JITE, 7 (1) July 2023 ISSN 2549-6247 (Print) ISSN 2549-6255 (Online)

JITE (Journal of Informatics and Telecommunication Engineering)

Available online http://ojs.uma.ac.id/index.php/jite **DOI : 10.31289/jite.v7i1.10114**

Received: 20 July 2023	Accepted: 28 July 2023	Published: 28 July 2023

Machine Learning Model for Language Classification: Bag-of-words and Multilayer Perceptron

Devi Hawana Lubis 1), Sawaluddin 2) & Ade Candra 3)

1,3) Master of Informatics Program, Universitas Sumatera Utara
2) Department of Mathematics, Universitas Sumatera Utara

*Coresponding Email: devihawana@gmail.com

Abstrak

Ketersediaan data saat ini telah menjadi aset besar bagi penelitian yang digunakan untuk berbagai keperluan seperti untuk pembelajaran mesin. Salah satu metode pembelajaran mesin dasar untuk pemrosesan bahasa alami adalah bag-of-words. Masalah dalam penelitian ini adalah sulitnya mengklasifikasikan teks karena teks masih memiliki karakteristik yang tidak terstruktur, sehingga penelitian ini akan menerapkan model untuk mengklasifikasikan bahasa teks. Teks akan ditempatkan dalam empat kategori, Inggris, Indonesia, Jerman dan Perancis. Penelitian dilakukan dengan menggunakan Bag-of-words dan Multilayer Perceptron untuk mengatasi masalah pembelajaran mesin terawasi ini. Penggunaan Bag-of-words untuk melakukan representasi teks untuk pola yang sederhana, pemrosesan yang mudah, dan kinerja yang baik. Di sisi lain, perceptron multilayer memiliki kemampuan untuk mempelajari pola data yang kompleks dalam bentuk gambar, teks, atau video. Penelitian ini akan mengumpulkan data dengan menggunakan teknik text mining yaitu crawling media sosial Twitter sebanyak 4000 record data. Penelitian ini menghasilkan model dengan akurasi 98 persen dengan loss 0,14 persen yang menunjukkan kinerja model yang baik dalam mengklasifikasikan bahasa berdasarkan data teks.

Kata kunci: klasifikasi teks, multilayer perceptron, bag-of-words, model pembelajaran mesin

Abstract

The availability of data today has become a great asset for research that is used for various purposes such as for machine learning. One of the basic machine learning methods for natural language processing is bag-of-words. The problem in this study is the difficulty in classifying texts because texts still have unstructured characteristics, so this study will apply a model to classify the language of texts. Texts will be placed in four categories, English, Indonesian, German and French. Research was conducted using Bag-of-words and Multilayer Perceptron to solve this supervised machine learning problem. The use of Bag-of-words to perform text representation for simple patterns, easy processing and good performance. On the other hand, a multilayer perceptron has the ability to study complex data patterns in the form of images, text or videos. This study will collect data using text mining techniques, namely crawling Twitter social media as many as 4000 data records. This study produces a model with an accuracy of 98 percent with a loss of 0.14 percent which shows good model performance in classifying languages based on text data.

Keywords: text classification, multilayer perceptron, bag-of-words, machine learning model

How to Cite: Lubis, D. H., Sawaluddin, S., & Candra, A. (2023). Machine Learning Model for Language Classification: Bag-of-words and Multilayer Perceptron. JITE *(Journal of Informatics and Telecommunication Engineering),* 7(1), 356-365.

I. INTRODUCTION

Moment interference is happening in the brain when multilingual communication takes place (Grundy et al., 2017);(Liu et al., 2019);(García et al., 2017). Switching language from one to another is resulting in a dynamic interaction that affects both the speaker and listener. Furthermore, studies said that there is an occurrence of general adaptation when switching language on cross-language communication(Wang et al., 2022)(Lubis et al., 2022b). Communication nowadays happens through offline and online media, let it by face-to-face, text-based or even video call. Twitter is one of the text-based social media that publishes more than 6000 tweets every second (Shrivastava & Kumar, 2021);(Al-Khowarizmi et al., 2017). Many researchers have an eye on natural language processing (NLP) to find out more about how the machine understands texts, analyze sentiment, classify, filter, summarize text and many more(Ishihara, 2021);(Yan et al., 2020);(Jeremy & Suhartono, 2021). Understanding text can be done with various methods. For example, machine learning that has multiple approaches to solve complicated problems. Here we will present two methods, Bag-of-words (BoW) and Multilayer Perceptron (MLP) to create a model that can do text classification to predict language of twitter data(Cai, 2021);(Lubis et al., 2022a). The use of these methods comes from the ability of the BoW to put text into fixed-size matrices on text with different lengths and MLP that can optimize parameters of text to find the best result for the prediction. The classification will be based on twitter data and will be divided into 4 languages, English, Bahasa Indonesia, German and France. Tensorflow is used to do the alghoritm of chosen machine learning methods. All learning process is prepared and ran in google colab. To learn more about the model, the python library Shapely Additive Explanation (SHAP) is used. This paper is structured as follows; the methodology section presents the elaboration of machine learning, BoW and multilayer perceptron. Proposed model section is about how the data collected and how the model created. The results and discussion section will show the model and its analysis. The last section is about the conclusion of this research.

II. LITERATURE REVIEW

A. *Related Research*

Based on previous research that has been done using the Bag-of-words technique and multilayer perceptron in language classification based on text data, there are studies as follows:

1. Previous research conducted by (Balamurali & Ananthanarayanan, 2020) applying Bag-of-words to extract text data will then collect data and apply it to the field of sentiment analysis using a multilayer perceptron which generates a model by generating word diagrams with word assessment modeling which is useful for the field of sentiment analysis
2. Previous research conducted by (Hájek, 2018) doing a combination of Bag-of-words and manual annotations in the field of sentiment to make stock predictions, this study produces a model that can make predictions with a combination of Bow can improve accuracy
3. Previous research conducted by (Diera et al., 2022) conducted a comparison of word extraction techniques such as bag-of-words, sequences and graphs in classifying single and multi-data words. The results of this study produce accuracy in each word representation
4. Previous research conducted by (Galke & Scherp, 2021) classify questions that apply a feature extraction model, namely bag-of-words, sequences and graphs. This study produces a kmodel that can classify questions related to data authenticity
5. Previous research conducted by (Araujo et al., 2020) carry out the implementation of Bag-of-words which will be combined with text preprocessing on unstructured data then will apply models from machine learning to carry out classifications, this research produces 93% accuracy in using Bag-of-words which will be combined with text preprocessing

B. *Bag-Of-Words (BoW)*

Bag-of-words is a fixed-sized input vector that transforms text with different length(Ofer et al., 2021). It can normalize vectors to show frequencies of every token (word) known as most-used-words. This method is a good predictor for supervised learning on classification problem. However, it may ignore the structure of the sentences that make this method not suitable for semantic problems. BoW works in 4 steps, which are accepting input documents, cleaning the text, tokenizing texts into matrices and creating vector that contain most-used-words and indicator of weather the word is happening on the text or not. Table I showing the illustration of how the BoW created matrices of text by the occurrence of words in the sentences. Words “de, die, ich, aku, ini, the” are the most-used-words. Number 1 indicates that words occur in the text and 0 says the opposite. This is how the fixed-sized matrices are created

Tabel 1 Example of Most-used-words BoW

tweet	Most-used-words					
	de	die	ich	aku	ini	the
kopf hoch ich hoffe es ist nicht all zu schlimm da ist die wl und fifa natrlich unwichtig hingegen	0	1	1	0	0	0
Heres a song for you the naughty ride (feat major lazer) via spotify	0	0	0	0	0	1
Pourquoi elle la appuy sur la pdale de frein cette conasse	1	0	0	0	0	0
Sini kak cek di turnitin haloo aku open jasa cek plagiasi	0	0	0	1	0	0

C. *Multilayer Perceptron (MLP)*

Multilayer perceptron is a machine learning algorithm that contains an input layer, one or more hidden layer and an output layer. There are nodes in every layer of the network. Every node in each layer is connected to the next one as shown on Fig 1. Each node has their own weight and bias(Turkoglu & Kaya, 2020);(Feng et al., 2020).

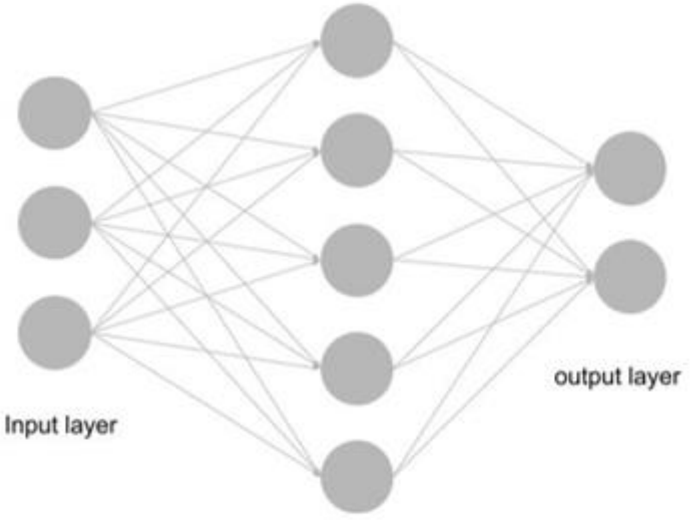


Figure 1 MLP network

MLP uses an activation function to put the output on to the label that is declared in advance for the supervised learning problems. The multilayer perceptron formula shown below has weight and bias as the parameter. Research has shown that this method has good performance on classification and pattern learning(Heidari & Shamsi, 2019);(Sharma et al., 2022);(Opěla et al., 2021).

$$g\left(\sum_{i=1}^{R} w_{ij}\ x_i + b_j \right) \tag{1}$$

Where,

g : activation function
i : input layer
R : total data
j : hidden layer
w_{ij} : weight on layer i, hidden layer j
x_i : neuron layer i
b_j : bias layer j

III. PROPOSED MODEL

Bag-of-words (BoW) is creating the vector to be used as the input layer of the Multilayer Perceptron. MLP is designed with one hidden layer and the sigmoid as the activation function. The data are taken from twitter API that contains 4000 tweets in four different languages

A. Collecting data

Data are collected by using python with the twitter developer account. As we register a new project in the portal, key and token will be given. It is used for authentication to the application programming interface. After that, tweepy library is used to fetches and cleans the data. Cleaning the data mean removing the mention and hashtag of the tweet data. Then the data will be saved into csv format file. Fig 2 shows process of gathering data and Fig 3 is sample data from twitter API

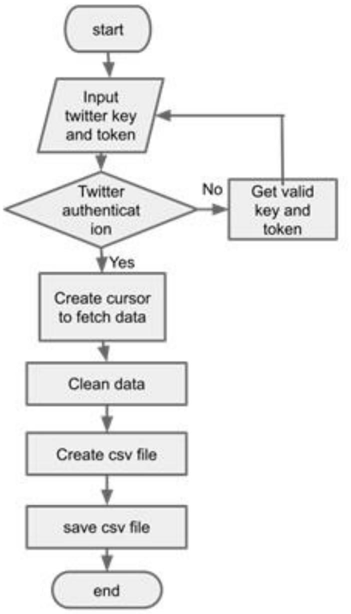
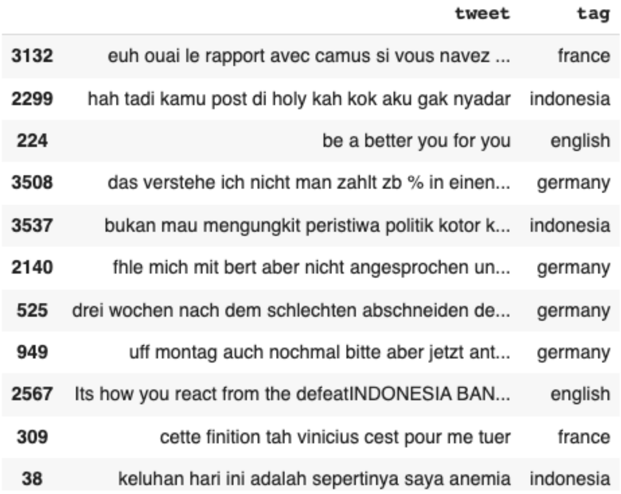


Figure 2 Collecting data flowchart

Caption in Figure 2 explains that data collection will use techniques from text mining which will use crawling data to obtain data from Twitter social media, technically will input tokens on Twitter social media to obtain data. After the data is obtained, data cleaning will be carried out by applying preprocessing text which consists of case folding, stop words, tokenizing then the file will be saved in csv format



	tweet	tag
3132	euh ouai le rapport avec camus si vous navez ...	france
2299	hah tadi kamu post di holy kah kok aku gak nyadar	indonesia
224	be a better you for you	english
3508	das verstehe ich nicht man zahlt zb % in einen...	germany
3537	bukan mau mengungkit peristiwa politik kotor k...	indonesia
2140	fhle mich mit bert aber nicht angesprochen un...	germany
525	drei wochen nach dem schlechten abschneiden de...	germany
949	uff montag auch nochmal bitte aber jetzt ant...	germany
2567	Its how you react from the defeatINDONESIA BAN...	english
309	cette finition tah vinicius cest pour me tuer	france
38	keluhan hari ini adalah sepertinya saya anemia	indonesia

Figure 3 Data

B. Creating Model

Figure 2 shows the process of building the model. It is programmed in Google Cloud Service (GCS) with python and tensorflow. Tensorflow provides machine learning algorithms that can be used to build models. Data from twitter will be split into training and testing data with a ratio of 80:20 percent. After that, data will be processed in custom BoW class to transform the text into matrices. Then a model will be created with tensorflow that contains an input layer, a hidden layer and an output layer. The summary of the model is shown at Fig 5. Then data are ready to train and the result is shown at Fig 4

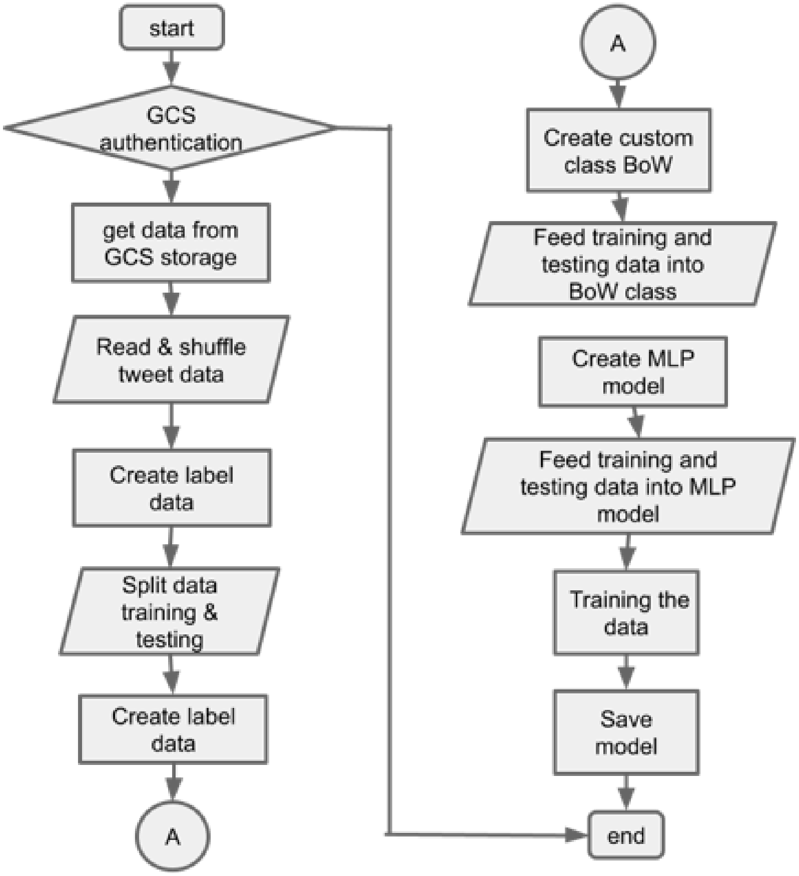
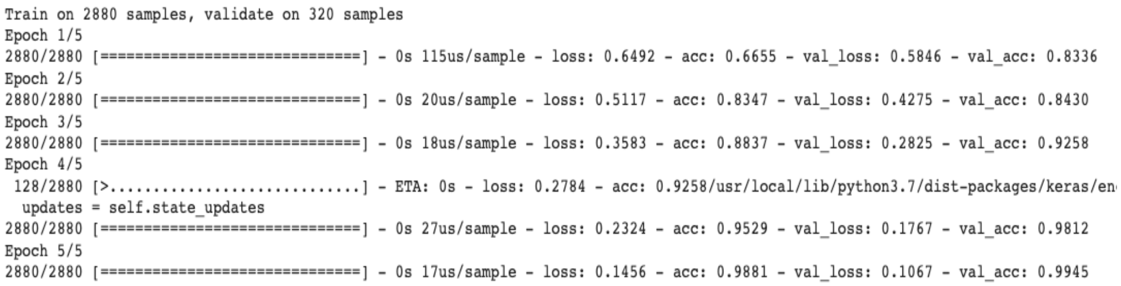

start
GCS authentication
get data from GCS storage
Read & shuffle tweet data
Create label data
Split data training & testing
Create label data
A
A
Create custom class BoW
Feed training and testing data into BoW class
Create MLP model
Feed training and testing data into MLP model
Training the data
Save model
end

Figure 3 Model Proposed

Caption in figure 3 is a model that will be submitted to be able to carry out language classification, in figure 3 there are steps for inputting data to read tweet data then labeling the data after that the process of dividing training and testing data is carried out then text data will be processed with the Bow model after feature extraction is obtained by applying a multilayer perceptron to classify languages based on text data. The following is the training process for the multilayer perceptron model which is shown in Figure 4 below



```
Train on 2880 samples, validate on 320 samples
Epoch 1/5
2880/2880 [==============================] - 0s 115us/sample - loss: 0.6492 - acc: 0.6655 - val_loss: 0.5846 - val_acc: 0.8336
Epoch 2/5
2880/2880 [==============================] - 0s 20us/sample - loss: 0.5117 - acc: 0.8347 - val_loss: 0.4275 - val_acc: 0.8430
Epoch 3/5
2880/2880 [==============================] - 0s 18us/sample - loss: 0.3583 - acc: 0.8837 - val_loss: 0.2825 - val_acc: 0.9258
Epoch 4/5
 128/2880 [>.............................] - ETA: 0s - loss: 0.2784 - acc: 0.9258/usr/local/lib/python3.7/dist-packages/keras/en
  updates = self.state_updates
2880/2880 [==============================] - 0s 27us/sample - loss: 0.2324 - acc: 0.9529 - val_loss: 0.1767 - val_acc: 0.9812
Epoch 5/5
2880/2880 [==============================] - 0s 17us/sample - loss: 0.1456 - acc: 0.9881 - val_loss: 0.1067 - val_acc: 0.9945
```

Figure 4 process training

IV. RESULT AND DISCUSSION

A. Model Analysis

Fig 5 shows the impact of words on the model with average magnitude from 0 to 5%. The plot says that the most common word in the tweet data is “de”. It is used to make prediction the most. The model has picked 200 most-used-words to make decision for the prediction, Fig 7 is only showing 20 of them. The impact value is saved inside the nested array of 4 categoies that represent each language. Each language has 200 array that represents each most-used-words and its prediction value. The model has a complex calculation in the training process

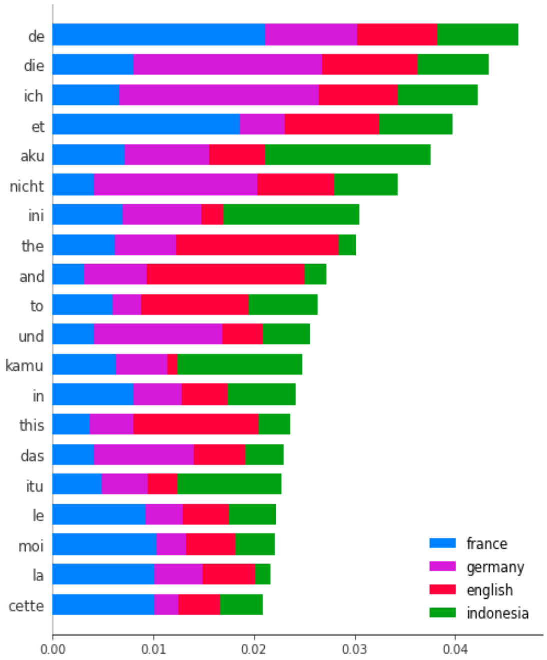


Figure 5 Words impact on classification

Figure 5 is showing the average impact of the words on a model where 20 words are used as the sample of most-used-words with the average magnitude from 0 to 5%. For example, the word “de” is used as the main word for model classification with an average value of more than 4% which has the polarity to France. As for the word “ich” which has an average impact of 4% as the German, word “aku” with average impact of 3.8% as Bahasa Indonesia and word “the” with average impact of 3.1% as an English

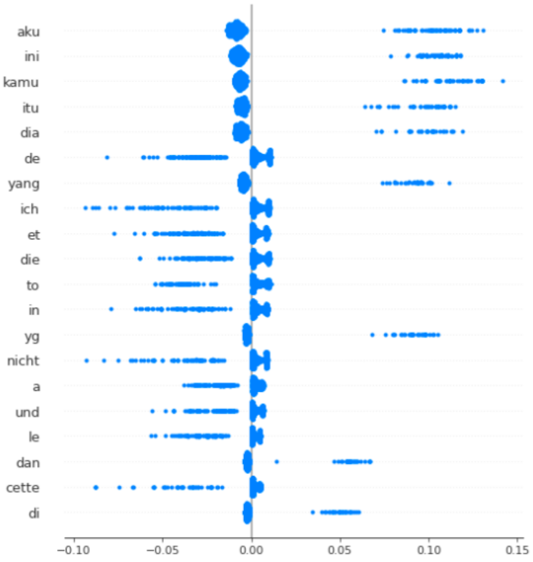


Figure 6 Word Impact on Bahasa Indonesia

Figure 6 is a plot that shows the word impact classification in Bahasa Indonesia. The word “aku, ini, itu, dia, kamu, dia, yang” has the positive impact. This helps the predictor to put tweets that contain those words to classify as Bahasa Indonesia. As for the negative value for the word “de, ich et, die, to, in, "nicht“ means that they are not Bahasa Indonesia

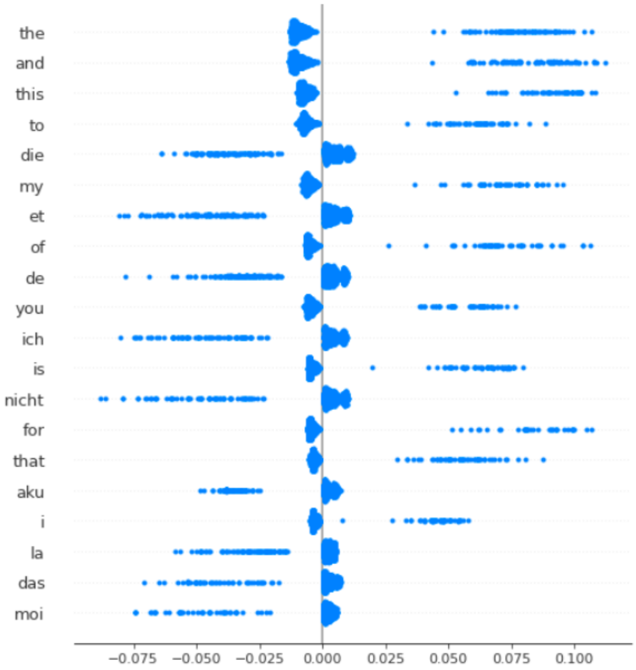


Figure 7 Word Impact on English

Plot in figure 7 is the opposite of figure 6. It shows the average impact of words in English. The word the” has a positive value with the average magnitude between 0 to 0.012 of all data value that means it can be classified as English

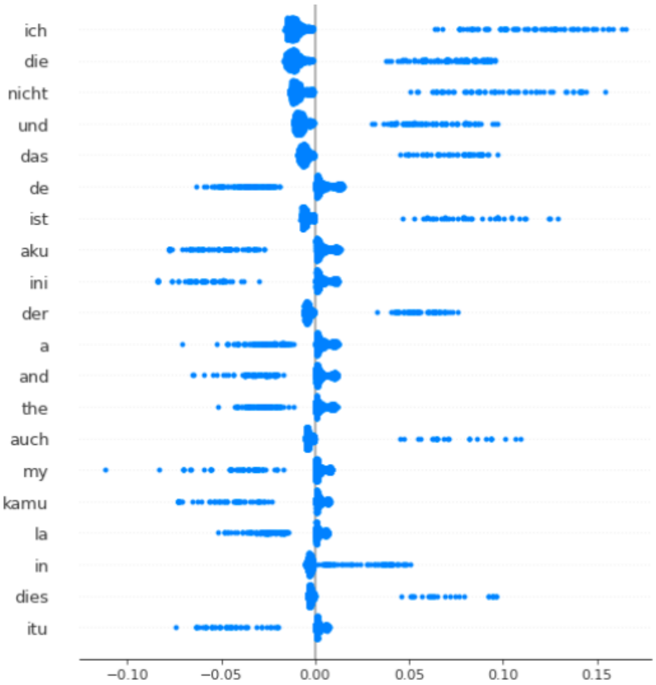


Figure 8 Word Impact on German

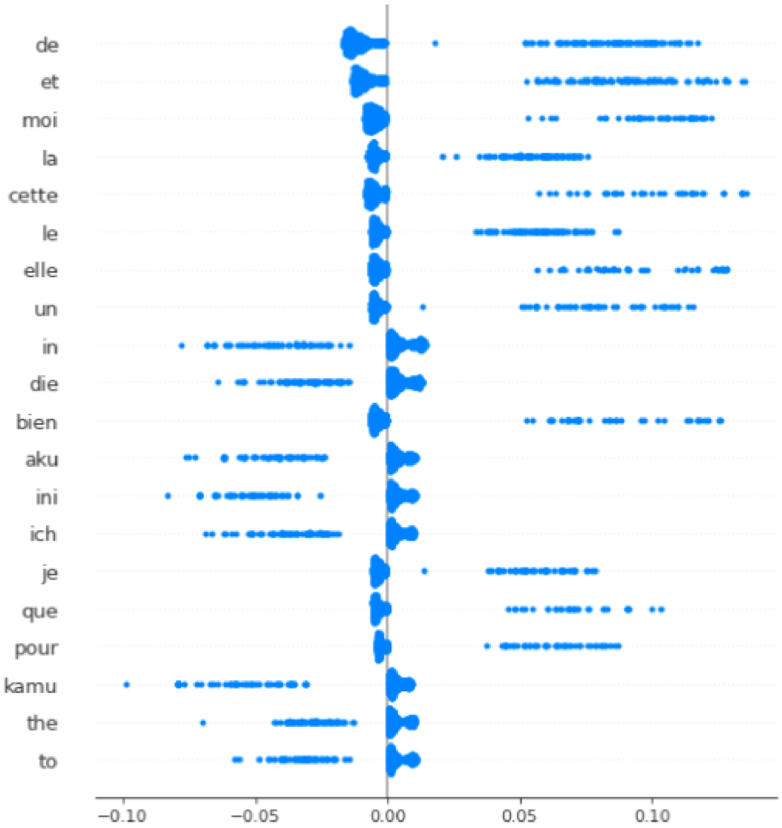


Figure 9 Word Impact on France

Plot at figure 8 and 9 show most used words to help classify the tweet. This plot put words "ich, die, nicht, und, das, ist, der, auth, in, dies" to German and words "de, et, moi, la, cette, le, elle, un, bien, je, que, pour" to France. Figure 10 and 11 shows the accuracy and loss value of the model. Loss plot shows decreasing number of values on each epoch and on the other hand, the value of accuracy plot keeps increasing on each epoch. It is indicating the good performance of the model.

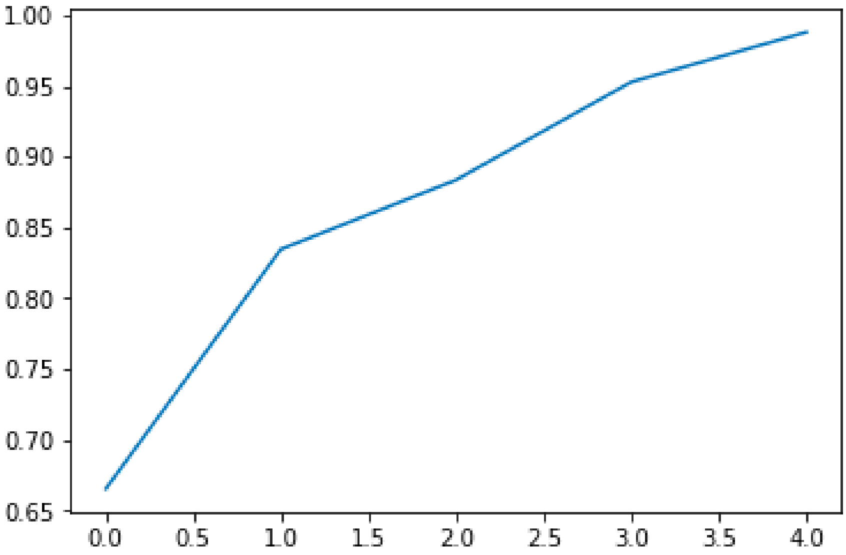


Figure 10 Model Accuracy

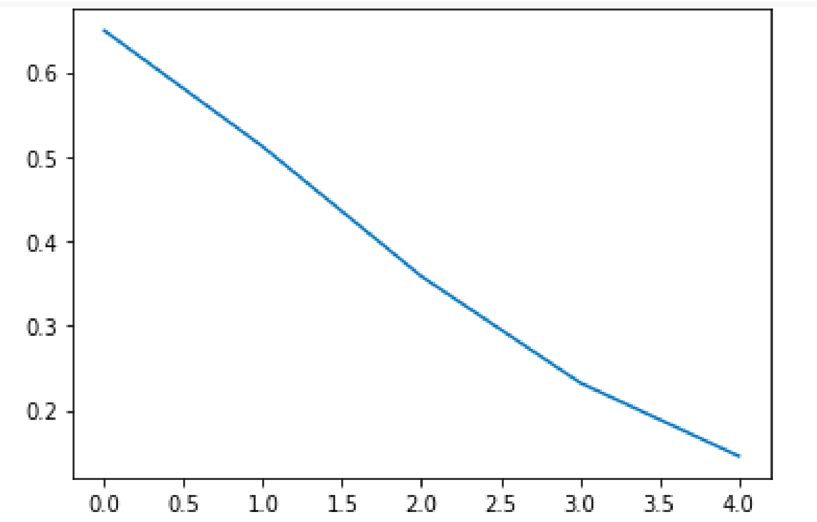


Figure 11 model Loss Accuracy

Fig 10 and Fig 11 shows the accuracy and loss value of the model. Loss plot shows decreasing number of values on each epoch and on the other hand, the value of accuracy plot keeps increasing on each epoch. It is indicating the good performance of the model. The model created has reached 99% accuracy and loss value of 0.14 which means that the model has a very good performance on text classification. This research can be used to make changes from unstructured text data using a model that the author built using the concept of machine learning

V. DISCUSSION

This study will use data originating from social media Twitter, data collection will be carried out using text mining stages with crawling techniques on Twitter social media which will retrieve 4000 data, then this research will use one of the basic natural language machine learning methods. processing, namely pockets of words. This study conducts language classification based on text data which will be placed into four categories, English, Indonesian, German and French. Using Bag-of-words to represent text into simple patterns, easy processing, and good performance. On the other hand, a multilayer perceptron has the ability to study complex data patterns in the form of images, text or videos. This study uses parameters such as epoch values, activation values, loss values, and the number of layers which produce a model with an accuracy of 98 percent with a loss of 0.14 percent which shows good model performance in classifying languages based on text data.

VI. CONCLUSION

The model created has reached 99% accuracy and loss value of 0.4% which means that the model has a very good performance on text classification. The research can be used as reference or further study and comparison on machine learning algorithms. Change of data, slang, cleaning data techniques and different methods of machine learning can be applied as the next topic for future research.

REFERENCES

Al-Khowarizmi, A., Sitompul, O. S., Suherman, S., & Nababan, E. B. (2017). Measuring the Accuracy of Simple Evolving Connectionist System with Varying Distance Formulas. Journal of Physics: Conference Series, 930(1), 0–6. https://doi.org/10.1088/1742-6596/930/1/012004

Asadulaev, A., Kuznetsov, I., Stein, G., & Filchenkov, A. (2020). Exploring and exploiting conditioning of reinforcement learning agents. IEEE Access, 8, 211951–211960.

Cai, M. (2021). Natural language processing for urban research: A systematic review. Heliyon, 7(3).

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. Information Processing & Management, 59(2), 102798.

Chen, H., Zhang, Y., Cao, Y., & Xie, J. (2021). Security issues and defensive approaches in deep learning frameworks. Tsinghua Science and Technology, 26(6), 894–905.

Feng, X., Ma, G., Su, S.-F., Huang, C., Boswell, M. K., & Xue, P. (2020). A multi-layer perceptron approach for accelerated wave forecasting in Lake Michigan. Ocean Engineering, 211, 107526.

García, M. A. M., Rodríguez, R. P., & Rifón, L. A. (2017). Wikipedia-based cross-language text classification. Information Sciences, 406, 12–28.

Grundy, J. G., Anderson, J. A. E., & Bialystok, E. (2017). Bilinguals have more complex EEG brain signals in occipital regions than monolinguals. NeuroImage, 159, 280–288.

Heidari, M., & Shamsi, H. (2019). Analog programmable neuron and case study on VLSI implementation of Multi-Layer Perceptron (MLP). Microelectronics Journal, 84, 36–47.

Hurwitz, J., & Kirsch, D. (2018). Machine Learning IBM Limited Edition. Retrieved March, 22, 2021.

Ishihara, S. (2021). Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. Forensic Science International, 327, 110980.

Jeremy, N. H., & Suhartono, D. (2021). Automatic personality prediction from Indonesian user on twitter using word embedding and neural networks. Procedia Computer Science, 179, 416–422.

Liu, H., Zhang, M., Pérez, A., Xie, N., Li, B., & Liu, Q. (2019). Role of language control during interbrain phase synchronization of cross-language communication. Neuropsychologia, 131, 316–324.

Lubis, A. R., Prayudani, S., Lubis, M., & Nugroho, O. (2022a). Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) Models on News Data. 2022 5th International Conference of Computer and Informatics Engineering (IC2IE), 314–319.

Lubis, A. R., Prayudani, S., Lubis, M., & Nugroho, O. (2022b). Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method. 2022 1st International Conference on Information System & Information Technology (ICISIT), 106–111.

Maass, W., & Storey, V. C. (2021). Pairing conceptual modeling with machine learning. Data & Knowledge Engineering, 134, 101909.

MR, G. R., Somu, N., & Mathur, A. P. (2020). A multilayer perceptron model for anomaly detection in water treatment plants. International Journal of Critical Infrastructure Protection, 31, 100393.

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal, 19, 1750–1758.

Opěla, P., Schindler, I., Kawulok, P., Kawulok, R., Rusz, S., & Navratil, H. (2021). On various multi-layer perceptron and radial basis function based artificial neural networks in the process of a hot flow curve description. Journal of Materials Research and Technology, 14, 1837–1847.

Sharma, R., Kim, M., & Gupta, A. (2022). Motor imagery classification in brain-machine interface with machine learning algorithms: Classical approach to multi-layer perceptron model. Biomedical Signal Processing and Control, 71, 103101.

Shrivastava, M., & Kumar, S. (2021). A pragmatic and intelligent model for sarcasm detection in social media text. Technology in Society, 64, 101489.

Turkoglu, B., & Kaya, E. (2020). Training multi-layer perceptron with artificial algae algorithm. Engineering Science and Technology, an International Journal, 23(6), 1342–1350.

Wang, Y.-C., Chuang, C.-M., Wu, C.-K., Pan, C.-L., & Tsai, R. T.-H. (2022). Cross-language article linking with deep neural network based paragraph encoding. Computer Speech & Language, 72, 101279.

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-based bag-of-words model for text classification. IEEE Access, 8, 82641–82652.