market while still preserving FRBSs user-friendly nature. The fuzzy sets are partitioned as Very Low, Low, Medium, High, and Very High. Then the resulting classification are used to signal whether a Cryptocurrency is bearish or bullish on the current day. The parameter used on the system yields an undesirable result of 53% accuracy with 25 Total Rule Length, however still producing the desired ease-of-use nature of FRBSs.**

Keywords— fuzzy inference system, fuzzy, fuzzy system, fuzzy rule-based system, cryptocurrency market, price forecasting, multi-objective evolutionary fuzzy system

I. INTRODUCTION

Cryptocurrency, such as Bitcoin, Ethereum, or the recent most popular one Dogecoin, has been the talk of the internet since the dawn of Bitcoin in 2009. The rise and fall of Cryptocurrency catch the interest of many people. Amongst them are those who are inexperienced yet eager to jump into the fray. Cryptocurrency also well-known for its volatility, instability, and frequent market fluctuation. Those new to the scene often find themselves at a loss and ended up losing their assets due to market instability.

II. LITERATURE STUDY

A. Exchange Market

Exchange Market is a platform mainly used to trade in Stocks, Foreign Currency, and now Cryptocurrency [1]. In this paper we aim only at Cryptocurrency Exchange Market. In general, an Exchange usually has this main terminology [2]:

1) *Low:* The lowest price of a currency, in this case, a Coin/Token of a Cryptocurrency, in one period,
2) *High:* The highest price of a Coin/Token in one period,
3) *Open:* The price position of a Coin/Token when the Exchange Market period started,
4) *Close:* The price position of a Coin/Token when the Exchange Market period ended,
5) *Volume*: The value amount of transaction that happened in the Exchange on one period,
6) *Market Cap:* The dollar value of Coin/Token available in circulation,
7) *Bull Market:* A condition in which market commodities prices are expected to rise,
8) *Bear Market:* A condition in which market commodities prices are expected to decline.

A Technical Analysis (TA) and Fundamental Analysis (FA) applies on Cryptocurrency Exchanges just like Stock Exchanges. Charts are displayed in candlestick form, and then patterns are analyzed based on price positions. These TAs and FAs are often too complicated for anyone new to trading. And thus, a solution for easier prediction methods is needed.

B. Crypocurrency

Cryptocurrency is a distributed and decentralized currency system based on cryptography that allows us to process transactions safely [3]. In Cryptocurrency, cryptography is used in transaction securities and transaction confirmations. Thus, boasting a secure way of processing transactions, and a sure way for validating the content of a transaction. All these Cryptocurrency transactions are contained within an environment called Blockchain, and the transacted currency is called coin or token. In this research paper, we are using Ethereum coin as the subject of the research. According to the United States of America's National Archives and Records Administration, the elements of a Cryptocurrency are:

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

C. Blockchain

1) Block: Similar to a ledger, it is made of a list of transactions that happened in a span of a time period. Any kind of settings or rules that relates to the Block is established when the network first created. These rules could include something like the maximum number of transactions per Block or the size of each Block.

2) Chain: As a Block approaches its maximum transactions, it is then linked into the preceding block through a hash. This hash value is inserted into the next block, thus creating a chain of Blocks. With this repeating chain, a new block will always have the same fixed-length value as the previous block due to using the same hash function. And due to this when a block is altered, modified, or tempered, the resulting hash will also be different. This then can be used to verify the validity of a block, whether it has been altered and can no longer be trusted or no.

3) Network: A Blockchain network is made up of nodes that contains all transactions record that exist on a Blockchain. Due to the decentralized nature of a Cryptocurrency Blockchain, there is no single official node on the network. Neither there is the most trustworthy node. The integrity is maintained by replicating the blockchain on all the nodes.

D. Token/Coin

A symbol representing a Cryptocurrency. In this paper for example, the token Ethereum coin is used as the subject.

III. Method

A. Fuzzy Rule-Based Systems

In general, a Fuzzy Rule-Based System is composed of the linguistic rule set also named Knowledge Base (KB) [4] and the fuzzy inference system. The KB are made of linguistic rules and parameters used to define the rules. While the inference engine works in generating the prediction or classification based on new inputs according to the rules in the knowledge base [5].

Let $Z = \{Z_1,...,Z_F\}$ be the set of input attributes and Z_{F+1} be the output attribute. Let $\boldsymbol{U_f}$, with $f = 1, ..., F + 1$, be the universe of the f^{th} attribute Z_f . Let $Z_f = \{A_{f,1},...,A_{f,Tf}\}$ be a fuzzy partition of T_f fuzzy sets on attribute Z_f. Finally, we define the training set $\{(z_1, z_{F+1,1}),...,(z_N, z_{F+1,N})\}$ as a collection of N input-output pairs, with $z_t = [z_{t,1}...,z_{t,F}] \in R^F$, $t = 1,...,N$.

In classification problems [4], Z_{F+1} is categorical and $z_{F+1,t} \in C$, where $C = \{C_1,...,C_K\}$ is the set of K possible classes. With the aim of determining the class of a given input vector, we can adopt a Fuzzy Rule-Based Classifier (FRBC) with an RB composed of M rules expressed as:

Rm : **IF** $\mathbf{Z_1}$ **is** $A_{1,jm,1}$ **AND ... AND** $\mathbf{Z_f}$ **is** $A_{f,jm,f}$ **AND ... AND** $\mathbf{Z_F}$ **is** $A_{F,jm,F}$ **THEN** $\mathbf{Z_{F+1}}$ **is** C_{jm} *with* RW_m

IV. Genetic Algorithm

Genetic Algorhitm (GA) is a part of Evolutionary Algorithm, which is an algorithm that takes natural evolution process as an example, where the main concept is survival of the fittest. The fitness of the population samples will be tested by a fitness function. Fitness in GA is defined as a visualization of how fit a solution towards a certain problem. Fitness function will produce a fitness value that will be the reference for the next process of the GA.

GA process starts by deciding the initial population which contains some chromosomes that is built by some genes that represents the solution candidates of a problem. The best candidates will be picked by a selection process based of the fitness value that already calculated for every chromosomes in the population. The picked candidates are the samples that will be used to fill the mating pool, which is a set where pairs of parents will be formed. In Evolutionary Algorithm, the principle of surviving arise because of the reproduction process. The offsprings that produced will carry the parents genes, which is why the parents picked to fill the mating pool are the best ones from the population. Therefore, the offsprings that being produced are the ones who have best genes from both of their parents.

GA is one of the algorithms that is being used for optimization process. In optimization, the optimal condition of the solutions got from the the main target that will be reached. But in optimization algorithm, local optimum often happens. Local optimum is a condition where the algorithm reached the highest or lowest value in some solution candidates. This contradicts the global optimum, which is when the algorithm reached the highest or lowest for all the solution candidates for a problem. Local optimum can happen when the population have a premature convergence. One way to prevent this to happen is to keep the variance of the chromosomes of the population. In GA, chromosome variance can be kept by implementing crossover and mutation operator.

Crossover is a recombinating operator that has a goal to get the best sample. Crossover operator will do a recombination from parent sets that picked randomly from the mating pool by the selection process. Crossover will produce a set of offspring with the variance kept for the next process, which is mutation. Mutation operator will keep the variance by switching one or more genes in the chromosome with the opposite value. For example, if our chromoseome have binary value 0 and 1, then if 1 is picked randomly for a mutation point, this value will be switched with 0, and the other way around. The result of mutation operator is a new offspring that will be tested in the fitness function to see the fitness value as a solution candidate from the given problem. The process of fitness, selection, crossover and mutation will be done recursively until one of the recursion control of the GA process fulfilled, which is iteration, convergence, or fitness value.

A. Multi-Objective Evolutionary Learning Schemes (MOELS)

The concept of Evolutionary Fuzzy System (EFS) has emerged since the early 2000s [6]. A study successfully adopted Multiple-Objective Evolutionary Algorithms (MOEAs) into Fuzzy Rule-Based Systems which yields MOEFSs [7] as a result. The use of MOEAs helped to optimizing the fuzzy set parameters, selection of rules and condition, or validating pre-defined partitions [4][8]. In short, MOEL uses chromosome coding in accordance with the type

of the FRBS, learning strategy, mutation, and crossover for generating offsprings. A candidate solution will then be evaluated through its accuracy on a training set, while its interpretability is evaluated by measured by the number of rules, or in this paper called Total Rule Length (TLR) [4].

B. SK-MOEFS' (2+2)M-PAES Algorithm Test

Accuracy test on dataset is done using SK-MOEFS' Memetic-Pareto Archived Evolution Strategy (M-PAES) [4] algorithm. While using Triangular Fuzzy Set, the Fuzzy Rule-Based Classifier (FRBCs) is generated with SK-MOEFS' Pareto Archived Evolution Strategy-Rule and Condition Selection-Fuzzy Decision Tree (PAES-RCS-FDT) [5][8]. The fuzzy sets MEDIAN solutions are shown in Triangular Sets. Each partition is labeled as Very Low, Low, Medium, High, and Very High. With by using Cryptocurrency Price History dataset, this paper aims to classify whether the current date signals a Bearish Market (Class 1) or a Bullish Market (Class 2). As stated in the Introduction, this paper will only test validity and ease-of-use of the SK-MOEFS and not about the validity of said algorithm used or any algorithm available in the SK-MOEFS toolkit. The test is done using hardware and software specification as stated in Table 1 below.

Using the specified hardware and software, the test is done using the default parameters available in the SK-MOEFS toolkit [9]. Said parameters have been altered in order to accommodate the hardware specification. While Probabilities for Operators has been modified to accommodate dataset values. These alterations have been kept minimum in order to test the default capabilities of the algorithm and Tool for users who has no knowledge in optimizing input parameters. Table 2 shows the parameters used for M-PAES Algorithm in the SK-MOEFS implementation.

TABLE I. HARDWARE & SOFTWARE SPECIFICATION USED IN TESTING

Type	**Specifications**	
	Brand and Type	***Details***
Processor	Ryzen 3 3100	3.60GHz, 4 Cores 8 Thread
RAM	-	2x8GB DDR4 3000MHz
Python	-	Version 3.8
SK-MOEFS	-	Version Commit db4fc89 2 Jul 2020

TABLE II. PARAMETERS USED IN TESTING

N_{val}	**Total number of fitness evaluations done**	**10000**
AS	**Archive Size for the M-PAES algorithm used**	**32**
M_{max}	**Maximum number of rules on each RB**	**10**
P_{Cr}	**Operator CR crossover probability**	**0.001**
P_{Ct}	**Operator CT crossover probability**	**0.5**
P_{MRB1}	**Operator CR first mutation probability**	**0.001**
P_{MRB2}	**Operator CR second mutation probability**	**0.7**
P_{MT}	**Operator CT mutation probability**	**0.2**
T_{max}	**Maximum number of fuzzy sets for each attribute**	**5**

V. RESULTS

The result shown on Fig.1 and Fig.2 are achieved by running the SK-MOEFS toolkit using the parameters described on Table 2. Although test resulted in alarming accuracy, but it is proven that the use of linguistic rules as explained in [10], will allow users to easily understand how a decision/classification is made. The Fuzzy Sets is shown on Fig.3, Fig.4, Fig.5, Fig.6, Fig.7, and Fig.8 and Rule-Base is shown on Table 3.

Train and Test results are displayed within a graph that is relatively easy to use. Fuzzy Sets are displayed with a standard Triangular Membership graph, it serves a detailed information about the Fuzzy Sets for people who understand and would like to delve on how the Rule Base and Classification are inferred. And finally, we have the final Rule Base in linguistic format. The Rule Base along with the Classification result, displayed in linguistic form, will offer the ultimate readability for users with no prior knowledge in either Market Trading or Fuzzy Logic.

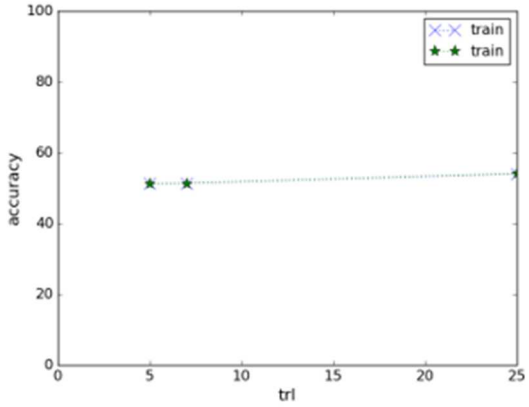


Fig. 1. Train and Test accuracy-TRL result

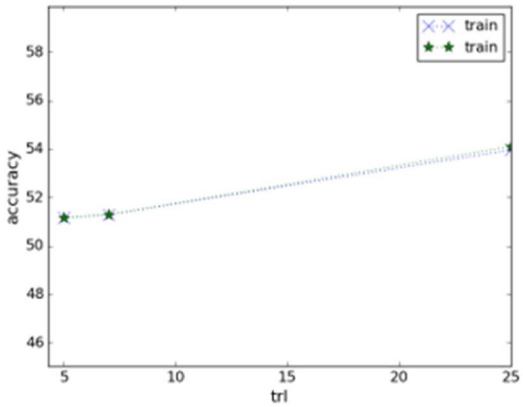


Fig. 2. Train and Test accuracy-TRL in Zoomed Result

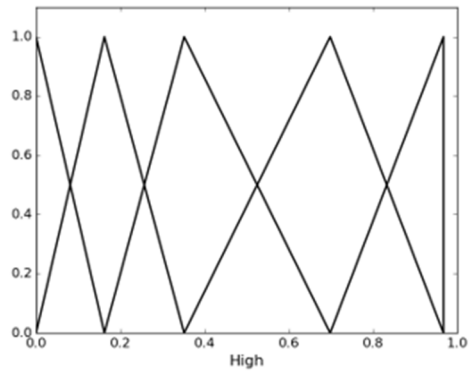


Fig. 3. Fuzzy Sets Label ‘High’

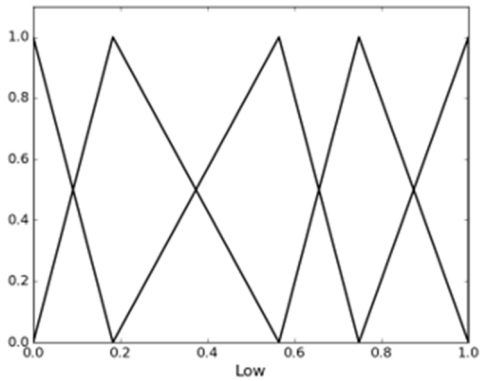


Fig. 4. Fuzzy Sets Label ‘Low’

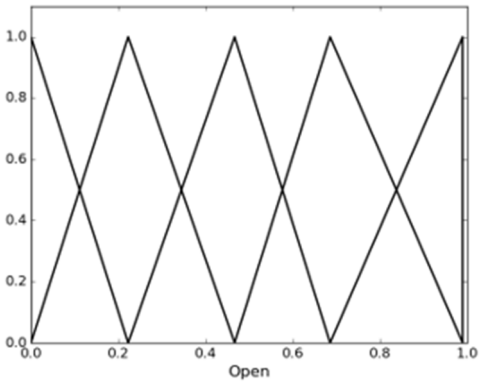


Fig. 5. Fuzzy Sets Feature ‘Open’

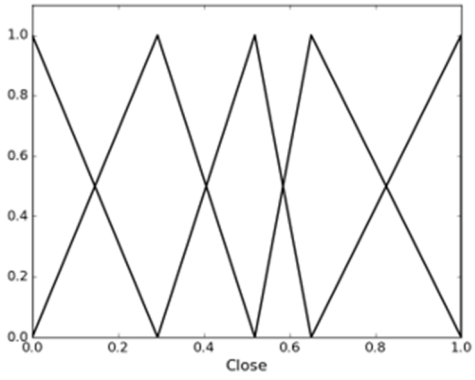


Fig. 6. Fuzzy Sets Feature ‘Close’

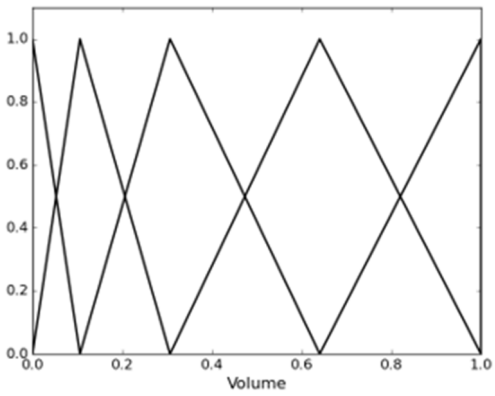


Fig. 7. Fuzzy Sets Feature ‘Volume’

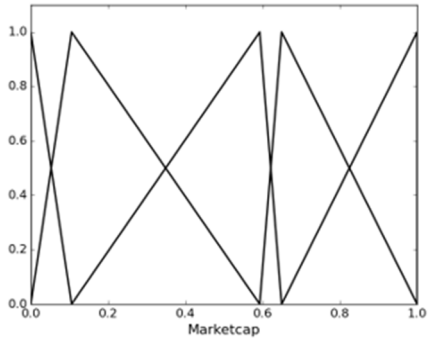


Fig. 8. Fuzzy Sets Feature ‘Marketcap’

TABLE III. RULE-BASE

	RULE BASE
1	IF Close is VL AND Volume is VL THEN Class is 2
2	IF Open is L AND Close is L THEN Class is 1
3	IF High is M AND Open is M AND Close is H THEN Class is 1

VI. CONCLUSIONS

The use of Fuzzy Rule-Based Systems (FRBS) as an easy-to-use and explainable AI model is validated in this paper with the help of a Python library called SK-MOEFS. SK-MOEFS toolkit, due to its open-source nature, offers easy expansion for custom methods. And by using the available Algorithms available in the SK-MOEFS toolkit, namely the (2+2) M-PAES Algorithm, we are able to produce prediction on a Cryptocurrency, in this case Ethereum Coin, based on its Price History. The result, as displayed on the previous chapter, yield more to be desired. However, due to testing it with bare default parameters, such result is acceptable. And this allows users without prior knowledge in AI Tools and Technical Analysis or Fundamental Analysis on Exchanges to act accordingly and responsibly. This paper also prove that the use of Fuzzy System is superior in term of readability. By utilizing SK-MOEFS, or Fuzzy System in general, with its Linguistic Rule-Base, users are able to understand and analyze the logic behind the prediction system. However, further tweaks and testing still need to be done with the parameters in order to achieve desirable results.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] P. Angelov, "Evolving Fuzzy Systems. Computational Complexity", 1053–1065. doi:10.1007/978-1-4614-1800-9_66, 2012

[2] E. Dourado, J. Brito, Cryptocurrency. The New Palgrave Dictionary of Economics. 10.1057/978-1-349-95121-5_2895-1. (2014).

[3] A.P. Faure, Foreign Exchange Market: An Introduction. (2013).

[4] L. Smigel, "What Is Open High Low Close in Stocks?" Analyzing Alpha, Analyzing Alpha, 17 Apr. 2020, analyzingalpha.com/open-high-low-close-stocks.

[5] Ducange, P., Mannara, G., & Marcelloni, F., Multi-objective evolutionary granular rule-based classifiers: An experimental comparison. 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). doi:10.1109/fuzz-ieee.2017.8015518, (2017).

[6] Knowles, J. D., & Corne, D. W. (n.d.). M-PAES: a memetic algorithm for multiobjective optimization. Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512). doi:10.1109/cec.2000.870313

[7] Gallo G., Ferrari V., Marcelloni FK., Ducange P. (2020) SK-MOEFS: A Library in Python for Designing Accurate and Explainable Fuzzy Models. In: Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1239. Springer, Cham

[8] Antonelli, M., Ducange, P., Marcelloni, F.: A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers. Inf. Sci. 283, 36-54. (2014)

[9] Fazzolari, M., Alcala, R., Nojima, Y., Ishibuchi, H., Herrera, F.: A review of the application of multiobjective evolutionary fuzzy systems: current status and further directions. IEEE Trans. Fuzzy Syst. 21(1), 45–65 (2012)

[10] Ducange, P., Marcelloni, F.: Multi-objective evolutionary fuzzy systems. In: Fanelli, A.M., Pedrycz, W., Petrosino, A. (eds.) WILF 2011. LNCS (LNAI), vol. 6857, pp. 83–90. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23713-3 11

The Effect of Educational Background on High Jobs and Income

Sopia Virgin[a1], Salma Luthfiyah Yulvi[a2], Shafira Khoirunnisa[a3], Linda Kurniawati[b4], Afrida Helen[ab5]
[a]*Department of Computer Science*
[b]*Research Center for Artificial Intelligence and Big Data*
Universitas Padjadjaran
Bandung, Indonesia
e-mail: [1]sopia19002@mail.unpad.ac.id, [2]salma19023@mail.unpad.ac.id, [3]shafira19012@mail.unpad.ac.id,
[4]linda.kurniawati@unpad.ac.id, [5]helen@unpad.ac.id

***Abstract*— In this era, people tend to assume that educational background is a string enough as consideration for opportunities, placements, and income of jobs. In fact, nowadays most companies also perceive that a good educational background will automatically bring out the maximum results for the company's performance. This refers to the assumption that the efficiency of workers is determined from the background itself, which also will boost the company's profit from human resources. However, to measure the level of influence of educational background on jobs and income, it needs a test that can map the percentage results accurately. Machine learning is an approach of AI (Artificial Intelligence) that is often used to help human problem solving or perform automation. Machine learning requires data to be studied, then it will classify the data based on how humans distinguish an object. Therefore, to complete the test above, it is necessary to use machine learning, namely KNIME. There are 3 algorithms of KNIME that can solve the above problems, which are Decision Tree, Logistic Regression, and Random Forest. Then, the highest level of algorithm accuracy is Random Forest, with a percentage of 84.9%. The accuracy result shows that the effect of educational background on employment and high income is 84.9%. So, it is true that the effect of educational background is very high. Instead, it is not only determined the high employment and income, yet it is still influenced by 25.1% other factors, such as productivity, work experience, etc.**

Keywords—income, jobs, education, random forest, decision tree, logistic regression

I. INTRODUCTION

Education is a demand or obligation that needs to be filled by everyone. Within the education process, we can prepare ourselves to gain more practices and knowledge as a foundation of working. Moreover, education is a form of consumption from society. Education is obligated to teach optimally and professionally, both in quality and quantity. This maximization is an important factor in investment in human resources and is also expected to reduce inequality in labor income (Lucas, 1988) [1].

Income inequality due to the influence of education has been widely analyzed in the labor economics literature. The result of the related analysis has explained that this income is an income (pay-off) from school fees that have been spent. Someone's income is determined by the education that has already been achieved (Card, 1999; Heckerman et al., 2006) [2]. However, this literature is still needing more consideration and wide discussion, because it cannot be calculated only from the school fees. Instead of it also needs some other factors that possibly influence this.

In 2018, in America, there were about 70% of high school graduates continued their studies to the college level. And from the age range of 18 to 20 years, the percentage of applicants to college increased by 6% from 35% to 41%. This increase occurred from 2000 to 2018. In addition, between 2000 and 2018, the total number of applicants in post-degree granting institutions increased from 13.2 million students to 16.6 million students, and this increase if calculated was 26%. It has been predicted that undergraduate enrolment will increase to 17.0 million students by 2029. Then from 2000 to 2017, it was found that student spending on public schools increased by 24%, reaching $12,794. Furthermore, tuition fees from 1993 to 2016, increased by 8%, reaching $23,902 (influenced by inflation) [3].

In America, the income of people with college degrees has increased by 5%, starting in 2000. And the fact is that people with college degrees earn 67% more than people with high school diplomas [3]. According to Becker's Human Capital Theory, education is an investment that makes individuals become more productive. Within the productiveness, someone will find easier way to be hired as a worker and get a higher income. And the return on investment can be calculated from the income of education. Policymakers in the whole world also accept the assumption that education is a good thing, and some of them are committed to investing in human capital, proposing for increasing economic growth and prosperity, thereby reducing inequality. However, some economists have another opinion, which is people who have a fairly good economic background tend to invest more in education (Spence & Arrow, 1973) [4].

The use of machine learning is basically often to be applied for economic researches and also can be used for HRM. The human resource management system (HRMS) is an important tool of enterprise management that plays an important role in the development of enterprises. Human resources today need to step up because the expectations have risen. Before machine learning has come to the rescue, HR managed data in a manual and semi-automated manner. To create analytics, it is necessary to gather, store, and process data. All of the above needs to be done in a short period of time because the data would quickly become irrelevant as the situation is changing and the data needs updating. Machine learning is able to help people finish their tasks [5]. It can assist and find out how the educational background affects the future job and income. This analysis was conducted by using the Decision Tree, Random Forest, and Logistic Regression. Those algorithms were used to help us in observing the statistics of either high or low effects for data validation. It is measured by the coefficient level to decide the accuracy.

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

The classification algorithm can be used to process data so that it can provide a predictive result to solve problems regarding the link between educational background and high income. In addition, if using the right classification method, it can provide predictive results with high accuracy. There are several types of classification algorithms which of course have different prediction functions and results. Decision Tree, Random Forest, and Logistic Regression are the classification choices to solve this research problem because they are easy to use and also produce a high level of accuracy. In addition, the tree algorithms can be used for categorical and continuous data types.

II. METHODS

A. *Machine Learning*

Machine Learning is the science for getting computers to learn and act like humans do, and improve their learning over time in an autonomous fashion, by feeding them data and information in the form of observations and real-world interactions [6]. Machine learning describes the capacity of systems to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks .

B. *Decision Tree*

Decision tree is a method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes [7]. The C4.5 algorithm is one of the algorithms used in the decision tree. Furthermore, a decision tree can be used to explore data, finding hidden relationships between a number of potential input variables and a target variable. A decision tree can overlap, especially if there are many classes and criteria used. And also, this algorithm makes it possible for over-fitting when there is noise in the data, and the model becomes unstable due to small data variations. The evaluation parameter used in this algorithm is the accuracy value.

C. *Logistic regression*

Logistic regression is a predictive model used to evaluate the relationship between the dependent variable (target), which is categorical data with nominal or ordinal scale, and the independent variable (predictor), which is categorical data with interval or ratio scale [8]. This algorithm does a classification based on the occurrence of events in data. Logistic regression also maps the probabilities of all available variables and determines the targeted optimization function. In this analysis, we use Node Partitioning to separate training data and test data, then use Node Logistic Regression Learner to create a model using training data. The training data we use is Salary(int) and use prediction(salary) as test data with Node Logistic Regression Predictor. The evaluation parameter used in this algorithm is the accuracy value.

D. *Random Forest*

Random forest is an ensemble classifier, which constructs a group of independent and non-identical decision trees based on the idea of randomization [9]. The random forest algorithm more accurately estimates the error rate compared with decision trees. More specifically, the error rate has been mathematically proven to always converge as the number of trees increases. Each decision tree uses a random vector as a parameter, randomly selects the feature of samples, and randomly selects the subset of the sample data set as the training set. The evaluation parameter used in this algorithm is the accuracy value

III. METHODOLOGY

Classification analysis is a part of predictive analysis. Predictive analysis is a method used to make future predictions. Predictive analysis can be done to reduce the risk that occurs in the future, this analysis also detects fraud committed by detecting patterns carried out by criminals. With predictive analysis, we can estimate the value of the unknown variable and can use a model that relates between variables to determine the value of an unknown variable. Predictive analysis is divided into two, the one using data intervals named Regression Analysis and Classification Analysis using nominal data. Regression is divided by linear and polynomial methods, and Classification Analysis can be solved with Decision Tree, Naïve Bayes, Logistic regression, Support Vector Machine (SVM) method, and many more. T. Data on adult incomes in the United States in 2021 comes from Kaggle. The attributes are listed in Table I.

TABLE I DATA DESCRIPTION TABLE

No	Dataset	Description
1	Age	The age range of the US population who received education.
2	Work class	The type of work the US people do, such as private and employee.
3	Estwgt	Estimated weight of US residences annual income.
4	Education	Education degree obtained by US people, like Doctorate, Prof-school, Highschool etc.
5	Occupation	Type of job that US people do, such as sales, professor specialist, manager, etc.
6	Race	Race of each people such as white, Asian-pacific and others.
7	Sex	Gender of each data like female or male.
8	Capital-gain	The amount of capital gain obtained.
9	Capital-spent	The amount of capital spent by each data.
10	Hours-per-week	The total time spent for working by each data.
12	Native country	The country of origin by each data.
13	Salary	The total yearly income.

This study implemented three algorithms, namely Decision Tree, Random Forest, and Logistic Regression. Fig. 1 is the stages of the flowchart of problem-solving:

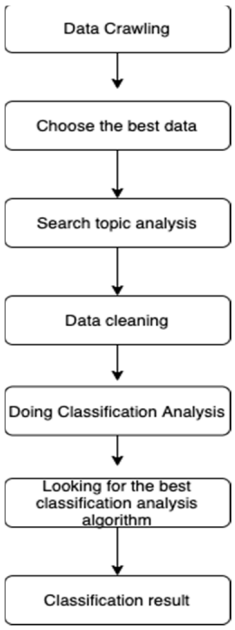


Fig. 1. Research Procedure

A. Data Cleaning

The first stage is data cleaning which is shown in Fig. 2, there are still many missing values in the initial data in each parameter, therefore we must clean the data first, to clean our data we using Knime, Through the nodes available in Knime we can easily clean up the data.

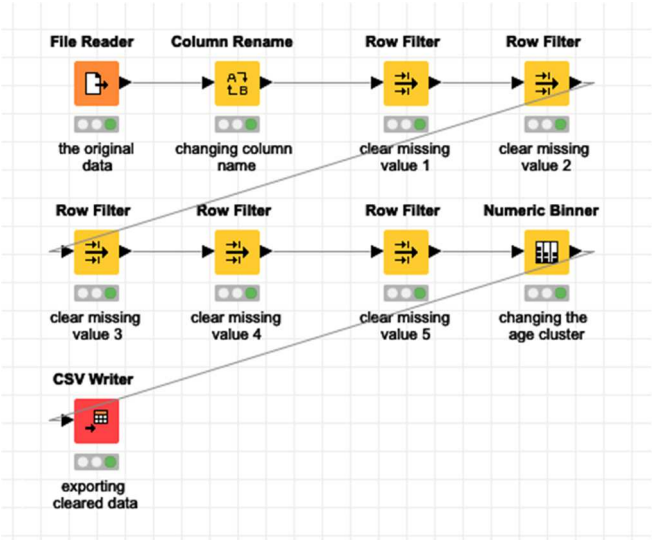


Fig. 2. Data cleaning workflow

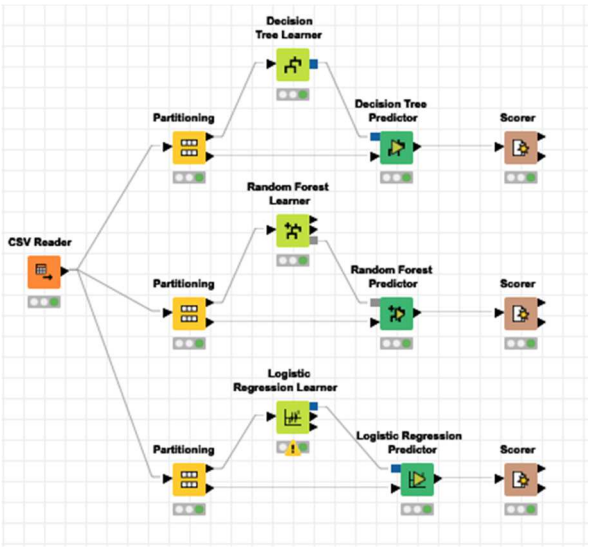
The first thing to do in data cleansing is to use the Node File Reader to read the file as a whole. After the data is read successfully, first look at which variables the missing value is located in. In this data, a lot of missing values are found, such as writing errors in attribute names, therefore, we must use Node Column Rename so that the error is successfully resolved. Every time we finish changing the settings on a node, we have to execute it so the node will work. After the attribute names have been successfully changed, we confirm the missing values in several variables by using the Node Row Filter, because there is more than one attribute that has a missing value, therefore we also use more than one Node Row Filter [15].

In the next step after removing the missing values, we provide age clustering by converting all ages 0-25 to young ages and also converting ages 26-100 to middle age by using the Numeric Binner Node on Knime. Furthermore, because the data is clean, the clean data is exported in the form of a CSV file using the Node CSV Reader, and the data is ready to be used for the next analysis process.

B. *Classification Analysis*

To solve the problem, we need to perform a classification analysis through related applications. This study implemented KNIME. Konstanz Information Miner (KNIME) is an application or platform that can perform various kinds of analysis, create reports and integrate data from any source[16]. It has nodes that can be easily used, besides that the analysis results from this platform are also easy to understand and the appearance or interface of the KNIME desktop version is also attractive and comfortable to use. The research scheme is illustrated in Fig. 3.

Fig. 3. Classification Workflow



C. *Comparison of Three Classification Algorithm*

We implemented some algorithms for the classification. The difference between them is described in Table II.

TABLE II COMPARISON OF THREE ALGORITHM TABLE

No	Methods	Data Type	Preliminary Process
1	Decision Tree	Polynomial , Numerical, Categorical [10].	Prepare the data, and select the attribute that will be the root for the tree. DT searching based on value comparison or grouping of each attribute[10].
2	Random Forest	Polynomial , nominal	Prepare the data for both regression and classification, and it is easy to see the relative importance it gives to input features.
3	Logistic Regression	Binary, nominal .	Classification based on the probability of occurrence in the logistic curve. Binary Logistic

No	Methods	Data Type	Preliminary Process
			Regression classifies data into two data classes

IV. Result and Analysis

A. Result from each Method

TABLE III RESULT TABLE

No	Methods	Acc (%)	Error rate (%)
1	Decision Tree Algorithms	80.019	19.891
2	Random Forest Algorithms	87,404	12.596
3	Logistic Regression Algorithms	65,909	34.091

Based on Table 3, Random Forest Algorithms has the highest level of accuracy that is equal to 87,404% with an error rate of 12,596%. The algorithm with the second-largest accuracy rate is the decision tree with an accuracy rate of 80.019% and an error rate of 19.891%. Meanwhile, the algorithm that has the lowest level of accuracy is Logistic Regression with 65.909% and an error rate is 34.091%. The three algorithms were tested using three variables, those are salary, education, and capital gain variables.

B. Comparison

After knowing the highest classification algorithm using salary, education, and capital gain variables, we conducted an experimental test by changing the capital gain variable into occupation one. Based on experiments conducted using different variables in Table V, the Random Forest Algorithms still have the highest level of accuracy rate that is 84,09% and 15,91% of error rate. Meanwhile, the same things happened with the previous experiment, Logistic Regression algorithms is still have the lowest level of accuracy rate, which is 74,909% and 25,091% of error rate.

TABLE IV COMPARISON TABLE

No	Methods	Acc (%)	Error rate (%)
1	Decision Tree Algorithms	78.417	21.583
2	Random Forest Algorithms	84.090	15.910
3	Logistic Regression Algorithms	74.909	25.091

V. Conclusion

The results of the analysis by conducting a test applied a decision tree, random forest, and logistic regression shows that the highest algorithm of accuracy rate is the Random Forest Algorithm, which is 87.404%. Meanwhile, the less effective algorithm to be used in solving this problem is the Logistics Algorithm Regression because it only gets a 65.909% accuracy rate with a fairly high error of 25.091%. According to our data, someone with higher education will potentially find a higher job level and earn bigger incomes than lower education.

It can also be concluded that there is a strong influence between educational background toward high employment and income. So, it can be reflected that the future of Americans will get a high income because they definitely have higher educational backgrounds, and are supported by the characteristics of Americans whose having more morals and motivation to reach their own dreams.

Acknowledgment

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

References

[1] J. S. Han, and J. W. Lee, "Demographic change, human capital, and economic growth in Korea", Japan and the World Economy, 53, 100984, 2020.

[2] F. Hampf, S. Wiederhold, and L. Woessmann, "Skills, earnings, and employment: exploring causality in the estimation of returns to skills", Large-scale Assessments in Education, 5(1), 1-30, 2017.

[3] J. Shaturaev, "Indigent Condition in Education and Low Academic Outcomes In Public Education System Of Indonesia And Uzbekistan", Архив научных исследований, 1(1), 2021

[4] C. Chadwick, "Toward a more comprehensive model of firms' human capital rents", Academy of Management Review, 42(3), 499-519, 2017.

[5] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[6] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019.

[7] B. Charbuty, and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning", Journal of Applied Science and Technology Trends, 2(01), 20-28, 2021.

[8] S. Nusinovici, et al, "Logistic regression was as good as machine learning for predicting major chronic diseases" Journal of clinical epidemiology, 122, 56-69,2020

[9] P. Probst, M.N. Wright,and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest" Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301, 2019.

[10] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm" International Journal of Advanced Computer Science and Applications, 11(2), 612-6, 2020

Wavelet Neuro-Fuzzy System (WNFS) in Predicting the Exchange Rate of the Rupiah Against the US Dollar

Asep Budiyana Muharam
Research Center for AI and Big Data
Department of Computer Science
Padjadjaran University
Jatinangor, Indonesia
asepbudiyana12@gmail.com

Muhamad Farid Ridho Rambe
Research Center for AI and Big Data
Department of Computer Science
Padjadjaran University
Jatinangor, Indonesia
muhamadfaridrr@gmail.com

Muhammad Risqullah Sudanta
Research Center for AI and Big Data
Department of Computer Science
Padjadjaran University
Jatinangor, Indonesia
rsudanta20@gmail.com

Akik Hidayat
Department of Computer Science
Padjadjaran University
Jatinangor, Indonesia
akik@unpad.ac.id

***Abstract*- In In this digital era, online trading is increasingly in demand and is often the community's first choice for investment. One type of online trading that is often chosen is *Foreign Exchange* (Forex) which is a buying and selling transaction of foreign currency. Forex investment is a type of investment that could return (more *returns* but also has a high level of risk because the fluctuations in currency values are very dynamic and significant. Therefore, in this study, a prediction technique for the IDR/USD exchange rate will be built. to be able to make decisions in the investment step using a combination of models wavelet, neural networks, and fuzzy systems called the Neuro-Fuzzy Wavelet model. The process in the neuro-fuzzy wavelet model includes the decomposition of Discrete Wavelet Transform (DWT), Projection scale values resulting from discrete wavelet transformations and a significant number of lags become inputs for the neural network, perform backward elimination on the lag variable, cluster data using Fuzzy C-Means, and conduct learning using a neural network. Prediction results using Wavelet Neuro-Fuzzy in determining the prediction of the rupiah exchange rate against the dollar American shows an average accuracy of 95%.**

Keywords— Forex, Predictions, Neuro-Fuzzy Wavelets

I. INTRODUCTION

In today's digital era, online trading activities are increasingly in demand and often become the main choice in investment. Foreign Exchange (Forex) is one of the investment models that are in great demand by the public because they have a high rate of return [1]. Forex is the activities of buying and selling the currency of a foreign country for profit or return [1]. The very dynamic and significant movement of currency values from time to time makes people tempted to make investment transactions in this model [1]. There are now seven currencies that are the main currencies in the eyes of investors, including the US dollar (USD), euro (EUR), Japanese yen (JPY), British pound (GBP), Swiss franc (CHF), Canadian dollar (CAD).), and the Australian dollar (AUD).

The IDR and USD pair is a pair that is relevant to the general public in Indonesia, IDR or Indonesian Rupiah itself is the currency of the Republic of Indonesia, and USD or United States Dollar is the most influential currency in the world. This currency has the most prospects because it is the main currency in various transactions in the world. Investing in dollars also gives people the opportunity to get a high return because it has the opposite nature of inflation. With the high value of the promised return, this will also be directly proportional to the level of risk faced. This investment model is categorized as an investment with high risk, the flow of funds moves so fast so that if the investor makes the wrong decision, there is a 100% chance of losing the investment value. herefore, we need a forecasting technique that can predict the IDR/USD exchange rate so that investors can make the right decisions when to buy and sell. A good forecasting technique requires a method that can provide a high level of accuracy with a small error [2].

Neuro-fuzzy wavelet is a prediction technique that combines discrete wavelet transform model, Artificial Neural Network (ANN) model, and fuzzy logic. Discrete wavelet transformation is a method that decomposes discrete data into several new data that correlate with the previous data [9]. An artificial neural network is a system that has certain characteristics similar to biological neural networks. Meanwhile, fuzzy logic is a knowledge representation constructed using IF-THEN rules [7].

Therefore, the researcher feels that the right method in this study is to use a neuro-fuzzy wavelet model to estimate the exchange rate of the IDR/USD pair in the future so that it can be a reference for investors in determining decisions when to sell and buy forex. The use of this method is based on the level of accuracy of the forecasts or predictions produced, besides that, the WNFS method can analyze non-smooth functions. Thus, the WNFS method is very suitable for modeling fluctuating data.

II. LITERATURE STUDY

A. Fundamentals of Foreign Exchange

Trade between the major currency markets in the world is called foreign exchange or is commonly referred to as Forex [3]. The Foreign Exchange Market itself will continue to run at any time during the working day without stopping. The Foreign exchange itself is also one of the instruments that are often used by many people to become an investment [3]. To get the maximum profit required support in the form of the right analysis in to get an efficient profit. The most

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

widely used foreign exchange today is the US Dollar which is usually denoted by USD in the forex market.

Apart from being a medium of exchange in every buying and selling transaction, the foreign exchange also has various functions, including as an international medium of exchange, a tool for controlling exchange rates, an international payment instrument, and a tool to facilitate international trade. As an international medium of exchange, foreign exchange can be used to carry out buying and selling transactions carried out abroad and foreign currency can also be used to exchange goods and services located in other countries, such as imports, exports, and hiring the services of someone residing in another country. Overseas [3].

As a tool to control the exchange rate, foreign exchange can be used as a comparison value for any existing currency, usually, we know this term as the exchange rate. The exchange rate of a currency can weaken and also strengthen depending on the situation and conditions at that time, a forecast is needed to be able to find out the value of a currency exchange rate against a currency at a certain time. As an international payment instrument, foreign exchange functions almost the same as an international medium of exchange, in that this function is usually carried out by a country to pay the country's debt to the country it owes. As an international trading tool, forex is used for transactions carried out by someone if it is constrained by the existing currency because it is different. With foreign exchange, one does not have to worry when making transactions between countries and can do it smoothly without any problems [3].

Two types of foreign exchange exist, namely physical and non-physical. For foreign exchange in physical form, it is usually in the form of physical money issued by banks from each existing country. For non-physical foreign exchange, it can be in the form of demand deposits or also securities. In foreign exchange itself has been given an international standard for the codes of each currency of each country. For each currency in the foreign exchange market, each currency has a three-letter code with the first two digits of the code being the country of origin of the currency and one digit at the end representing the name of the currency.

B. Wavelet Transformation

Transform is a conversion function that can be used to divide a function or signal into different frequency components, which can then be studied according to their scale [4]. In simple terms, the wavelet transform is used to convert a function in the time domain into the frequency domain [4]. These changes are needed to facilitate the analysis to be carried out, in the field of signal processing, the changes are carried out on the signal and the system. A function converter that can be used to divide functions and signals into components with different frequencies is called a wavelet transform [4]. The components that have been changed by the Wavelet Transform can then be understood according to their respective scales [4]. The wavelet transform can be used to transform a function that has a time domain into a function with a frequency domain. The transformation that was done earlier needs to be done so that the analysis to be carried out can take place more easily. In the case of signal processing, the transformation will be carried out on the signal and the associated system.

Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform are the two types of existing Wavelet Transform, whereas the name implies Continuous Wavelet Transform is used for functions that have an origin in real numbers that are above the x-axis, while Discrete Wavelet Transform is used for a function that has an integer origin [4]. Because the data from the time series of the IDR to USD exchange rate has an integer origin, this study will use the Discrete Wavelet Transform.

Continuous Wavelet Transformation has a way of working through calculating the convolution of a signal to a modulation window at any time with a certain scale because this modulation window has an independent scale, it is usually referred to as a wavelet parent. In this transformation recognize the terms translation and scale, where the translation is where the modulation window is shifted along with the sending of the signal, this relates to time information, then scale relates to signal frequency where the scale that is not low has relevance to the overall information of a signal, while the scale is not high has a relationship with detailed information [4].

Compared to Continuous Wavelet Transformation, Discrete Wavelet Transformation is assumed to be relatively easy to apply. The basis of the Discrete Wavelet Transformation is a way to take the time and scale representation of a signal with digital filtering techniques. At first, the signal will pass through the filter high pass and low pass, then a portion of each output will be sampled. In its application, Continuous Wavelet Transformation *is* more widely used for scientific research, as opposed to *Discrete Wavelet Transformation* which is used for applications in engineering and computer fields.

C. Neural Network

A part of the human-made intelligence system that is used to process information designed by replicating how the human brain can solve problems is called a Neural Network or what we usually call an Artificial Neural Network [5]. To be able to solve these problems, a learning process by the Artificial Neural Network is needed which requires a change of synaptic weights [5]. Information is received at the input layer of the neuron. The neural network at another layer then receives input from the input layer as an output from information processing [5]. The information that was processed earlier will be input to the neurons in other layers and so on until finally, the information reaches the outer layer or the outer layer of an artificial neural network. The artificial neural network itself is inspired by how the ability of the neural networks in the human brain to work, where when humans are given a stimulus or the neural network is usually called input, then the stimulus is passed on through other neural networks that are in the human body until finally, the person can determine what he should do after getting the stimulus.

Some of the functions of an artificial neural network, among others:

1. To determine a pattern
2. To describe the received pattern from input to a new pattern on the output
3. To remember patterns that will be used Back to
4. describe similar patterns
5. Streamline the process of a problem
6. To predict a term,

In general, an artificial neural network has three main structures, i.e enter human-like dendrites, axons outcomes were similar as in humans, and also activation-like synapses in humans. An artificial neural network is formed from various networks that are connected by a direct connection. The connection from one network to another has a numerical weight in which the weight will determine the strength and also as an indicator of a relationship [5].

D. Fuzzy Set

A set *fuzzy* is a principle that forms the basis for the creation of logic *fuzzy* [6]. Each member in the set *fuzzy* has a certain degree of membership which in the process of determining it will be determined by the membership function [6]. The Membership function is a curve that shows the mapping of data input points into their membership value (degree of membership). One way that can be used to get the membership value is through the function approach. The classical set is previously known in advance generally distinguishes each member in the set with a value of zero or one. As an example of a set, namely, the set of living things where there is a set of animals, a set of one-celled creatures, a set of fungi, a set of plants, and also a set of bacteria. A fuzzy set or fuzzy set is an expansion of normal sets in which the existence of a variable is not only true or false but can also be true if the variable has a membership degree in the range 0 to 1 [7]. However, what distinguishes fuzzy sets and probabilities is that fuzzy sets provide a measure for opinions and decisions, where probability signifies a proportion to the frequency that something can be true in a period time [7].

E. Fuzzy Inference System

Fuzzy Inference System is a computational framework based on fuzzy set theory, fuzzy rules in the form of IF-THEN and fuzzy reasoning. The fuzzy inference system consists of three processes, namely fuzzification, inference, and defuzzification. Fuzzification changes the input in the form of crisp into the form of fuzzy input in the form of linguistic values that are determined based on the membership function. The Inference is a process where the input that has been converted into fuzzy input is sent to a knowledge base that contains fuzzy rules to be defined beforehand so that fuzzy output or α-predicate is generated. Defuzzification is the last step in a fuzzy logic system where the goal is to convert each result from the inference engine which is expressed in the form of a fuzzy set to a real number.

F. Neuro-Fuzzy Wavelet

To be able to ascertain how the IDR exchange rate against USD in a certain period time, we need a forecast which in this study will use the Neuro-Fuzzy Wavelet method. Neuro-Fuzzy Wavelet itself is a combination of Wavelet Transform and Neuro-Fuzzy methods, which aims to combine the advantages of both methods and cover the shortcomings of each method with the advantages of other methods [8]. The initial process of the method is to manage changes to existing data, then determine how many input models will be entered later [9]. After determining the number of input models, the next step is to select the variables that will be used as inputs and outputs in the training. Next, group the data for training into each group. After being grouped into each group, then learning is carried out in the incident section by following the rules of fuzzy inference. Then the neural network will do learning which will then continue learning on the consequent part of the neural network. After learning is done [10], the results will be simplified with the backward technique, then a final output will be determined and calculate the error rate in the forecasting that has been done [11].

III. METHODOLOGY

A. Data Sources and Research Variables

In this study, the data used is secondary data sourced from Daily Updated Forex Data in Indonesian Rupiah per US Dollars. The data is daily data from January 1, 2021 to May 8, 2021. The amount of data used is 128 data with 2 columns, namely the date and the value of the rupiah.

B. Data Analysis Phase This

The research was conducted using *Google Collaboratory*, with the following computer specifications:

- Laptop Lenovo IP110-15ISK LMID
- Processor AMD A8-7410 Quad Core APU
- RAM 4 GB DDR3L
- Memory 1 TB HDD

From the diagram in Fig. 1 the steps of data analysis carried out are as follows:

- Inputting exchange rate data.
- Creating a Neuro-Fuzzy System (WNFS) Wavelet program.
- The processing

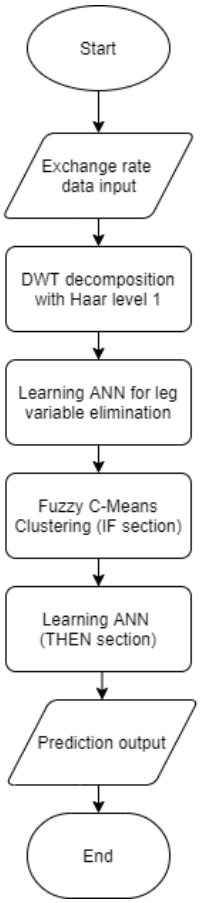


Fig. 1. WNFS Process Flowchart.

The preprocessing steps carried out in the WNFS program are as follows:

- Decomposing Discrete Wavelet Transform (DWT) using Haar level 1.
- The projected scale value of the discrete wavelet transforms, and a significant number of lags are input for the neural network. In this case, the significant lag is shown by the graph of the Autocorrelation Function (ACF).
- Performing backward elimination on the lag variable used.
- Clustering the data with Fuzzy C-Means then converting the results of the clustering to fuzzy sets 0 and 1 (IF part).
- Conducting learning using a neural network (THEN section).
- Calculates the final output to make predictions.

IV. RESULTS AND DISCUSSION OF THE

A. Wavelet Process

1) Comparison of DWT Results.

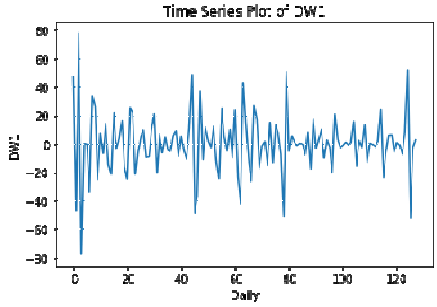


Fig. 2. Time Series DW1 Results from DWT Haar Level 1.

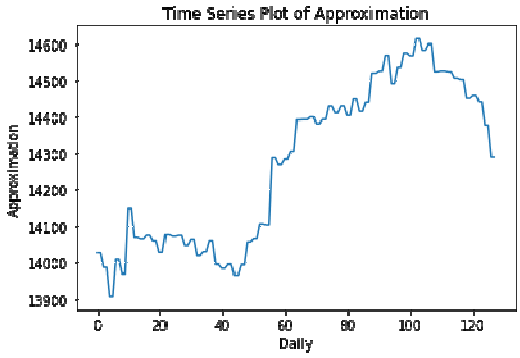


Fig. 3. Time Series Approximation of DWT Haar Results Level 1.

To determine the appropriate input for the Neuro-Fuzzy model in this study, the researchers compared the correlation coefficient between DW1 and *Approximation* with the actual data. The results of the calculations *approximation* and DW1through DWT are shown in the form of plot graphs in Fig.2 and Fig.3. From Table 1 the most significant correlation coefficient is found in the data *approximation*, so it will be used as input for the Neuro-Fuzzy model.

TABLE I. CORRELATION COEFFICIENT BETWEEN DW1 AND APPROXIMATION WITH ACTUAL DATA.

DWs	CORRELATION COEFFICIENT
DW1	0.0972
APPROXIMATION	0.9953

.

2) Determination of Many Input Variables.

Furthermore, determining the number of input variables used can be seen from the number of *lags* that exceed the significance line limit on the ACF.

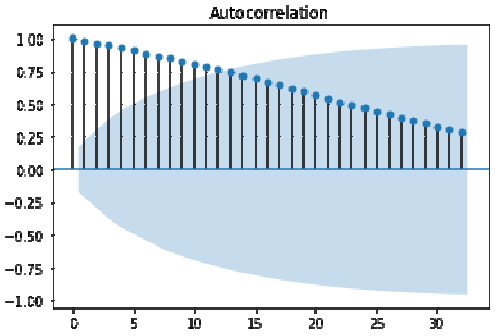


Fig. 4. ACF diagram on Approximation data.

Based on Fig. 4 above, 12 lags come out of the significance line, so the model to be built uses 12 input variables with 116 data pairs.

B. Neuro-Fuzzy System Process

1) Parameter

TABLE II. ZERO ORDER NFS MSE TRAINING WITH 12 INPUT VARIABLES

Many Neurons	MSE	Many Neurons	MSE
1	0.01686	11	0.00384
2	0.00754	12	0.00492
3	0.00657	13	0.00549
4	0.00723	14	0.00456
5	0.00511	15	0.00413
6	0.00520	16	0.00575
7	0.00514	**17**	**0.00376**
8	0.0051	18	0.00586
9	0.00466	19	0.00518
10	0.00493	20	0.00444

Parameters used to build the model *neural network* that will be used for elimination *backward* of input variables is using a backpropagation algorithm with one *hidden layer*, a maximum epoch of 1000, a *learning rate of* 0.001, and an error tolerance of 10^{-7}. The MSE value in each experiment the number of neurons in the *hidden layer* can be seen in Table 2. The smallest MSE value is when using 17 neurons, so the model to be built uses 17 neurons on *hidden layers*.

2) Comparison of MSE Values

Then, this study compared the MSE values if a variable was eliminated. MSE obtained at the time of variable x_9 eliminated quite small so removing x_9 will give a better model.

3) Clustering

The *clustering is* training data grouped into three clusters using the Fuzzy C-Means clustering method.

4) Determination of Membership Value

Each resulting *cluster* obtained, determined the value of membership in each *cluster* for all data. The result is the conversion of membership values from fuzzy sets to firm sets 0 and 1. The conversion results obtained are trained with a neural *network* to produce membership values for the antecedent part.

5) Learning with Algorithms

Then learning continues to the consequent section with the same algorithm.

6) Results

Graph of the suitability of the output with the target can be seen in Fig. 5 following. From this training, an accuracy value of 95% was obtained. So from the results of this study, it can be concluded that the Neuro-Fuzzy Wavelet can predict the IDR exchange rate against USD

TABLE III. MSE EXPERIMENTS ELIMINATION

Variables are eliminated	MSE	variables are eliminated	MSE
-	0.00458	x_7	0.00401
x_1	0.00447	x_8	0.00405
x_2	0.00393	**x_9**	**0.00367**
x_3	0.00387	x_{10}	0.00385
x_4	0.00472	x_{11}	0.00440
x_5	0.00378	x_{12}	0.00559
x_6	0.00382	-	-

.

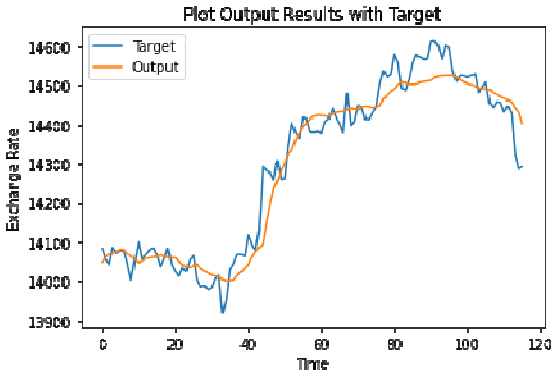


Fig. 5. Plot the WNFS Network Output Results with the Output Target.

V. CONCLUSIONS

Based on the analysis and discussion that has been carried out, the following conclusions can be drawn:

1. By using DWT Haar level 1, network learning using backpropagation, and using the sigmoid activation function, the neuro-fuzzy wavelet system built can be used to predict prices. IDR exchange rate against USD.
2. The wavelet transform data used as input is Approximation data which is the scale component of the DWT level 1 transformation, with the input variables used include $X_1 - X_{12}$
3. Prediction results using WNFS show an average accuracy of 95%.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] M. Bahrepour, "An adaptive ordered fuzzy time series with application to FOREX", Expert Systems with Applications, 38, 475-485, 2011

[2] A. Setiaji, "Application of wavelet-neuro-fuzzy model to predict exchange rate euro against the US dollar," PhD Thesis, UNY, 2014

[3] H. C. Berlianta "Getting to Know Foreign Exchange:, Yogyakarta: Gadjah Mada, 2005

[4] N. Dubey, and H. Modi, "A Robust Discrete Wavelet Transform Based Adaptive Watermarking Scheme in YCbCr Color Space against Camcorder Recording in Cinema/Movie Theatres" Engineered Science, 15, 116-128, 2021.

[5] Y Chihab,et al, "Algo-trading strategy for intraweek foreign exchange speculation based on random forest and probit regression. Applied Computational Intelligence and Soft Computing, 2019.

[6] T.J. Ross, "Fuzzy Logic with Engineering Applications (3 ed.), 2010.

[7] Sielvy E., "Application of Neuro-Fuzzy Model to Predict Gold Price," Thesis, UNY, Yogyakarta, 2013.

[8] S. J. Hadi, and M. Tombul, "Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study" Water Resources Management, 32(14), 4661-4679, 2018.

[9] H. Badrzadeh, R. Sarukkalige, and A. W. Jayawardena, "Intermittent stream flow forecasting and modelling with hybrid wavelet neuro-fuzzy model" Hydrology Research, 49(1), 27-40, 2018

[10] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[11] Nurmitasari, "Application of the Neuro-Fuzzy Wavelet Model to Predict Bengawan Solo River Flood". Numerical: Journal of Mathematics and Mathematics Education, Vol. 2, No. 1, 2018

Text Categorization of Job Vacancy Using Recurrent Neural Network Method

Sarah Hasna Azzahra
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
sarahhasna2210@gmail.com

Intan Nurma Yulita
Research Center for AI and Big Data
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
intan.nurma@unpad.ac.id

Akik Hidayat
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
akik@unpad.ac.id

***Abstract*— The need for information on job vacancies will increase over time. This results in information on job vacancies being updated very quickly and the data collected will be more numerous and varied. One way to group information is to tag each job vacancy. It can be possible that a job vacancy contains the characteristics of more than one category or is commonly referred to as multilabel This research aims to categorize job vacancies so that job vacancies can be grouped into appropriate categories automatically. The categories chosen to be categorized are Administration, Finance, IT, and Marketing. This research uses the Recurrent Neural Network method. In obtaining the optimal model, this research tested the hyperparameters. Testing on the Word2vec hyperparameter produces a dimension of 150 and a window size of 15. In the Recurrent Neural Network hyperparameter, the optimal model is obtained when the number of units in the Recurrent Neural Network is 200, the batch size is 32, and the threshold is 0.2. This research resulted in a Hamming score of 83.07% and an F1 score of 86.15%.**

Keywords*— Recurrent Neural Network, Multilabel, Category, Job Vacancy Text

I. INTRODUCTION

The development of technology towards all-digital is increasing rapidly, changing the lifestyle of humans that cannot be separated from electronic devices. The more sophisticated digital technology can make significant changes to the world and give birth to various kinds of increasingly advanced digital technologies. Technology has now become a tool that can help most human needs and has made it easier for various groups to access information in many ways. One of them is in the search for job vacancies [1]. In today's development of the digital information world, companies compete to meet the needs of human resources. Human resources in a company organization are important for companies to support progress, quality, and achieve company goals. In search of vacancies, usually job seekers search the print media, word of mouth information, or go directly to the company. This process certainly has drawbacks, these files can be exposed to water, torn, lost, or other things that we don't want. To overcome this, companies can't be far from technological advances because the recruitment process can be done through the digital world [2].

The current trend is website-based activities that can help human activities so that space and time are no longer obstacles in finding and getting the information needed with a relatively shorter time and relatively faster process [2]. The use of websites can help job vacancies reach a larger candidate population and candidates often explore several job vacancies on websites through job categories, so that job categorization becomes a challenge for job vacancies websites [3]. Automatic categorization can be understood as a learning process in which the program recognizes the characteristics that distinguish each category from other categories. Figuerola et al. suggest using an automated categorization system to increase success in information retrieval [4]. However, the categorization of job vacancies using Indonesian is still rare.

There are various ways to process data and perform categorization, one of which is using the algorithm found in Machine Learning. Machine Learning (ML) is a branch of computer science that examines how a machine can solve problems without explicitly programming [5]. The algorithm used for this research is the Recurrent Neural Network (RNN), which is one of the parts of the Neural Network. Recurrent Neural Network (RNN) is designed to process sequential data, where the current step has some kind of relationship with the previous step. Recurrent Neural Network (RNN) works well if sequential information is considered important such as job vacancies because meaning can be misinterpreted or grammar can be wrong if not using sequential [5].

II. LITERATURE REVIEW

A. Multilabel Classification

Multilabel classification is a classification to design a model that serves to carry out labeling on each label independently in a collection of many labels. In the past, multilabel classification was used for text categorization and medical diagnosis. Text documents usually have more than one conceptual class, such as articles about the reaction of the Christian church to the Da Vinci Code film that has been released which can be categorized as religion and arts. Currently, multilabel classification is increasingly being used for the needs of modern applications such as protein function classification and music categorization [6]. In general, multilabel classification is divided into two categories, Problem Transformation (PT) and Algorithm Adaptive (AA). The Problem Transformation method changes the classification task into one or more single label classifications or regression problems. While the Adaptive Algorithm adapts to the learning algorithm to be able to manage instances with multiple outputs or what is called multi-label. One of the algorithms included in the adaptive algorithm is neural networks [7].

B. Text Preprocessing

Text preprocessing is the stage to prepare text into data that will be processed in the next stage. At this stage, several things are done, including tokenizing, transforming case, filter tokens, filter stopword, and stemming [8].

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

1) Tokenizing

Tokenizing is the process of cutting the input string based on each word that composes it. Splitting sentences into single words is done by scanning sentences with white space separators.

2) Case Folding

Case Folding is a process to change the form of words, all characters are made into a lower case.

3) Filter Stopword

Stopword filter is the process of removing words that often appear but do not have any effect on text classification.

4) Stemming

Stemming is the process of changing the form of a word into a basic word. This method is a process of changing the form of the word into a basic word that adapts to the structure used in the stemming process.

C. Word2vec

Word2vec uses a large text set (corpus) as training data to build vocabulary and generates a vector space that can number several hundred dimensions, with each unique word in the corpus in the form of a vector where the vector formation applies the Skip-gram model and the CBOW model (Continuous Bag- of-Words) [9]. In the CBOW model, the model will predict one target word as output in the given context as input. While the skip-gram model is the opposite of CBOW model, the model will predict the context as the output of the target word as input [10]. A hyperparameter is something whose value and type are set by the researcher. The following is the hyperparameter of word2vec that will be used in this research.

1) Dimension Size

Dimension size hyperparameters are usually decided on a rule of thumbs, ranging from 50 to 300. We recommend that dimension sizes be selected based on corpus statistics. Dimensions that are too large will be very susceptible to overfitting because there are additional vectors that do not represent words. Likewise, too few dimensions will allow it to not adequately represent semantics [9].

2) Window Size

The window size is a hyperparameter to define the word around the target used for model training. The window size that is too small will cause the resulting word context to be less. Meanwhile, for the larger window size, more contexts of words are generated so that the probability of the pair of words will appear and the similarity value will be better. However, a window size that is too large does not always produce a good value because the larger the window size, the context of the word will be too much and cause the similarity value to be weak [11].

D. Recurrent Neural Network

Recurrent neural network (RNN) is an artificial neural network that repeats itself, which allows information to persist in the network and has feedback linkages to the network itself allowing activations to flow back in a loop of learning previous sequences and information. Recurrent Neural Network (RNN) is a category of deep learning because it has many layers, namely hidden layers, and can be applied to NLP research and text mining [12].

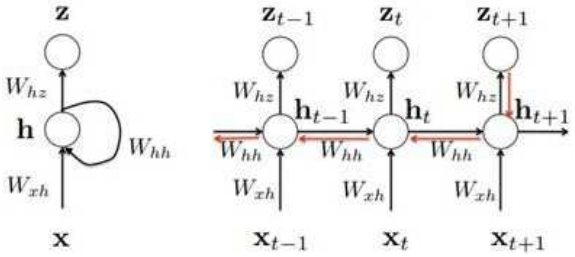


Fig. 1 Closed and Open Loop RNN

Fig. 1 is a visualization of an example of a snippet of an RNN which is a looping condition for its architecture where xt is the sequence and zt is the result of the output. The looping process allows information to be thrown from one step to the next. RNN models a dynamic system, where the hidden state does not only depend on the current state but also depends on the previous hidden state [13]. The following is the equation used in the hidden state:

$$h_t = tanh\,(W_{hh}\,h_{t-1} + W_{xh}\,x_t) \quad (1)$$

Where:
Ht = Hidden state at timestep t
Xt = Input at timestep t
Zt = Output at timestep t

The Recurrent Neural Network model by default uses the tanh activation function and the following is the prediction equation for each time step [14].

$$z_t = \text{softmax}\,(W_{hy}\,h_t + b_y) \quad (2)$$

Where:
Ht = Hidden state at timestep t
By = Bias
Zt = Output at timestep t

In RNN, hyperparameters are parameters used to control the learning process. The success of the RNN depends on the optimal hyperparameter. The following are the hyperparameters that will be used in this research.

1) Number of Units

The hidden layer is a set of units that have an activation function applied as well as an intermediate layer between the input layer and the output layer. If the data can be separated linearly, there is no need to use a hidden layer because the activation function can be directly implemented in the input layer. On the other hand, if the data cannot be separated linearly, then a hidden layer is needed. The number of units stored in the hidden layer needs to be taken into account. If the number of units is less than the complexity of the data, underfitting will occur. Underfitting occurs when there are too few units in the hidden layer to detect patterns in a complex data set. If the number of neurons is too much, there will be overfitting. Determining the number of neurons is very dependent on the amount of training data and the complexity of the data [15].

2) Threshold

In categorization with unbalanced classes, ROC is considered the gold standard of categorization ability. However, using only ROC in selecting a potentially optimal classification is not sufficient. Godbole's research [13]

proposed a new method to find the right predictive probability threshold from the test set. The experimental evaluation was carried out using predefined benchmarks and the results showed that the proposed method could effectively improve the prediction performance compared to the more commonly used methods.

Therefore, determining the right threshold is as important as the perfect ROC curve for making predictions. In most classifications, the prediction probability threshold has a default of 0.5. However, the threshold of 0.5 does not work well for the prediction of an unbalanced classification. Although researchers have attempted to increase the AUC value in previous studies, this investigation ignores the predictive probability threshold for the unknown test and data. As a result, classification performance, including recall, precision, and F-score, remains imperfect although the AUC score can be rather high [13].

3) Batch Size

Batch size is used to determine the number of observations or samples made before changing the weights, which will be distributed in a neural network. A batch size that is too large will result in a decrease in the quality of the model. This is because batch sizes that are too large tend to blend with the sharp minimizer of training and testing or can be called sharp minima leading to poorer generalization [16].

E. Metrics Evaluation

In multi-label classification, the predictions generated for each instance is a set of labels, therefore, the prediction can be completely true, partially true (with different levels of accuracy), or completely false. Evaluation for multi-label data can be grouped into three groups, namely, evaluating partitions, evaluating ranking, and using a label hierarchy. Evaluating partitions is evaluating the quality of classification into classes, evaluating ranking is evaluating whether the classes are ranked in order of relevance, and label hierarchy is evaluating how effectively the learning system can take into account the existing label hierarchical structure [17]. To obtain partially correct predictions, Sorower proposed a series of definitions for accuracy, precision, recall, and F1 size. Assume Y is the actual label and Z is the predicted label. Accuracy is measured by an asymmetrical Hamming score that measures how close T is to S [17].

1) Accuracy

Accuracy for each instance is defined as the proportion of labels predicted to be correct to the total number (predicted and actual) of labels for that instance. The overall accuracy is the average of the accuracy of all instances.

$$Akurasi = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (3)$$

2) Precision

Precision is the proportion of correctly predicted labels to the total number of predicted labels, averaged over all instances.

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|} \quad (4)$$

3) Recall

Recall is the proportion of correctly predicted labels to the total number of actual labels, averaged over all instances.

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad (5)$$

4) F1 Score

F1 Score is a weighted comparison of the average precision and recall.

$$F1\ Sore = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \quad (6)$$

As in the single-label multi-class classification, the higher the accuracy, precision, recall, and F1 scores, the better the learning algorithm's performance.

5) Hamming Loss

Hamming loss reports the number of times, on average, the relevance of an instance to the class label is incorrectly predicted. Therefore, the hamming loss takes into account incorrectly predicted labels and missing errors (relevant labels are not predicted), normalized over the number of classes and the total number of instances.

$$HL = \frac{1}{kn}\sum_{i=1}^{n}\sum_{l=1}^{k}[I(l \in Z_i\, l \wedge l \notin Y_i) + I(l \notin Z_i\, l \wedge l \in Y_i)] \quad (7)$$

I is an indicator function. The smaller the value of the hamming loss, the better the performance of the learning algorithm.

III. Analysis

A. Data Collection

The data used in this research is data on job vacancies using Indonesian which are obtained online. The data collected consists of 1,588 rows and 4 categories with different numbers of vacancies for each category.

B. Preprocessing

The data that has been collected is semi-structured and still has noise because there are still symbols, so the data that has been obtained must go through the preprocessing stage. The preprocessing stage has an important role because it can convert data into structured data and can have an influence on data quality for further processing. The following are the preprocessing steps carried out in this research in Fig. 2.

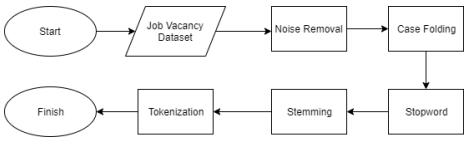


Fig. 2 Flowchart Preprocess

C. Training, Validation, and Test Data

Data are grouped into training data, validation data, and test data. The amount of data for model training is 80% of the total data, which is 1270 data. The training data is used for fitting the model to be trained, and the model will be validated using the validation data. In conducting model training, this research uses the 10-fold cross-validation technique. The model that has been trained will be tested using 318 test data, this aims to test the performance of the model that has been made.

D. Word2vec

Building the word2vec model uses the gensim library in the python programming language and uses the skip-gram algorithm. The skip-gram algorithm studies the probability distribution of words in context with a window that is set to predict the context as output and uses the target word as input.

E. Recurrent Neural Network

In this section, the calculation process performed by RNN is described. The number of the hidden layer is 1 and the activation function used is tanh.

IV. Results

A. Hyperparameter Tuning

The role of hyperparameters is very important because hyperparameters directly control the training behavior of the algorithm. The selection of the appropriate hyperparameters will have an important impact on the performance of the model being trained to get optimal results. Therefore, it is necessary to test several hyperparameters Word2vec, the hyperparameters to be tested are window size and dimension size.

TABLE I. Results Of The Word2vec Hyperparameter Experiment Based On Hamming Score

Hamming Score	**Dimension**				
Window	***50***	***100***	***128***	***150***	***200***
2	79.09%	78.72%	78.93%	79.40%	80.19%
5	78.93%	79.98%	80.19%	80.66%	79.72%
10	80.50%	80.77%	82.08%	80.50%	80.97%
15	80.97%	82.39%	81.08%	82.70%	81.29%
20	80.29%	81.29%	81.40%	80.03%	80.82%

1) Dimension Size

In this test, based on Table 1 and Table 2, it can be seen that the dimension size of 150 is the optimal dimension in representing a word. Dimensions that are too large will be very susceptible to overfitting because there are additional vectors that do not represent words. Likewise, the dimensions that are too small will allow them to not adequately represent the semantics.

2) Window Size

In the window size test, based on Tables 1 and 2, it can be concluded that the most optimal result is to use a window size of 15 because it can capture the context of the word well. The window size that is too small will cause the context of the resulting word to be smaller so that it may be possible not to capture some of the contexts that are considered important. The window size that is too large does not always produce a good value, the more contexts the word will cause the similarity value to be weak.

TABLE II. Results Of The Word2vec Hyperparameter Experiment Based On F1 Score

F1 Score	**Dimension**				
Window	***50***	***100***	***128***	***150***	***200***
2	81.34%	80.97%	80.92%	81.66%	82.81%
5	81.13%	82.02%	82.39%	82.70%	82.08%
10	82.99%	83.07%	83.96%	82.39%	83.23%
15	82.91%	84.48%	83.18%	84.49%	83.34%
20	82.55%	83.44%	83.46%	82.29%	83.02%

3) Dimension Size

In this test, based on Table 1 and Table 2, it can be seen that the dimension size of 150 is the optimal dimension in representing a word. Dimensions that are too large will be very susceptible to overfitting because there are additional vectors that do not represent words. Likewise, the dimensions that are too small will allow them to not adequately represent the semantics.

4) Window Size

In the window size test, based on Tables 1 and 2, it can be concluded that the most optimal result is to use a window size of 15 because it can capture the context of the word well. The window size that is too small will cause the context of the resulting word to be smaller so that it may be possible not to capture some of the contexts that are considered important. The window size that is too large does not always produce a good value, the more contexts the word will cause the similarity value to be weak.

TABLE III. Test Results for Number of Units and Threshold

F1 Score	**Number of Units**				
Threshold	***50***	***100***	***150***	***200***	***225***
0.16	85.31%	85.69%	85.60%	85.34%	85.59%
0.19	84.82%	85.52%	85.28%	86.02%	85.74%
0.2	84.93%	85.47%	85.18%	86.15%	85.35%
0.25	85.27%	85.85%	85.18%	84.91%	85.53%
0.38	85.59%	84.33%	84.59%	85.12%	84.38%

TABLE IV. BATCH SIZE TEST RESULT

No	Batch Size	Hamming Score	F1 Score
1	16	81.97 %	85.01 %
2	32	83.07 %	86.15 %
3	128	82.75 %	85.68 %
4	256	82.76 %	85.86 %
5	512	82.18 %	85.57 %

5) Number of Units

Based on Table 3, the number of units of 200 is the most optimal number of units because the number of units is proportional to the complexity of the data. The number of units that are considered less or lower than the data complexity will result in underfitting. While the number of units that are too large can result in overfitting because the network has so much information processing capacity, the information contained in the training dataset is so limited that it is not sufficient to train all units.

6) Threshold

The final prediction result is determined by the prediction probability and the default threshold value is 0.5. If the prediction probability is more than or equal to 0.5 then it will be rounded up to 1, whereas if the prediction probability is less than 0.5 it will be rounded up to 0. However, the value of 0.5 is not always ideal in some cases, especially for datasets with unbalanced or imbalance classes. Therefore, the threshold value needs to be adjusted to get the best F1 score in improving model performance. Different threshold values will produce different F1 scores. In this research, testing was carried out on the threshold value from 0.01 to 0.99 to get the best F1 score. Based on Table 3, the threshold of 0.2 with 200 units has the highest F1 score of 86.15%. So this study uses a value of 0.2 as the threshold value.

7) Batch size

This research tested the hyperparameter batch size with sizes 16, 32, 128, 256, and 512. Table 4 is the result of testing the hyperparameter batch size, the highest Hamming score is found in batch size 32 so that batch size 32 is the optimal batch size. Using a batch size that is too large can reduce the quality of the model that has been made, so that batch sizes 64 and 128 in this test are considered too large. In this test, batch size 32 can achieve the best training stability.

V. CONCLUSION

A. Conclusion

This section contains conclusions and suggestions based on research that has been carried out, namely:

- The implementation of the Recurrent Neural Network (RNN) algorithm in categorizing job vacancies resulted in a Hamming score of 83.07% and an F1 score of 86.15% which can be seen in Table 4.
- The Word2vec model built in this research is the result of the best hyperparameter testing. The dimension size used is 150 because it is the optimal size in representing a word. While the window size used is 5 because it produces a fairly good word context which can be seen in Table 1 and Table 2.
- The Recurrent Neural Network (RNN) model that is built is the result of the best hyperparameter testing to produce an optimal model. The best model of Recurrent Neural Network (RNN) is obtained when using the number of units of 200 which can be seen in Table 4.3 because it is by the complexity of the data so that it can reduce the occurrence of overfitting and underfitting, and the threshold with a value of 0.2 produces the best F1 score which can be seen in Table
- Batch the optimal size is 32 which can be seen in Table 4 because it can achieve the best training stability.

B. Suggestion

The suggestions from the research that has been done so that it can be considered in further development are:

- At the data collection stage, it is expected to increase the number of datasets that have multilabel.
- In future research, it is expected to try other categories and increase the number of categories.
- Further research can try to use feature selection

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] P. G. Lovaglio, M. Mezzanzanica, and F. Colombo, "Comparing time series characteristics of official and web job vacancy data", Quality & Quantity, 54(1), 85-98, 2020.

[2] H Suswanto, et al, "Development of Mobile Academic Exhibition Information System to Support Achievement of Job Hiring Graduate Vocational High School", Journal of Physics: Conference Series (Vol. 1028, No. 1, p. 012080), 2018. .

[3] E. Malherbe, M. Cataldi, and A. Ballatore, " Bringing order to the job market: Efficient job offer categorization in E-recruitment", SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1101–1104, 2015.

[4] C. G. Figuerola, A. F. Z. Rodríguez, and J. L. A Berrocal, " Automatic vs manual categorisation of documents in Spanish. Journal of Documentation, 57(6), 763–773, 2001.

[5] J. Han, and M. Kamber, Data mining: concepts and techniques. In Morgan Kaufmann, 2005.

[6] T. Katte, " Recurrent Neural Network and its Various Architecture Types" International Journal of Research and Scientific Innovation, V(III), 124–129, 2018.

[7] N. I. Widiastuti, " Deep Learning - Now and Next in Text Mining and Natural Language Processing", IOP Conference Series: Materials Science and Engineering, 407(1), 2018.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, " Efficient Estimation of Word Representations in Vector Space" 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 1–12, 2013.

[9] C. Allen, and T. Hospedales, "Analogies explained: Towards understanding word embeddings", International Conference on Machine Learning (pp. 223-231), 2019.

[10] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, “Word2vec model analysis for semantic similarities in English words”, Procedia Computer Science, 157, 160–167, 2019.

[11] F. Herrera, F, Charte, A. J. Rivera, and M. J. del Jesus, “Multilabel classification: Problem analysis, metrics and techniques”, Multilabel Classification: Problem Analysis, Metrics and Techniques, 2017.

[12] P. R. Vlachas, et al, “Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics”, *Neural Networks*, *126*, 191-217, 2020.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016.

[14] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, “On large-batch training for deep learning: Generalization gap and sharp minima”, 5th International Conference on Learning Representations, 1–16, 2017.

[15] M. Zhang, and K. Zhang, “ Multi-Label Learning by Exploiting Label Dependency”, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 999–1008, 2010.

[16] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, “ Finding the Best Classification Threshold in Imbalanced Classification”, Big Data Research, 5, 2–8, 2016.

[17] M. Sorower, “ A literature survey on algorithms for multi-label learning”, Oregon State University, 2010.

Twitter Social Media Sentiment Analysis Of Telecommunications Service Provider Using Long Short-Term Memory Method

Fauzi Faruq Nabbani
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
fauzifaruqn97@gmail.com

Intan Nurma Yulita
Research Center for AI and Big Data
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
intan.nurma@unpad.ac.id

Ino Suryana
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
ino.suryana@unpad.ac.id

***Abstract*—As one of the telecommunication companies in Indonesia, Provider X needs to pay attention to the services that their products provide for their users. This needs to be done so that their consumers don't switch to a different provider. To get the answers or information regarding the improvement of services, companies can process or analyze the users' sentiment toward their products through the Twitter social media. The purpose of this research is to analyze the sentiments of Provider X's tweet data of 5.375 on December 24, 2019 to February 24, 2020. The collected data is classified into two classes, negative and positive. This research uses the Long Short-Term Memory method. Hyperparameter testing is done to get the most optimal model. The test produced Word2vec hyperparameter including dimensions of 50, window size 5, epoch 50 and learning rate 0.025. In LSTM hyperparameter, optimal conditions are obtained when the number of LSTM layers is one, five LSTM units, the activation function is sigmoid, the optimizer is Nadam, the recurrent dropout size is 0.2 and the batch size is 32. The LSTM model is evaluated using K- fold cross validation and produces an average accuracy of 97.53%, f1score 97.39% and a loss of 10.36%.**

Keywords—sentiment analysis, long short-term memory, telecommunication company, twitter, word2vec

I. INTRODUCTION

Twitter is a social media that is quite popular among the public, including Indonesia. Twitter is able to express writing in 140 characters, informal message format, a lot of text messages every day, comes from individuals with various backgrounds, and is easy to use. Sentiment analysis is computational research of opinions, sentiments, and emotions that are expressed textually. With sentiment analysis, a company or individual can obtain a public opinion about its products, images, and services. Sentiment analysis is an efficient alternative because they do not need to conduct expensive conventional surveys and focus group discussions. This study analyzes Indonesian-language tweets about the brand of telecommunications service providers in Indonesia, namely the company Provider X. Telecommunication service provider companies are competing to improve their services. One of them is Provider X. This company needs to pay attention to what are the advantages and disadvantages of its service products. Provider X must also pay attention to other companies as competitors so that consumers do not switch to other service providers. Therefore, to find out how the level of user satisfaction with Provider X, it is necessary to analyze the sentiment on Twitter of users who have provided opinions or comments on the company.

In classifying sentiment analysis problems, there are several methods or algorithms that can be implemented. One of them is machine learning. This technique has several methods including using the Multilayer Perceptron (MLP). However, MLP has shortcomings, because in the case of sentiment analysis the input data is sequential data and the input from MLP must have a fixed input length measure, while in sentiment analysis problems the input length measurement is not always the same. In addition, MLP does fast forward based on activation without saving feedback.

To deal with this MLP problem, there is another method, namely Recurrent Neural Network (RNN). This method is a deep learning architecture where the processing is called over and over again. The way RNN can store information from the past is by looping through its architecture, which automatically keeps information from the past stored. However, the RNN has a disadvantage when processing longer sequence conditions. The RNN has a short-term memory problem. When the RNN carries information from the initial time step to the next time step, the RNN can forget the information in a longer sequence.

To solve RNN problems, there is a development of RNN, namely Long Short-Term Memory (LSTM) which is able to overcome short-term memory problems [7]. LSTM is also used to solve vanishing gradient problems or situations where the gradient value is 0 or close to 0 with a gate mechanism and an exploding gradient problem where the gradient loss results in a very large update of the model weight during training [2]. Therefore, this sentiment analysis research uses the Long Short-Term Memory (LSTM) method. The input of this method is in the form of words that are converted or extracted in the form of numbers, so this research also uses a word embedding called Word2vec to prepare words or features that will be processed in training.

II. LITERATURE REVIEW

A. Sentiment Analysis

Sentiment analysis is a subjective text mining and analysis technology used to derive useful knowledge and information from texts, as well as to classify trends in text sentiment [10]. Sentiment analysis aims to classify the polarity of the text in a sentence, document, feature or aspect level. Then determine the opinion expressed in the sentence, document, feature or level of that aspect whether it is positive, negative, or neutral [2].

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

B. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) has been an important focus of research and development during the 1990s. This method is designed to study sequential (sequential) or time-varying patterns [6]. It is called a recurring neural network because the value of neurons in the previous hidden layer will be reused as input data. The use of neurons in the hidden layer will be stored in a layer called the context layer.

C. Long Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a development of the RNN architecture which shares a repeating network. LSTM was proposed in 1997 by Sepp Hocreiter and Jurgen Schmidhuber [4]. A computation unit of an LSTM network is called a memory cell or cell state. This is what differentiates LSTM from RNN. In RNN the network loop uses only one simple layer, namely the tanh layer. In each cell state, there is a computation process consisting of several gates, namely the forget gate, input gate and output gate [1].

Forget gate will read the input values xt and s-1. This gate is also known as the sigmoid gate which contains parameters and bias. These two parameters will be studied in the training process. The result of this gate process is a number between 0 and 1. The number 0 means that the information will be deleted or forgotten and the number 1 means that the information will be stored. This gate computation process can be calculated with (1).

$$f_t = \sigma(W_f . x_t + U_f . s_{t-1} + b_f) \quad (1)$$

The next step at the input gate is to determine new information that will be used in the cell state (Ct). This stage has two parts, namely the sigmoid layer and tanh. The sigmoid layer is used to decide the value to be updated. The tanh layer will create a new candidate value vector C̃t. This gate computation process can be calculated with (2) and (3). The two parts will be combined to make an update to memory or state.

$$i_t = \sigma(W_i . x_t + U_i . s_{t-1} + b_i) \quad (2)$$

$$c_t = tanh(W_c . x_t + U_c . s_{t-1} + b_c) \quad (3)$$

Next, change the memory state or cell state from Ct-1 to cell state Ct, calculated using (4).

$$ct = ft * ct{-}1 + it * ct \quad (4)$$

After updating the memory cell, the last stage is entering into the output gate process as a sigmoid gate which will determine the next hidden state (st).

$$ot = \sigma(Wo . xt + Uo . st{-}1 + bo) \quad (5)$$

To generate information from the hidden state (st) which will be forwarded to another node, it can be obtained from multiplying the sigmoid gate output (ot) with the new cell state (Ct) modified with tanh activation.

$$st = ot * \tanh(ct) \quad (6)$$

D. Word2Vec

Word2vec, introduced by Thomas Mikolov, is a neural network built on the CBOW or Skip-Gram architecture to reconstruct the linguistic context of words. Word2vec takes a large corpus of text input and generates a vector space, usually of several hundred dimensions, with each unique word in the corpus given a corresponding vector in space. To build a Word2vec feature model, there are three processes involved, namely a vocabulary builder, a context builder, and a neural network.

E. Cross Validation

Cross validation is a statistical method for evaluating and comparing learning algorithms by dividing data into two segments: one is used to study or train a model and the other is used to validate the model. One of the techniques of cross validation is K-fold cross validation, which breaks the data into k pieces of data sets with the same size.

In K-fold cross validation, the data is divided into k segments which have the same or nearly the same ratio. The data is training and validated k times with each iteration taking a different segment as test data or validation and the remaining k-1 segments as training data to then take the average value of the results of each iteration

F. Confussion Matrix

The confusion matrix is a method of evaluating the performance of the classification model in making predictions. This matrix contains information that represents the labels predicted by the model, and the actual labels. For example, there is a model predicting two classes, namely positive or negative words. Then there is a calculation of the confusion matrix, positive class classified as positive (true positive), negative class classified as negative (true negative), positive class classified as negative (false positive), and negative class classified as positive (false negative).

G. Hyperparameter

In RNN and LSTM, hyperparameter is a parameter whose value is used to control the learning process. The success of the RNN and LSTM depends on the optimal hyperparameter. Hyperparameters are variables that affect the model output. Hyperparameter values are not changed while the model is being optimized. The following is the hyperparameter used in the research.

1) Activation function

The activation function is a function used in LSTM to calculate the sum of weights and biases, which are used to decide whether a neuron should be activated or not. The activation function also helps the process of normalizing the output of each neuron to have a value in the range between 1 and 0, or between -1 and 1. Research using sigmoid activation. This function is a non-linear function. The input part is a real number that has an output with a value range of 0 to 1.

2) Loss function

Loss function is a function for calculating the losses associated with all possibilities generated by a model by calculating the difference in actual and predicted output [5]. In this study, using the binary cross entropy loss function. Binary cross entropy is used because the form of the classification is a binary classification problem with a target value of 0 or 1.

3) Optimizer

Optimizer is an algorithm or method used to change neural network attributes such as weight and learning rate to reduce losses from the training process [3]. There are several types of optimizers that are often used in research experiments including Adam, SGD, Nadam, RMSprop, Adagrad, and Adadelta.

4) Batch size

The batch size is used to determine the number of observations or samples made before changing the weight, which will be distributed in a neural network and determined relative to computer specifications [9]. For example, there are 10,000 data numbers and a batch size of 10, it will take 1,000 iterations for all data to be spread across the neural network.

5) Epoch

The iteration unit that is often used is the epoch. One epoch can be interpreted as the number of iterations carried out to train the neural network in one round of all training data [8].

III. METHODOLOGY

A. Object of Research

The case study of this research is Provider X. This provider is a leading provider of cellular telecommunications and digital data services in Indonesia. The data used in this research is data sourced from the social media Twitter of the company Provider X. The data used is public opinion about Provider X. The raw data from Twitter still has noise and there are many tweets that use slang words or the writings are not in accordance with the Kamus Besar Bahasa Indonesia (KBBI), so this research is needed to correct words or sentences that are not suitable.

B. Research Stages

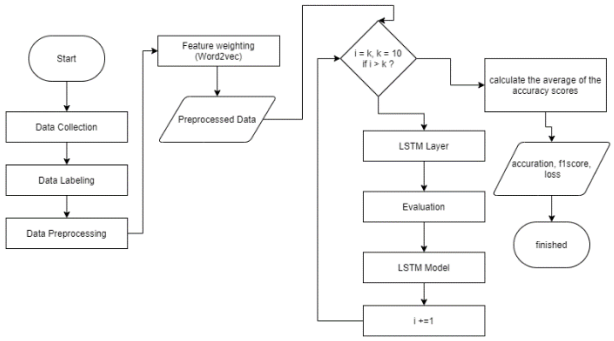


Fig. 1. Visualization of research stages

1) Data collection

The research stages are illustrated in Fig. 1. The data collected is text data in the form of tweets from Twitter social media with online crawling techniques using API from Twitter, namely by accessing the API key, API secret key, Access token and Access token secret. The crawling process is done by searching for the username keyword from the company Provider X, namely @Provider X.

Data collection or crawling was carried out starting from December 24, 2019 to February 24, 2020 and for the data used in building the model in this study, there were 5,375 tweets with two columns, namely tweet and date in the form of * .csv file format. From the 5,375 data, labeling is carried out into the positive sentiment class and negative sentiment class. The number of positive class data used in this study was 2,661 data while for the negative class amounted to 2,714 data.

2) Data labelling

Tweet data obtained from the crawling process that has been saved into a file format with the extension * .csv. Furthermore, the data will be given a label or labeling. Labeling data is by giving a score to each tweet which is done manually by the author. The score used in the data labeling process is by giving a score of 1 as a positive tweet and a score of 0 as a negative tweet.

3) Data preprocessing

This stage is very important to have an influence on the quality of the data that will be processed at the next stage. At this stage it will be cleaned of unnecessary components such as links, emoticons, and affixes. The stages taken in the data preprocessing process is explain as follows.

a) Case folding: Process changing tweet sentences that contain capital letters to lowercase. The purpose of case folding is to reduce data redundancies.

b) Noise removal: This process cleans the tweet data from unnecessary characters. These characters are in the form of punctuation marks, website addresses, numbers, too many spaces, and other characters besides the alphabet.

c) Stopword removal: This stage aims to clean or remove words that have no meaning and influence on the classification process or have no sentiment value.

d) Tokenization: The final stage in preprocessing is tokenization which separates or cuts tweet data in the form of phrases, clauses, or sentences into words or known as tokens.

4) Word2vec layer

The formation of this Word2vec model uses a library called gensim in the python programming language. The Word2vec training process in this study uses the Skip-Gram. The Skip-Gram algorithm is used to predict the context of a word by looking at the closeness of a word to another word whose position is before or after the word. So at this stage it will convert the word into a vector form that will be used in the LSTM input layer.

In building the Word2vec model, this research uses words that have been built previously which consist of features or vocab of more than one word using the Bigram technique. A feature or vocab that includes Bigram is a combination of two words that have the appearance of at least 5 times in the data. The number of words or vocab used in this Word2vec model is 34,744 phrases out of a total number of words of 53,539.

5) Making LSTM models using keras

This research built an LSTM model with a library from the Python programming language, namely Keras, and the evaluation of the LSTM model was carried out using K-fold cross validation with a value of k, namely 10. The following are the steps for building an LSTM model:

a) Define the model: This stage will define a network architecture consisting of a sequential layer, an embedding layer, an LSTM layer, a dense layer, and an activation function layer.

b) Compile network: To train the model, it is necessary to compile by defining the loss function, optimizer and the metrics used are accuracy and f1score.

c) Fit the model: After defining and compiling, then fitting or training the model on the training data by adjusting the hyperparameters, that epoch and batch size.

d) Perform network evaluation: The model that has been trained can then be evaluated using data validation to produce an evaluation score in the accuracy metric and f1score.

e) Saving the model: the model that has been created will be saved in the form of * .H5 file format, which can be loaded back to perform sentiment predictions and evaluate predicting data using confusion matrix techniques.

To determine the optimal hyperparameter, this research conducted some tests on hyperparameter Word2vec (window size, and dimensions), the number of layers LSTM, LSTM unit, optimizer and dropout. After getting the model, the best model will be selected to be evaluated using confusion matrix techniques, against predicting data (data that is not included in the training process or real data)

IV. RESULT

A. Word2vec Hyperparameter Testing Analysis

In building a Word2vec model, there are hyperparameters to be tested, including dimensions and window sizes

1) Dimension Size

TABLE I. THE TEST RESULT OF THE DIMENSION SIZE

Dimension	Learning Rate	Epoch	Accuracy (%)	F1score (%)
10	0.025	50	95.18	95.02
50	0.025	50	97.17	97.08
100	0.025	50	97.10	97.05
200	0.025	50	97.10	96.86

Dimensions of 50 provides the best accuracy and f1score values. The number of dimensions with a size of 50 is the optimal vector size that represents a word in the data. Table 1 shows that when using dimensions of size 50, dimensions that are too large result in the appearance of additional vectors but do not represent the data or may result in overfitting. In the small dimension, the meaning of a vector used does not show the meaning of the data or every word.

2) Window Size

TABLE II. THE TEST RESULT OF THE WINDOW SIZES

Dimension	Window	Learning Rate	Epoch	Accuracy (%)	F1score (%)
10	5	0.025	50	95.18	95.02
50	5	0.025	50	97.17	97.08
100	5	0.025	50	97.10	97.05
200	5	0.025	50	97.10	96.86

Table II shows the differences in the results of the accuracy score and the f1score that are not too far away. The most optimal results are when the window size is 5. When using window 5, this results in a fairly good blend of surrounding or context word predictions for the target word

B. Analysis of the Number of LSTM Layers

The next test is carried out on the layer architecture of the LSTM. The number of layers to be tested were 1, 2, 3 and 4 LSTM layers. Best results occur when using only one layer. Table III is the test results on the number of LSTM layers.

TABLE III. THE TEST RESULT OF THE NUMBER IN LSTM LAYERS

Layer	Unit	Optimizer	Batch Size	Dropout	Accuracy (%)	F1score (%)
1	2	Adam	32	0.20	97.23	97.09
2	2	Adam	32	0.20	96.82	96.68
3	2	Adam	32	0.20	96.61	96.39
4	2	Adam	32	0.20	96.22	96.09

Adding layers may not necessarily improve the performance or score of accuracy and f1score. In this research, using only one LSTM layer resulted in the best accuracy and f1socre scores compared to using other layer schemes. In general, adding layers to the neural network will increase the learning capacity of the model. This causes the model to find more complex patterns in the given dataset

C. LSTM Unit Analysis

Testing on this hyperparameter uses unit sizes of 2, 5, 10 and 50. Table IV shows the test results of unit changes in the LSTM layer, namely by trying several unit sizes gives results that have very little effect on changes in the accuracy score and f1score on each test. Based on the test results, 5 units have the highest accuracy and f1score and an increase in the previous test was 97.36% accuracy and 97.22% f1score.

TABLE IV. THE TEST RESULT OF UNIT CHANGES IN LSTM LAYERS

Layer	Unit	Optimizer	Batch Size	Dropout	Accuracy (%)	F1score (%)
1	2	Adam	32	0,20	97.23	97.09
1	5	Adam	32	0,20	97.36	97.22
1	10	Adam	32	0,20	97.25	97.04
1	50	Adam	32	0,20	97.25	97.10

D. Optimizer Analysis

Table V shows is the test results of the optimizer hyperparameter. The best results were shown when testing using the Nadam optimizer which gave an increase in the accuracy score and f1score from the previous hyperparameter testing, which had an accuracy of 97.53% and an f1score of 97.39%. Nadam optimizer that is easier to implement, computationally efficient, requires less memory, suitable when handling large amounts of data.

TABLE V. THE TEST RESULT OF UNIT CHANGES IN LSTM LAYERS

Layer	Unit	Optimizer	Batch Size	Dropout	Accuracy (%)	F1score (%)
1	5	Adam	32	0.20	97.36	97.22
1	5	SGD	32	0.20	97.08	96.90
1	5	Rmsprop	32	0.20	97.25	97.12
1	5	Nadam	32	0.20	97.53	97.39
1	5	Adadelta	32	0.20	97.19	97.07
1	5	Adagrad	32	0.20	97.08	96.94

E. Dropout Analysis

This research implements the use of dropouts in the LSTM layer by conducting experiments by testing the dropout sizes of 0.1, 0.2, 0.4 and 0.5. Table VI shows the results of the hyperparameter dropout test, and the best accuracy and f1score results are obtained when the dropout value is at a rate of 0.2. Based on the aim of the dropout which serves to improve the neural network by reducing overfitting.

TABLE VI. THE RESULT OF THE HYPERPARAMETER DROPOUT TEST

Layer	Unit	Optimizer	Batch Size	Dropout	Accuracy (%)	F1score (%)
1	5	Nadam	32	0.1	97.26	97.19
1	5	Nadam	32	0.2	97.53	97.39
1	5	Nadam	32	0.4	97.00	96.85
1	5	Nadam	32	0.5	96.82	96.58

Table VII shows the result of training to build a model with 10 k-fold cross validation. The 9th fold of the 10 folds shows the highest or optimal score when making the model

in, with accuracy of 98., 32%, f1score 98.28 and the smallest loss value is 0.0734

TABLE VII. THE RESULT OF TRAINING TO BUILD A MODEL WITH 10 K-FOLD CROSS VALIDATION

K-Fold	Accuracy (%)	F1score (%)	Loss (%)
1	97.03	96.89	15.45
2	96.84	96.72	13.57
3	97.40	97.18	9.80
4	97.03	97.03	12.26
5	97.77	97.33	11.32
6	97.58	97.17	9.26
7	97.58	97.49	7.86
8	97.58	97.68	8.86
9	98.32	98.28	7.34
10	98.14	98.08	7.89
Average	***97.53***	***97.39***	***10.36***

F. Testing with Predicting Data

The best model is used in testing predicting data. The predicting data amounted to 1,344 data. This test will use a confusion matrix technique. Table VIII is the result of the matrix. From this matrix, it produced an accuracy score of 97.54%, positive class f1score of 97.70% and negative class f1score of 97.50%.

TABLE VIII. THE TEST RESULT WITH PREDICTING DATA

Real Class	Prediction Result	
	Negative	***Positive***
Negative	646	12
Positive	21	665

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] J. Brownlee, "Long Short-Term Memory Networks With Python," Machine Learning Mastery, 2017.

[2] Dehaff, "Sentiment Analysis for Business, Finance, and Social Media," New York Symposium, February 2012.

[3] S. Doshi, "Various Optimization Algorithms For Training Neural Network," Medium.com, 2019.

[4] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory,"

[5] Neural Computation, vol. 9, 1997.

[6] K. Janocha, W. M. Czarnecki, "On Loss Functions for Deep Neural Networks," London: arXiv:1702.05659, 2017.

[7] L. Medsker, L. Jain, "Recurrent Neurl Network Design and Applications," CRC Press, 2001.

[8] A. Pulver, S. Lyu, "LSTM with working memory," International Joint Conference on Neural Networks (IJCNN), 2017, pp. 845–851.

[9] B. Santosa, A. Umam, "Data Mining dan Big Data Analytics," Yogyakarta: Penebar Media Pustaka, 2018.

[10] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining" ACM Computing Surveys (CSUR), 50(2), 1-33, 2017.

[11] Y. Zhang, W. Ren, and T. Zhu, "MoSa: A Modeling and Sentiment Analysis System," MDPI- Symmetry, 2019.

LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, And Random Forest

Moch. Lutfi
Master of Science Management,
Faculty of Economics
Padjadjaran University
Bandung, Indonesia
lutfimoch1403@gmail.com

Sheilla Putri Agustin
Master of Science Management,
Faculty of Economics
Padjadjaran University
Bandung, Indonesia
sheilla099@gmail.com

Intan Nurma Yulita
Research Center for AI and Big Data
Department of Computer Science
Universitas Padjadjaran
Bandung, Indonesia
intan.nurma@unpad.ac.id

***Abstract*— Researchers have carried out stock predictions in various ways. Stock is one type of investment that is quite attractive to investors. However, the risk of relatively large stock price fluctuations raises many questions among investors and researchers (academics) about how investors can predict stock prices in the future. There are two types of factors that affect stock price movements. There are external factors and internal factors. This study predicts stock prices using linear regression, SMO regression, and random forest algorithms. and compare the three models to find the most effective model in predicting stock prices. The results show that the most efective modeling in this study is using SMOReg which has the smallest error value among other models with a metric with an MAE number of 7.4758; MSE 157.0944; RMSE 12.5337; and MAPE 0.524. This means that the model is reliable enough to be used to predict future stock prices. However, SMOReg still has weaknesses in studying sequential data and causes overfitting.**

Keywords*— *Stock Price, Predictions, Algoritma linear regression, SMO regression, dan random forest

I. INTRODUCTION

Stock is one type of investment that is quite attractive to investors. However, the risk of relatively large stock price fluctuations raises many questions among investors and researchers (academics) about how investors can predict stock prices in the future. There are two types of factors that affect stock price movements. First external factors, such as economic factors and political factors. Stock prices reacted when Donald Trump was elected president of the United States [1]. The Second internal factor, which is caused by the company itself, is the fall in stock prices, which is influenced by the risk of financial distress [2].

Stock prices are challenging to predict, but many researchers have tried to predict stock prices in various ways. In 1988, White predicted changes in IBM's daily stock returns using NN [3]. Zhang combined the auto-regressive integrated moving average (ARIMA) model with artificial NN (ANN) to predict time series and conduct comparative studies. The results showed that ANN is more profitable in analyzing and processing nonlinear data [4]. Based on the theory of the relationship between volume and price, Wu et al. [5] conducted an empirical study of the relationship between volume and cost in the Chinese stock market using the conditional generalized exponential auto-regressive model and backward propagation NN. Li [6] predicts the error series using empirical mode decomposition with a support vector machine (SVM). Compared to conventional machine learning algorithms, deep learning (DL) algorithms c to solve nonlinear problems [7] more satisfactorily. The established DNN-based predictive models display a higher predictive capacity than other models [8].

Chen [9] uses the mean-variance model to predict stock prices which shows that the proposed method "eXtreme Gradient Boosting (XGBoost) with the improved firefly algorithm (IFA)" is superior to the traditional method (without stock prediction) and becomes a benchmark. measure in terms of return and risk. Cakra [10] predicts Stock Prices Using Linear Regression Based on Sentiment Analysis which results that the prediction model using previous stock prices and hybrid features as predictors provides the best predictions with coefficients of determination 0.9989 and 0.9983.

Based on this background, it shows that several different methods have been used by researchers to predict stock prices. This study aims to predict stock prices using linear regression, SMO regression, and random forest algorithms. and compare the three models to find a model that is effective in predicting stock prices.

II. LITERATURE REVIEW

A. Linear Regression

Linear regression method is a statistical tool used to determine the effect of one or several variables on one variable. The benefits of linear regression include regression analysis that is more accurate in conducting correlation analysis, because the analysis is difficult to show the level of change of a variable to other variables (slop) can be determined. With forecasting regression analysis or estimates of the value of the dependent variable on the value of the independent variable, it is more accurate. In addition, this analysis is to determine whether the direction of the relationship between the dependent variable is positive and negative and to predict the value of the dependent variable if the value of the independent variable increases or decreases and the independent variable. The data used is interval or ratio scale data [11].

B. SMO Regression

Sequential minimal optimization (SMO) is an algorithm for solving quadratic programming (QP) problems that arise during training of support vector machines (SVM). It was discovered by John Platt in 1998 at Microsoft Research. SMO is an iterative algorithm, which updates only two Lagrange multipliers at each step in a way that ensures convergence to the effective solution. For the derivation of the SMO

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

algorithm, form the following matrix of the convex non-smooth optimization problem [12].

C. Random Forest

Random forest is a classification and regression based method where there is a decision tree aggregation process [13] Random forest uses a number of Decision Tree classifications or decision trees. These trees are constructed using a random sample from the complete training data set, which results in potential differences between each tree, because the ranking of importance of features is different for different trees. Each tree is trained on an instance of the bootstrap training data and at each node the algorithm only searches in a random subset of the variables to determine the split. Bootstrapping is a common technique that iteratively trains training data and evaluates classifiers to improve their performance [14].

III. METHOD

Stock data from May 17, 2016 to May 11, 2021 (daily) consists of 6 attributes which are then taken 2 attributes for testing, because basically most of the data used in stock prediction research is date and close price, with a total of 1821 data. The data is divided into two parts, namely 70% for training data and 30% for test data. The data is taken from the site https://finance.yahoo.com/. Table 1 shows the data used.

TABLE I. DATA OF SHARE LQ45

Date	Open	High	Low	Close	Volume
17/05/2016	811	812	806	808	10.5406.500
18/05/2016	807	812	805	810	59.276.500
19/05/2016	810	811	800	802	26.878.200
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
07/05/2021	2.630	2.650	2.610	2.620	133.080.200
10/05/2021	2.640	2.700	2.620	2.680	157.887.500
11/05/2021	2.600	2.690	2.600	2.670	164.247.600

Fig. 1. is a chart of the LQ45 stock price for the period 17 May 2016 – 11 May 2021.

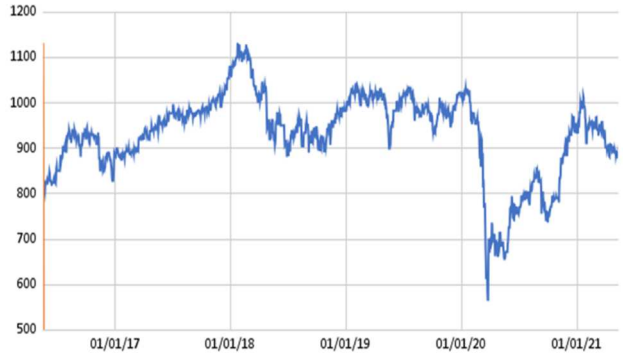


Fig. 1. LQ45 Stock Price Chart for the Period 17 May 2016 – 11 May 2021.

After data collection is done, then the next stage is the implementation of pre-processing of the data that has been obtained. At this stage, substitutions are made where there are empty dates because the stock market is only open on weekdays so holidays are filled in using the last day's close data. It is shown in Table 2 and 3.

TABLE II. CLOSE SHARE LQ45 DATA

Date	Close
05/17/16	808
05/18/16	810
05/19/16	802
05/20/16	805
05/23/16	812
05/25/16	819
05/26/16	821
05/27/16	826

TABLE III. PRE-PROCESSING DATA CLOSE SAHAM LQ45

Date	Close
05/17/16	808
05/18/16	810
05/19/16	802
05/20/16	805
05/21/16	805
05/22/16	805
05/23/16	812
05/24/16	806
05/25/16	819
05/26/16	821
05/27/16	826

IV. RESULT

Tests carried out consist of Comparison of Hyperparameters, Comparison of Models, Testing of Data Amounts and also Testing of Learning Rate. The models used for comparison include Linear Regression (LR), SMO Regression (SMOReg) and Random Forest (RF).

A. Hyperparameter

A hyperparameter is a variable defined by the model builder. Differences in hyperparameters have an impact on the treatment of the model during the training and prediction process. Hyperparameter testing is done to get the effective hyperparameter.

1. Linear Regression Hyperparameter

The hyperparameter used in this stage is the attribute Selection Method, to find out which Linear Regression is the effective. Table IV shows the results of the Linear Regression hyperparameter testing that has been carried out on the attribute Selection Method.

TABLE IV. HYPERPARAMETER LINEAR REGRESION

Attribute Selection Method	MAE	MAPE	RMSE	MSE
0	8,5562	1,0605	12,9858	168,6321
1	8,5562	1,0605	12,9997	168,9933
2	**7,5785**	**0,9328**	**12,6005**	**158,7725**

Based on the experimental results, the use of the attribute Selection Method with a value of 2 or the Greedy method provides effective performance with the lowest error value of MAE 7.5785; MAPE 0.9328; RMSE 12,6005; and MSE 158.7725. Greedy's algorithm uses a problem approach by finding the maximum value at each step (*step by step*), which

can be used to obtain optimum values and optimization problems. So that the hyperparameter attribute selection method is taken as much as 2.

2. Hyperparameter SMOReg

The hyperparameters used in this stage are epsilon parameter and tolerance, to find out which SMOReg is the effective. Table V shows the results of the SMOReg hyperparameter testing that has been carried out on the epsilonParameter and tolerance.

TABLE V. HYPERPARAMETER SMO REGRESION

epsilonParameter	tolerance	MAE	MAPE	RMSE	MSE
0,001	0,001	**7,7341**	**0,9044**	**12,4701**	**155,5042**
0,001	0,05	10,1260	1,2531	14,3910	207,0997
0,001	0,1	26,1224	3,1194	29,5602	873,8043
0,05	0,001	12,1082	1,4683	15,9012	252,8473
0,05	0,05	17,8090	2,1419	21,6681	469,5069
0,05	0,1	14,6747	1,7663	19,1099	365,1888
0,1	0,001	15,9148	1,8695	19,7980	391,9607
0,1	0,05	15,9429	1,9906	21,2197	450,2737
0,1	0,1	30,1929	3,7776	37,0539	1372,9879

Based on Table V, showing the experimental results, the use of epsilonParameter 0.001 and tolerance 0.001 provides effective performance with the lowest error value of MAE 7.7341; MAPE 0.9044; RMSE 12,4701; and MSE 155.5042. The epsilon value determines the level of accuracy of the model used, therefore it is necessary to choose an epsilon that provides a certain accuracy. From table V, the smaller the epsilon parameter and tolerance, the smaller the error value. However, epsilon with a value of 0, can cause overfitting. So that the epsilon hyperparameter is taken, the parameter is 0.001 and the tolerance is 0.001.

3. Hyperparameter Random Forest

The hyperparameters used in this step are bagSizePercent, numIterations and minimum variance for split, to find out which SMOReg is the effective. Table VI shows the results of the Random Forest hyperparameter testing that has been carried out on bagSizePercent, numIterations and the minimum variance for split.

TABLE VI. HYPERPARAMETER RANDOM FOREST

bagSizePercent	numIterations	minimum variance for split	MAE	MAPE	RMSE	MSE
80	100	0,001	**48,8544**	**6,6405**	**75,6892**	**5728,8542**
80	100	0,05	51,1847	6,9633	79,2023	6273,0016
80	100	0,1	53,5220	7,2656	81,8799	6704,3185
80	150	0,001	49,3019	6,7009	76,3012	5821,8720
80	150	0,05	50,9087	6,9244	78,7854	6207,1466
80	150	0,1	53,7551	7,2977	82,2178	6759,7587
80	200	0,001	49,2084	6,6896	76,2201	5809,5047
80	200	0,05	50,2688	6,8399	78,0054	6084,8428
80	200	0,1	53,3498	7,2445	81,7302	6679,8335
90	100	0,001	51,6044	7,0080	79,3422	6295,1907
90	100	0,05	53,3521	7,2492	81,9026	6708,0398
90	100	0,1	55,1598	7,4847	84,0008	7056,1273
90	150	0,001	50,8389	6,9045	78,2640	6125,2615
90	150	0,05	51,5136	7,0054	79,5765	6332,4239
90	150	0,1	53,8029	7,3043	82,2583	6766,4224
90	200	0,001	50,4144	6,8453	77,6534	6030,0552
90	200	0,05	51,3404	6,9818	79,3437	6295,4177
90	200	0,1	54,3064	7,3709	82,8917	6871,0410
100	100	0,001	53,6954	7,2884	82,0654	6734,7321
100	100	0,05	51,4928	6,9995	79,4265	6308,5658
100	100	0,1	49,6192	6,7401	76,6459	5874,5935
100	150	0,001	50,2035	6,8182	77,4017	5991,0197
100	150	0,05	51,4469	6,9922	79,3476	6296,0389
100	150	0,1	53,5251	7,2642	81,7882	6689,3083
100	200	0,001	51,5011	6,9900	79,0111	6242,7461
100	200	0,05	51,7457	7,0334	79,7832	6365,3574
100	200	0,1	54,0430	7,3339	82,4748	6802,0906

Based on the experimental results, the use of bagSizePercent 80, numIterations 100 and minimum variance for split 0.001 provides effective performance with the lowest error value MAE 48.8544; MAPE 6.6405; RMSE 75.6892; MSE 5728,8542. BagSizePercent shows how big the percentage of the training data set is. In the numIterations data (the number of trees in the random forest), according to Table VI the small iteration value indicates a small error value and the minimum variance for split indicates the minimum variance value from the train data, the smaller the variance value, the closer the data is to the expected value.

B. Model Comparison

This test is done to get the effective hyperparameter. Tests were conducted to find the mean absolute error (MAE), mean squared error (MSE), rooted mean squared error (RMSE) and mean absolute percentage error (MAPE). Table VII shows the results of the comparison of the performance of each model in predicting stock prices:

TABLE VII. MODEL TEST RESULT

Model	MAE	MAPE	RMSE	MSE
Linear Regression	**7,5785**	0,9328	12,6005	158,7725
SMOReg	7,7341	**0,9044**	**12,4701**	**155,5042**
Random Forest	48,8544	6,6405	75,6892	5728,8542

Each model has its own advantages and disadvantages. Based on Table VII can be seen that RF is not able to make predictions well, as evidenced by high error data. In general, the RF model is used for classification cases. But it can still be used in regression cases, including time series forecasting. In this case RF is not able to provide good performance, because the input data is too much. RF computing is not able to recognize data conditions that have a sequence, so it is difficult to predict future stock prices in Fig. 2.

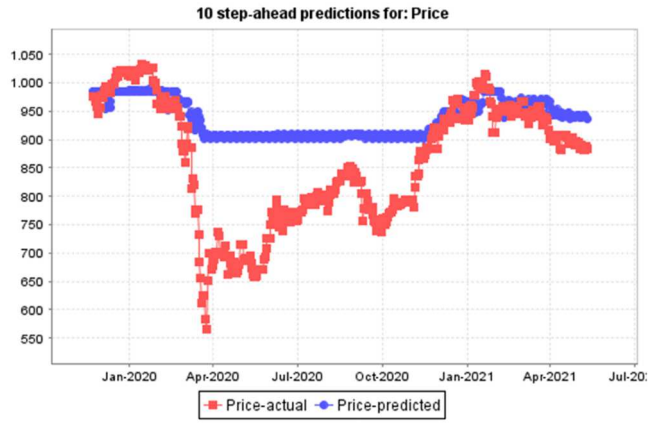


Fig. 2. RF Testing of Test Data

As the model with the simplest computation, LR provides a fairly good performance. However, this model has difficulty predicting non-linear data, as shown in Fig. 3, the problem to be solved is about the time series. When the time span used on the test data is getting longer, the prediction results are increasingly inaccurate. This happens because LR assumes that the attribute data used is independent and only

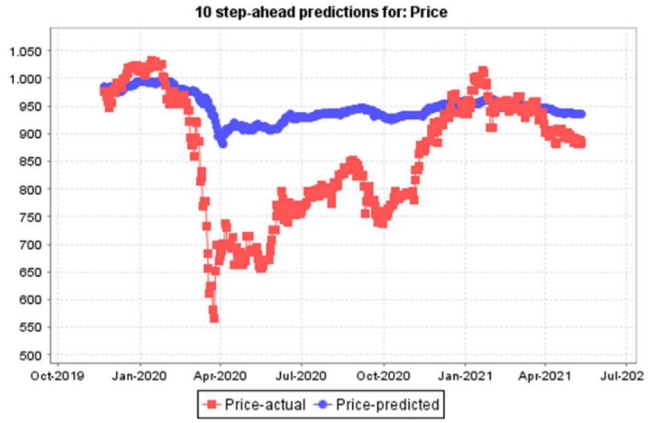


Fig. 3. LR Testing of Test Data

sees an increase in numbers (in this case the nth day) as a result of its prediction

The SMOReg model has the best ability in predicting the results of this study. As shown in Fig. 4, the prediction results can be done non-linearly by SMOReg. According to the prediction results of the training data, SMOReg gets a metric with an MAE number of 7.4758; MSE 157.0944; RMSE 12.5337; and MAPE 0.524.

Fig. 4. SMOReg Testing of Test Data

C. Model Test Result

In this section, LQ45 stock price predictions are made using the processed SMOReg model. The model provides output for the next 30 days. The results of price predictions are presented in Fig. 5 and Table VIII.

Fig. 5. LQ45 Stock Price Prediction for the Next 30 Days

TABLE VIII. LQ45 STOCK PRICE PREDICTION RESULTS FOR THE NEXT 30 DAYS

Date	Price
12/05/2021	882
13/05/2021	882
14/05/2021	882
15/05/2021	882
16/05/2021	881
17/05/2021	881
18/05/2021	880
19/05/2021	880
20/05/2021	880
21/05/2021	880
21/05/2021	880
22/05/2021	880
23/05/2021	879
24/05/2021	880
25/05/2021	879
26/05/2021	879
27/05/2021	879
28/05/2021	879
29/05/2021	878
30/05/2021	878
31/05/2021	878
01/06/2021	877
02/06/2021	877
03/06/2021	877
04/06/2021	877
05/06/2021	876
06/06/2021	876
07/06/2021	876
08/06/2021	876
09/06/2021	875
10/06/2021	875

These results indicate that from May 12, 2021 to June 10, 2021, the stock price continued to experience an insignificant decrease. According to these results, there will be a gradual decline but still at a minimum price of 875. Please note that the predictions of the SMOReg model are approximate value and cannot represent the actual stock price in the future market. Thus, this figure can only be used as a reference by investors to see the estimated condition of the stock market in Indonesia.

V. SUMMARY AND CONCLUSIONS

Based on the results obtained, the effective modeling in this research is using SMOReg which has the smallest error value among other models.

- The Linear Regression model provides an effective value by using the Greedy method. However, the longer test data causes the prediction results to be inaccurate because LR assumes that the attribute data used is independent and only sees an increase in numbers
- The Random Forest model is not able to provide good performance so that it is difficult to predict stock prices due to not being able to recognize data conditions that have a sequence, so that it is difficult to predict future stock prices.
- This means that the SMOReg model is reliable enough to predict future stock prices. However, SMOReg still has weaknesses in studying sequential data and causes overfitting.

ACKNOWLEDGMENT

The authors thank the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work is also supported by a Conference Grant funded by the Directorate of Research & Community Service, Universitas Padjadjaran, Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] A.F. Wagner, R.J. Zeckhauser, & A. Ziegler. Company stock price reactions to the 2016 election shock: Trump, taxes, and trade. Journal of Financial Economics, 130(2), 428-451. (2018).

[2] C.K. Andreou, P.C. Andreou, & N. Lambertides. Financial distress risk and stock price crashes. Journal of Corporate Finance, 67, 101870. (2021).

[3] White H. Economic prediction using neural networks: the case of IBM daily stock returns 451–458. (1988).

[4] Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175. (2003)

[5] Wu Q, Wang C, Tang Y. Empirical research on volume- price relationship based on GARCH models and BP neural net- work. J Sichuan Univ Nat Sci Edn 50(04):703–708 (in Chinese). (2013)

[6] Li X, Zhang Z. Support vector machine method for financial time series prediction based on simultaneous error prediction. J Tianjin Univ Sci Technol 47(01):86–94 (in Chinese). (2014).

[7] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 521 (7553):436. (2015)

[8] P. Yu, & X. Yan, Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6), 1609-1628. (2020).

[9] W. Chen, H. Zhang, M. K. Mehlawat, & L. Jia. Mean–variance portfolio optimization using machine learning-based stock price prediction. Applied Soft Computing, 100, 106943. (2021).

[10] Y.E. Cakra, & B. D. Trisedya. Stock price prediction using linear regression based on sentiment analysis. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 147-154). IEEE. (2015, October).

[11] D. Yendriani, M. Si, and U. N. Wisesty, "Prediksi Harga Saham Menggunakan Hidden Markov Model (HMM) dan Fuzzy Model," vol. 2, no. 2, pp. 6592–6599, 2014.

[12] Balasundaram S, Gupta D, Kapil (2014) Lagrangian support vector regression via unconstrained convex minimization. Neural Netw 51:67–79

[13] M. Dhawangkhara, E. Riksakomara. "Prediksi Intensitas Hujan Kota Surabaya dengan Matlab Menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya)". Jurnal Teknik ITS. 2017; 6(1): 94-99

[14] W. T. Aung and K. H. M. S. Hla, "Random forest classifier for multi-category classification of web pages," 2009 IEEE Asia-Pacific Serv. Comput. Conf. APSCC 2009, pp. 372–376, 2009.

Data Mining Implementation Using Frequent Pattern Growth on Transaction Data for Determining Cross-selling and Up-selling (Case Study: Cascara Coffee)

Aditya Rizky Fadillah[1], Intan Nurma Yulita [2], Aditya Pradana[3], Mira Suryani[4]
Department of Computer Science, Universitas Padjadjaran
Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran
Jatinangor, Sumedang 45363 Indonesia
Email : [1]aditya16016@mail.unpad.ac.id, [2] intan.nurma@unpad.ac.id, [3] aditya.pradana@unpad.ac.id, [4]mira.suryani@unpad.ac.id

Abstract— **The development and competition of the coffee shop business is increasingly popular nowadays. Cascara Coffee is one of the business shops that must have the right marketing strategy so that the shop's business can survive. This paper presents a study on the use of the FP-Growth association algorithm to process transaction data in order to provide best association parameter in cross-selling and up-selling recommendations for coffee sales. The data used is sales of Cascara Coffee in one year as much as 16,579. Based on the experimental results, the highest lift ratio value obtained is 2,789. In addition, the best association rule is to use a minimum support parameter of $0,20\ x\ 10^{-2}$ and a minimum confidence of 0.3. The association rules can be a recommendation for the company to carry out cross-selling and up-selling marketing strategies.**

Keywords—Association Rules, Cross Selling, FP-Growth, Lift ratio, Up Selling

I. INTRODUCTION

Cascara Coffee is a coffee shop that has been established since 2018. This shop is located in the West Bandung Regency area. The name Cascara Coffee is inspired by the coffee cherry skin which is a waste from processing to become coffee beans which can be reprocessed into something that can be consumed. Cascara Coffee offers coffee, tea and other side dishes.

But on the other hand, the development of business people in the coffee shop sector is growing rapidly. The proliferation of coffee shops is increasingly required to be able to face competition to attract consumers. The right marketing strategy and always developing the business is the right thing to keep the shop business afloat. Data related to sales can be used as a system to meet the needs of business decisions. One of the systems that is often used for business decisions is the way companies recommend what products are available and which are most often purchased together. With the formation of this system, the shop party can make marketing strategies that are right on target. To solve the problems that Cascara Coffee has, it requires methods and algorithms that can build a system.

Data mining is an activity to extract important information or knowledge from a large data set within certain technique [1]. Information or knowledge generated from data mining can be used to improve decision making [1]. Data mining is one of the fastest growing fields in the computer industry. Once a field of minor interest in computer science and statistics, it quickly developed into a field of its own. One of the greatest strengths of data mining is reflected in the various methodologies and techniques that can be applied to a number of problem sets including in the business field [4].

One of the tasks in data mining is association rules. Association rules or affinity analysis is about "what with what". The results of the association rule method will provide output in the form of knowledge of what menus are often ordered together [1]. In this study, one of the algorithms of association rules, namely Frequent Pattern Growth (FP-Growth) is used.

Frequent Pattern Growth (FP-Growth) is a method in data mining to find frequent itemsets without using candidate generation [2] After the system is successfully created, the results of the system can be used as a marketing strategy. One strategy that is suitable to be used from the system results is cross-selling and up-selling. Cross-selling and up-selling can convert into actual sales so as to increase profits [3]. For this reason, it is necessary to implement data mining with the FP-Growth algorithm to determine cross-selling and up-selling on sales patterns at Cascara Coffee.

II. LITERATURE REVIEW

A. *Preprocessing*

Preprocessing is the process of converting raw data into an understandable format. Real data is sometimes incomplete, inconsistent, redundant, and a lot of noise. This process is the first step to help convert raw data into a format that can be processed and utilized for further processing [2].

- Data Cleaning. Data cleaning is the process of detecting damaged and less accurate data from a collection of record data sets. The main use of the cleaning step is based on the detection of incomplete, inaccurate, inconsistent and irrelevant words so that the data can be modified or deleted if necessary [2].
- One Hot Encoding. One hot encoding is the most widely used encoding scheme. It compares each level of a categorical variable to a fixed reference level. Each of

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

these coding schemes converts each variable into a binary variable. Each variable is changed its value to (1) when it shows there is and (0) when it doesn't exist [5].

B. Frequent Pattern Growth (FP-Growth)

Association rules or affinity analysis in data mining is about "what with what" [1]. There are two key terms, namely support and confidence, which are used in calculating the correlation between variables [6]. Support is usually expressed as a percentage or can also be expressed as a number of occurrences. It is called support because it measures how much support it has for the validity of the developed rules. Meanwhile, confidence is the ratio between the number of transactions that include all items in the antecedent.

In this study, the FP-Growth algorithm is used as an important step in the association to see patterns that often appear in the dataset [2]. The FP-Growth works is how to find all items in the data set that meet the minimum support. Support itemset (I) is defined as part of the data set $T\{T1 \ ... \ Tn\}$ which contains I as a subset denoted as sup(I). when the support is equal to or more than the minimum support, then it is called a pattern that often appears.

$$Sup(A) = \frac{Total\ transaction\ contains\ Subset\ A}{Total\ Transaction} \qquad (1)$$

The total amount in the general pattern depends on the minimum level of support. The less the minimum support, the higher the value of the more frequent pattern. On the other hand, the higher the minimum support, the less frequent the pattern. After support is found, the next step is to look for association rules that meet the minimum requirements for confidence by calculating the confidence of association rules A then B. The confidence value of these rules is obtained from the following formula.

$$Confidence = \frac{The\ number\ of\ transactions\ contains\ A\ B}{Jumlah\ transaksi\ mengandung\ A} \qquad (2)$$

The FP-Growth algorithm is used to reduce the number of scans of the data set. FP-Growth only needs to scan the database twice. On the first scan, a database of frequent itemsets will be generated. In the second scan the itemset is used to generate an FP-Tree by filtering out rare items [7].

In detail, here's the FP-Growth mechanism and the illustration support by Fig. 1:

- Input the dataset is the first step of the FP-Growth algorithm stage as the data will be searched for patterns of association rules.
- Enter the minimum support and minimum confidence parameters to make the lower limit of the result of a rule.
- Calculation of the support value for items in each dataset line that meets the requirements for the minimum support.
- The formation of the FP-Tree begins with an empty root. The algorithm will scan transactions and enter them into the FP-Tree sequentially. If the itemset has a high support value, it will be at the top. When it has the same node it will increase the value of the FP-Tree but when it is different it will create a new branch.
- In the conditional pattern base search, the data is sorted from the lowest node. Then it will look for the path formed by the node
- Conditional FP-Tree¬ is formed by looking for a collection of items that have been formed from the conditional pattern base by showing the minimum support.
- The process of finding frequent patterns is a continuation of the Conditional FP-Tree. The minimum confidence value is used to determine the rules that will appear in the final result.

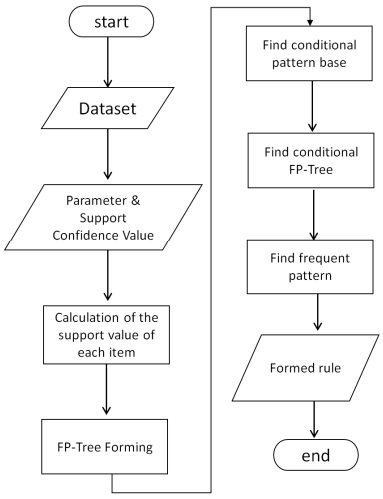


Fig. 1. Stages of the FP-Growth Algorithm.

C. Cross-selling and Up-selling

Cross-selling and up-selling can convert into actual sales thereby increasing profits. Cross-selling and up-selling are arguably the two most widely used strategies in marketing or selling activities. These two strategies used to be found in fast food restaurants and hotels, but along with technological advances, these techniques are also widely applied in online business and other business fields [3].

Cross-selling is called bundling. Cross-selling is about offering customers additional products or services to the basic product they purchased [9]. Cross-selling as a range-bound process designed to sell additional items that are different from those the customer has already purchased or has expressed an interest in buying before. Items sold can be products, services, or a combination of both [10].

In contrast to cross-selling, the up-selling technique involves increasing the volume of orders either by selling more units of the purchased item or upgrading to a more expensive version of the item purchased [11]. More simply, cross-selling is selling product A by offering products B, C, D and E together. While up-selling is the art of selling product A, and in order to increase sales, make it A+, A++ or A+++ usually done by adding the size from the menu.

III. METHODOLOGY

A. Research Methodology

The research stages as shown in Fig. 2 serves to describe the flow of the research starting from data collection and the process used in the system to determine the frequent item set or consumer habit patterns in buying products using the FP-Growth method which functions to calculate and determine the most frequent data set (frequent itemset). At the end of the stage there are evaluation activities carried out.

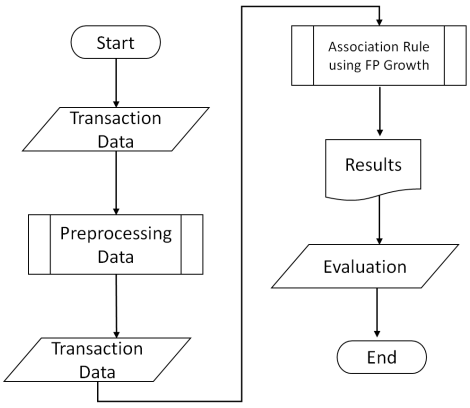


Fig. 2. Research flow to determine rule for cross-selling and up-selling recommendations using FP-Growth.

B. Data Acquisition

The transaction data collected is the result of transactions that occurred during January 1, 2020 to December 27, 2020 in Cascara Coffee. The amount of data obtained from the results of these interviews amounted to 16,579 data. The data obtained also has 7 features. It is listed in Table 1. Table II shows examples of transaction data obtained.

TABLE I. FEATURES IN DATASET

No.	Feature	Description
1.	*Outlet*	The Outlet feature is a store where transactions occur.
2.	*Date*	Date is an attribute of the date the transaction occurred.
3.	*Time*	Time is an attribute of the time the transaction occurred.
4.	*Receipt Number*	Receipt Number is a transaction code on the transaction that occurred.
5.	*Items*	Items is the name of the menu attribute ordered at the time of the transaction. There are 58 different menus on the Items attribute.
6.	*Event Type*	Event Type is an attribute of the event type. The contents of the Event Type attribute of this research data set are only "Payment".
7.	*Payment Method*	Payment Method is an attribute of the payment method on the transaction.

TABLE II. EXAMPLE OF TRANSACTION DATA IN CASCARA COFFEE

Outlet	Date	Time	Receipt Code	Items	Event Type	Payment Method
Outlet 1	9/12/ 2020	20:24: 38	6MRDOZ	Crème Brule, Vietnam Drip	Payment	Cash
Outlet 1	9/12/ 2020	19:41: 27	6MRDOY	Caffe Latte, Vietnam Drip, Red Choco Velvet	Payment	Cash
Outlet 1	9/12/ 2020	18:34: 56	6MRDOW	Cappuccin o, Caffe Latte	Payment	Cash
Outlet 1	9/12/ 2020	18:33: 53	6MRDOV	Crème Brule, Matcha Latte	Payment	Cash

C. Preprocessing

In this study, data cleaning and one hot encoding were used at the pre-processing stage. The data cleaning results is presented in Table 3.

- Data Cleaning. In this process the data cleaning uses the regex library. The word that is deleted in this preprocessing is the number of menus ordered. Then the data will be split one item for one row.

TABLE III. EXAMPLE OF PREPROCESSING RESULTS

Outlet	Date	Time	Receipt Number	Event Type	Payment Method	Items
Outlet 1	9/12/ 2020	20:24:38	6MRDOZ	Payment	Cash	Creme Brulee
Outlet 1	9/12/ 2020	20:24:38	6MRDOZ	Payment	Cash	Vietnam Drip
Outlet 1	9/12/ 2020	19:41:27	6MRDOY	Payment	Cash	Caffe Latte
Outlet 1	9/12/ 2020	19:41:27	6MRDOY	Payment	Cash	Vietnam Drip
Outlet 1	9/12/ 2020	19:41:27	6MRDOY	Payment	Cash	Red Choco Velvet
Outlet 1	9/12/ 2020	18:34:56	6MRDOW	Payment	Cash	Cappucino
Outlet 1	9/12/ 2020	18:34:56	6MRDOW	Payment	Cash	Caffe Latte
Outlet 1	9/12/ 2020	18:33:53	6MRDOV	Payment	Cash	Creme Brulee
Outlet 1	9/12/ 2020	18:33:53	6MRDOV	Payment	Cash	Matcha Latte

- *One Hot* Encoding. At this preprocessing stage, one hot encoding will only require the receipt number and items feature. Table 4 shows the results.

TABLE IV. EXAMPLE OF ONE HOT ENCODING RESULTS

Affogato	Ayam Rica-rica	Babycino	Bandrek	Caffe Latte	Caffe Mocha	Cappucino	Cascara Tea
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

D. Implementation and Evaluation

The preprocessed data then becomes input that is ready to look for patterns that often appear using FP-Growth. The hope is that from this pattern that often appears, items that are often purchased together can be found to provide cross-selling recommendations. In addition, up selling can be recommended based on frequently purchased beverage items. At this stage the FP-Growth algorithm as described in Section II is implemented.

After the model is successfully created, lift ratio and validation of association rules were used as evaluation techniques.

- Lift ratio. Lift ratio is an important parameter that must be considered in association rules in addition to support and confidence. This parameter is used to calculate the strength of the random event rules that occur in each combination. Before calculating the lift ratio value, the expected confidence value must be calculated first.

Itemsets that have a value greater than or equal to the minimum support will be categorized in the high-frequency itemset row. The itemset will later be recalculated with the confidence value it has with the confidence and minimum confidence formula. Then the value will be calculated to prove the strength of the association rules formed by calculating the lift ratio value.

- Association Rule Validation. The dataset has been divided into two, namely training data and testing data. In the training data, the process of determining the association rules formed from transaction data has been carried out. After the rules are formed, it will be tested on data testing how many frequencies can be obtained from each rule to find out which rules are effective. Association rules are divided into 9 months for training data and 3 months for testing data.

IV. Results and Discussions

A. Analysis of Minimum Support

The minimum support analysis will test the effect of the minimum support results that have been carried out using the FP-Growth algorithm on the rules that occur in transaction data at Cascara Coffee. Testing is done by trying different minimum support parameters with a fixed minimum confidence parameter. In this study, trials have been carried out using the minimum support parameter from $0{,}05\ x\ 10^{-2}$ to $0{,}5\ x\ 10^{-2}$. However, in this study, only the sample is used. The minimum confidence parameter used in this study is 0.3. while the minimum support parameters used are $0{,}08\ x\ 10^{-2}$, $0{,}1\ x\ 10^{-2}$, and $0{,}2\ x\ 10^{-2}$. Fig. 3 shows the experimental results of three types of minimum support parameters used in the study.

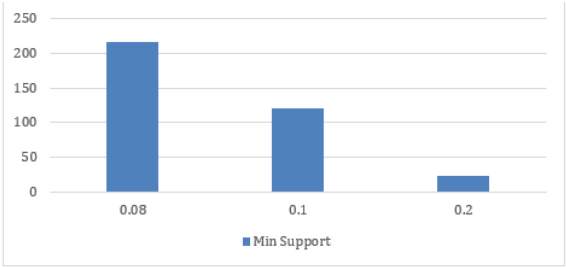


Fig. 3. Minimum support analysis results.

At the minimum support parameter $08\ x\ 10^{-2}$ formed as many as 217 rules that occur in transaction data. The lift ratio formed shows a positive correlation because the lift ratio test results for each rule have a value greater than 1. Next is the result of the association rule of the parameter $0{,}1\ x\ 10^{-2}$. In these results, 121 combinations of rules were formed. The results of the lift ratio show a positive correlation. However, the average of the frequency results showed an increase. Then the last experiment is the result of the minimum support parameter $0{,}2\ x\ 10^{-2}$. The results of these parameters produce 24 rules. As in the previous experiment, the lift ratio value shows a positive correlation. The frequency column shows the results of the emergence of rules in the last three months which are quite large on average compared to the previous experiment.

B. Minimum Confidence Analysis

The minimum confidence analysis will test how much influence the minimum confidence parameter in the FP-Growth algorithm has on the results formed from transaction data that has occurred at Cascara Coffee. In this analysis, the minimum confidence parameter will be tested against the minimum support parameter. In this study, trials have been carried out using the minimum confidence parameter from 1.0 to 0.3. However, in this study, only the sample is used. The minimum confidence parameters to be tested are 0.3, 0.4, and 0.5. While the minimum support parameter to be used is $08\ x\ 10^{-2}$. Figure 4 shows the results of the analysis of the minimum support.

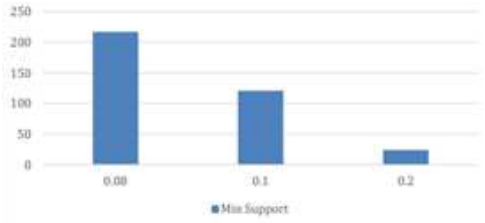


Fig 4. Minimum support analysis results.

At the minimum confidence parameter 0.3 and the minimum support $08\ x\ 10^{-2}$, 217 rules are formed. These rules have a lift ratio that is positively correlated because the lift ratio value is greater than 1. Next is the result of the rules formed at the confidence parameter of 0.4. There are 76 rules formed from these parameters. The lift ratio in this parameter also has a positive correlation. The last experiment conducted in Table V is the result of the minimum confidence parameter of 0.5. The results of these parameters form 22 rules that are formed. The value of the lift ratio in this parameter also has a positive correlation.

TABLE V. Example of Association Rule Results with Minimum Confidence 0.5

No	Support $x\ 10^{-2}$	Confidence	Lift Ratio	Item A	Item B	Freq
1.	0.099	0.500	2.001	Creme Brulee, Kopi Susu Original, Gorengan Platter	Caffe Latte	0
2.	0.115	0.538	2.154	Creme Brulee, Red Choco Velvet, Gorengan Platter	Caffe Latte	1
3.	0.132	0.533	2.913	French Fries, Gorengan Platter, Caffe Latte	Red Choco Velvet	6

C. Analysis of Lift Ratio Results Against Association Rule Results

From the results of the association that has been formed, then the lift ratio is calculated which proves the correlation formed from each rule. Table VI is an example of the appearance of the lift ratio. With these parameters, the average lift ratio value is 1.510, which means that the average of these rules has a positive correlation. Fig. 4 shows the comparison of the minimum, maximum, and average values on the results of the association rules formed with the minimum support parameters of $0{,}08\ x\ 10^{-2}$, $0{,}1\ x\ 10^{-2}$ and $0{,}2\ x\ 10^{-2}$. At the minimum support value $0{,}1\ x\ 10^{-2}$ and $0{,}2\ x\ 10^{-2}$ have the same minimum value of 1.204 while at

$0{,}08\ x\ 10^{-2}$ has a smaller minimum value of 1.201. At the maximum value of the parameter $0{,}2\ x\ 10^{-2}$ has a fairly low value of 2.703 compared to other parameters which have a value of 4.506. While for the average all parameters have different values, namely 2.039 for $0{,}08\ x\ 10^{-2}$, 1,935 for $0{,}1\ x\ 10^{-2}$ and 1,65 for $0{,}08\ x\ 10^{-2}$.

TABLE VI. EXAMPLE OF LIFT RATIO RESULTS

No.	Item A	Item B	*Support*	*Confidence*	*Lift*
1.	Kopi Susu Original, Gorengan Platter	Caffe Latte	0.404	0.386	1.547
2.	Gorengan Platter, Caffe Latte	Kopi Susu Original	0.404	0.320	1.664
3.	Creme Brulee, Gorengan Platter	Caffe Latte	0.330	0.364	1.455
4.	French Fries, Gorengan Platter	Red Choco Velvet	0.297	0.333	1.820
5.	Gorengan Platter, Pure	Caffe Latte	0.222	0.321	1.286

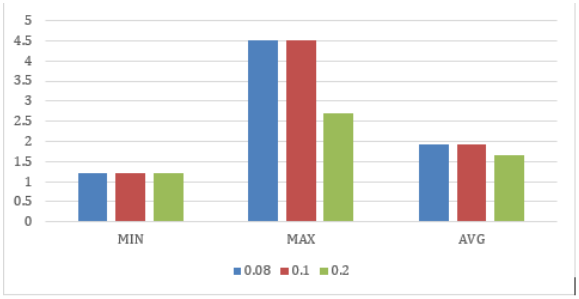


Fig. 4. Lift ratio analysis on minimum support parameter.

Fig. 5 shows the comparison of the minimum, maximum and average analysis results from the results of the association rules formed from the minimum confidence parameters of 0.3, 0.4 and 0.5. It can be seen in the figure that the minimum result from the image parameter 0.5 has the highest minimum value of 2.001 while the others have a value of 1.630 for parameter 0.4 and 1.200 for parameter 0.3. For the maximum value, all tested parameters have the same value, which is 4,506. Meanwhile, for the value of the average, it can be seen that it has an increase in the parameters of 0.3, 0.5 and 0.5 with the lift ratio values being 2.039, 2.327, 2.789, respectively.

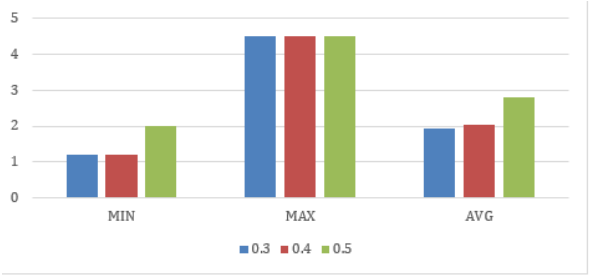


Fig 5. Lift ratio analysis on minimum confidence parameter.

From the results of the analysis, the big picture that can be drawn is that the lift ratio is strongly influenced by the parameters used. The higher the minimum confidence and if the minimum support parameter used is low, it can increase the average value of the lift ratio. But the most important thing about the lift ratio is the correlation value. Therefore the result obtained must be more than 1.

With the results that have a positive tendency, it can be concluded that the correlation can be used as a recommendation for cross-selling and up-selling marketing techniques because a positive correlation is a relationship between 2 variables where an increase in one variable causes an increase in the value of the other variable. Or vice versa, the smaller the value of a variable, the value of other variables will also go down. It can also be said, this correlation is a unidirectional relationship so that when you find a transaction to buy *item a*, there is also a tendency to buy *item b*.

D. Association Rule Validation

The rules that have been formed using the FP-Growth algorithm will be tested on data testing. The rule will calculate the occurrence or frequency of the testing data. Testing data is transaction data for the last three months. Table 8 is an example of the frequency results contained in the rules formed 8.

TABLE VII. FREQUENCY IN DATA TESTING

No.	Item A	Item B	Freq.
1.	Kopi Susu Original, Gorengan Platter	Caffe Latte	12
2.	Gorengan Platter, Caffe Latte	Kopi Susu Original	12
3.	Creme Brulee, Gorengan Platter	Caffe Latte	9
4.	French Fries, Gorengan Platter	Red Choco Velvet	12
5.	Gorengan Platter, Pure	Caffe Latte	3
6.	Matcha Latte, Gorengan Platter	Red Choco Velvet	5
7.	Cappucino, Gorengan Platter	Caffe Latte	9
8.	Cappucino	Caffe Latte	121
9.	Cappucino, Red Choco Velvet	Caffe Latte	19
10.	Creme Brulee, Cappucino	Caffe Latte	19
11.	Cappucino, French Fries	Caffe Latte	14
12.	Pisang Goreng, Pure	French Fries	5
13.	Matcha Latte, Pisang Goreng	French Fries	7
14.	Roti Bakar Perancis	Caffe Latte	16
15.	Cappucino, Pure	Red Choco Velvet	13
16.	Cappucino, Pure	Caffe Latte	8
17.	Cappucino, Mi Kuah	Caffe Latte	9
18	French Fries, Taro Latte	Matcha Latte	12
19.	Pure, Taro Latte	Red Choco Velvet	12
20.	Choco, Caffe Latte	Red Choco Velvet	7
21.	Choco, Pure	Red Choco Velvet	9
22.	Caffe Mocha	Red Choco Velvet	35
23.	French Fries, Makaroni Keju	Caffe Latte	7
24.	Creme Brulee, Makaroni Keju	Caffe Latte	6

In Table VII, the results of the minimum support parameters are $0{,}20\ x\ 10^{-2}$ and the minimum confidence is 0.3. It can be seen that the number of occurrences of each association rule is different. This is because consumer patterns vary. By validating the results of these rules, it can be seen that the rules actually work according to the real data.

From the parameters that have been tested, the minimum support $0{,}2\ x\ 10^{-2}$ and minimum confidence 0.3 have the best frequency results compared to other parameters that have been tested. For more details, see the Table VIII.

TABLE VIII. FREQUENCY COMPARISON

No.	Support $x\ 10^{-2}$	Confidence	Min	Max	Average
1.	0.08	0.5	0	6	2,000
2.	0.08	0.4	0	9	2.273
3.	0.08	0.3	0	121	4.465
4.	0.1	0.3	0	121	5.950
5.	0.2	0.3	3	121	15.875

E. Determining Cross Selling and Up Selling

After the results of the association rules using the algorithm that have been made are formed, in the end the right business decisions can increase sales. One way that can be used as an appropriate technique to implement this research into marketing is cross-selling and up-selling.

In this study, the system only recommends the association rules that are formed. However, the recommendations given can give an idea of the success or failure of the marketing techniques that will be used. In this study, the best parameter to see what rule options are likely to be successful in cross-selling and up-selling marketing techniques is the result of the minimum support parameter of $0{,}20\ x\ 10^{-2}$ and the minimum confidence of 0.3.

V. CONCLUSION

The results of the association rules can be seen from the lift ratio results. In this study, there is a tendency for a positive correlation so that it can be concluded that the correlation can be used as a recommendation for cross-selling and up-selling marketing techniques. In the experiment, the highest average lift ratio was 2,789 from the minimum value of 2,001 and the maximum value of 4,506.

From the results of the analysis in this study, it can be concluded that the association rules using a minimum support parameter of $0{,}20\ x\ 10^{-2}$ and a minimum confidence of 0.3 which displays 24 rules are the best parameters in this study because they produce fairly high frequency validation results. However, the results of the validation are different due to the various patterns of customers, but by validating the results of the rules, it can be seen that the rules actually work in accordance with the real data. So that these parameters can be used as material for cross-selling and up-selling marketing techniques.

ACKNOWLEDGMENT

The author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service Contract, Universitas Padjadjaran No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] B. Santosa and A. Umam, Data Mining dan Big Data Analytics, Penebar Media Pustaka, Penebar Media Pustaka, 2018.

[2] R. Anggrainingsih, N. R. Khoirudin and H. Setiadi, "Discovering drugs combination pattern using fp-growth algorithm", International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 4(September), pp. 735–738, 2017, doi: 10.11591/eecsi.4.1085.

[3] R. Cui and H. Shinb, "Sharing aggregate inventory information with customers: Strategic cross-selling and shortage reduction", Management Science, 64(1), pp. 381–400, 2018, doi: 10.1287/mnsc.2016.2600.

[4] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition, Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition, 2011, doi: 10.1002/9781118029145.

[5] K. Potdar, T.S. Pardawla, and C. D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers", International Journal of Computer Applications, 175(4), pp. 7–9, 2017, doi: 10.5120/ijca2017915495.

[6] M. A. Rashid, M. T. Hoque, and A. Sattar, "Association Rules Mining Based Clinical Observations," arXiv preprint arXiv:1401.2571, 2014.

[7] L. Wang et al. "Mining data association based on a revised FP-growth algorithm", Proceedings - International Conference on Machine Learning and Cybernetics, 1(1), pp. 91–95, 2012, doi: 10.1109/ICMLC.2012.6358892.

[8] J. Han, J. Pei, and Kamber, Data Mining: Concepts and Techniques, Third Edition, p. 847, Elsevier, 2011.

[9] J. Kwiatkowska, "Cross - selling and up - selling in a bank", 7(4), 2018

[10] C. Schmitz, Y. C. Lee, and G. L. Lilien, "Cross-selling performance in complex selling contexts: An examination of supervisory- and compensation-based controls", Journal of Marketing, 78(3), pp. 1–19, 2014.

[11] Kakamura, "Approaches to Customer Segmentation", Journal of Relationship Marketing, 2667(October), pp. 117–130, 2008, doi: 10.1300/J366v06n03.