Sentiment Analysis of Text Memes: A Comparison Among Supervised Machine Learning Methods

Endah Asmawati
Department of Informatics
Institut Teknologi Sepuluh Nopember
7025211030@mhs.its.ac.id
Department of Informatics Engineering
Universitas Surabaya
Surabaya, Indonesia
endah@staff.ubaya.ac.id

Ahmad Saikhu*
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
saikhu@if.its.ac.id

Daniel Siahaan
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
daniel@if.its.ac.id

***Abstract*—Meme is a new form of content in social media. A meme contains sentiment towards a particular issue, product, person, or entity. Memes can be in the form of text, images, or images that contain text. Memes are entertaining, critical, sarcastic, and may even be political. Traditional sentiment analysis methods deal with text. This study compares the performance of four sentiment analysis methods when used on Indonesian meme in the form of text and images that contain text. Firstly, the extraction of text memes was carried out, followed by the classification of the extracted text memes using supervised machine learning methods, namely Naïve Bayes, Support Vector Machines, Decision Tree, and Convolutional Neural Networks. Based on the experimental results, sentiment analysis on meme text using the Naïve Bayes method produced the best results, with an accuracy of 65.4%.**

Keywords—sentiment analysis, memes, supervised machine learning.

I. INTRODUCTION

Sentiment analysis is the process of analyzing people's opinions, sentiments, evaluations, judgments, attitudes, and emotions towards an entity such as a product, service, organization, individual, issue, event, and a specific topic and determining whether the polarity of the sentiments are positive, neutral, or negative [1]. We can learn the public mood through sentiment analysis so that phenomena can be predicted [2].

Many studies related to sentiment analysis have been carried out, including in the fields of politics [3], tourism [4,5], government intelligence [6], and other areas. Sentiment analysis in previous researches has been applied at different levels, namely document level, sentence level, or aspect level using data from various social media. Memes in the form of text data can be analyzed for sentiment analysis.

Research that uses memes as a data source to predict or analyze sentiment is still limited to date [7], so research opportunities in this field are still open. A meme is an idea, behaviour, or style that spreads quickly. Memes can be in the form of images from television shows, movies, or homemade images with the addition of words for humor [8]. Usually, memes are created when something becomes a trending topic. The rise in social media users has led to the rapid emergence and distribution of memes. Sentiment analysis is one of the analytical methods that can be used to quickly learn public opinion through the abundance of memes on social media [9].

Research on text sentiment analysis with datasets from Twitter or social media has previously been carried out. The methods used for sentiment analysis in these researches include machine learning methods (CNN [10], LSTM [11], SVM [12], Naïve Bayes [13]), Lexicon-based or hybrid methods [14]. These methods' average accuracy is relatively high, above 80%.

Research in meme sentiment analysis has not been widely carried out. In these studies [15,16,17], sentiment analysis was carried out on memes based on text and images. In [15], textual features were extracted from text data using lexicon-based approaches (SentiWordNet and Contextual Dictionary). Text analysis was done on four analysis levels: word, phrase, comment, and discussion. While emotions from facial expressions in images were extracted using CNN. The proposed method in this study resulted in an accuracy of 86.53%.

Meanwhile, [16] performed sentiment analysis with input from multimodal text data (text, image, and infographic). Classification is divided into five classes: highly positive, highly negative, positive, negative, and neutral. The proposed text classification method is a hybrid method based on SentiCircle and CNN. The accuracy for text classification of the proposed method is 87.8%. In [17], text classification was carried out using a hybrid method based on SentiWordNet and Gradient Boosting, in which the proposed method obtained an accuracy of 82.21%.

Three previous studies carried out sentiment classification in text memes using hybrid methods on non-Indonesian datasets [15,16,17]. On the other hand, for text datasets from social media but not from meme texts, based on these researches [10,11,12,13], individual supervised machine learning methods for sentiment classification also achieved high accuracy. Therefore, this study analyzes and compares four supervised machine learning methods for classifying the sentiment of textual information residing in memes. The study focused on memes' text that is written in Bahasa. This study is part of bigger research focusing on developing methods for classifying sentiment based on graphical and textual information residing in memes. The selection of supervised machine learning methods for comparison is based on the performance of the methods in previous studies on sentiment analysis.

II. RESEARCH METHOD

The research methodology consists of 5 stages: dataset collection, pre-processing, classification model formation, training and testing, and performance evaluation. Figure 1 shows the flow of the research methodology used in this study.

A. Dataset Collection

The dataset used in this study consists of memes in the form of text written in Bahasa. The dataset was obtained from Instagram, Twitter, Facebook, and meme websites (example:

memegenerator.id). The number of comments in the form of text on social media is much more than the number of memes. Therefore, the number of datasets used in this research is only 460. The Indonesian language keywords used to search for the memes within the online sources were *ramadhan, puasa, lebaran, thr, mudik, 17 Agustus, libur, minyak goreng, mandalika*, and *ppkm*.

The authors carried out dataset collection due to the unavailability of an Indonesian meme dataset. The dataset consists of text memes. So the memes collected are memes that contain only text or memes that consist of images and text. The collected data is labeled by a language expert. This expert has a bachelor's degree in Indonesian and has been an Indonesian teacher for over five years. Labeling is done based on text without looking at the image from the meme. The meme's text is grouped into three sentiment labels: positive, negative, and neutral. Figure 2 shows several text memes within the dataset with positive, neutral, and negative labels. Table I shows the characteristics of the dataset. The data set on each label will be seen for its characteristics. Table I shows that the characteristics of text memes with different labels are almost similar with respect to the minimum, maximum, and average word count. So, there is no difference in characteristics on each label even though the amount of data is not the same. Generally, data with a neutral label dominate the number of datasets, while positive labels have the least amount.

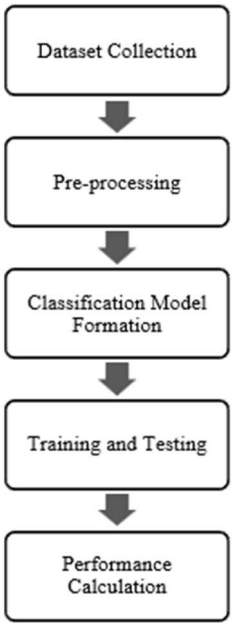


Fig. 1. Research Method

(A) (B) (C)

Fig. 2. Memes are labeled as Positive (A), Negative (B), and Neutral (C).

TABLE I. CHARACTERISTICS OF THE DATASET

Descriptions	Amount of Data		
	Positive	Negative	Neutral
Total number of data	89	155	216
The minimum length of the sentences	3	3	2
The maximum length of the sentences	29	30	32
The average length of the sentences	9.7	10	10.4

B. Pre-processing

Before pre-processing, the collected data was first labelled. In this study, labelling was carried out by experts, namely Indonesian language experts who possess, at minimum, a bachelor's degree in Bahasa.

Furthermore, data pre-processing was carried out by means of data cleaning, case folding, stop word removal, tokenization, and stemming. Then, feature extraction was performed using the TF-IDF (Term Frequency-Inverse Document Frequency) method. Feature extraction is done by assigning a value to each word in the training data. The TF-IDF value of a word *w* depends on term frequency (TF), which is the frequency of occurrence of the word *w* in document *d* compared to the total number of words in document *d*, and inverse document frequency (IDF), which is the proportion of documents that contain the word *w*. TF, IDF, and the TF-IDF value of a word are calculated as eq (1)-(3).

$$TF_{wd} = wf_d / Nwords_d \quad (1)$$

$$IDF_w = \log(Ndocs/df_w) \quad (2)$$

$$TF - IDF_w = TF_{wd} \times IDF_w \quad (3)$$

Where wf_d is the frequency of occurrence of word *w* in document *d*, $Nwords_d$ is the number of words in document *d*, df_w is the number of documents that contain the word *w*, and *Ndocs* is the total number of documents [18].

C. Classification Model Formation

In general, there are three principal sentiment analysis approaches machine learning, lexicon-based, and hybrid approaches [2,14,19]. The methods used in this research are machine learning methods. The selection of methods to be compared in this study follows the model given by [20,21,22], as shown in Figure 3. The methods that are compared in this study are Nave Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), and Convolutional Neural Network (CNN).

D. Training and Testing

In this stage, the dataset is divided into two, namely training data and testing data. Data splitting is done to determine the minimum training data needed to produce good accuracy results. In the first scenario, the dataset is split into 90% training data and 10% testing data. In the second scenario, the dataset is divided into 80% training data and 20% testing data until the ninth scenario, where the dataset is split into 10% training data and 90% testing data. In each scenario, the split of the dataset into training and testing data is done randomly. The testing stage is carried out to validate the model that had been trained previously.

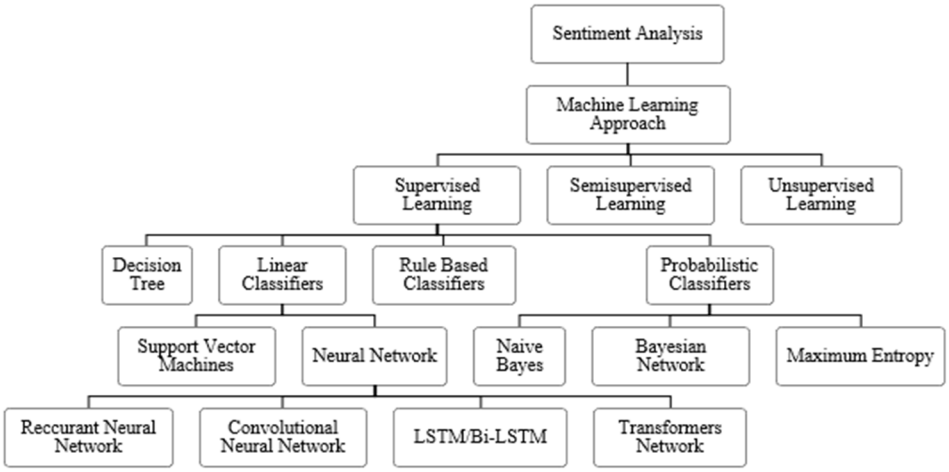


Fig. 3. Different methods of machine learning for sentiment analysis.

In the CNN method, validation is carried out using the testing data at each epoch. The model that had the best performance was stored using the *Keras* library. The CNN was trained for 100 epochs using the Adam optimizer.

E. Performance Evaluation

The performance evaluation in this study was carried out using a confusion matrix for multiclass classification [23], with three classes, namely positive, neutral, and negative. The accuracy, precision, recall, and F1-score values for each model were calculated based on the confusion matrix. These values for each model were calculated for each scenario of data splitting. To test the consistency of the models, for each scenario of data splitting, each model was tested ten times. During testing, the amount of data for each class were randomly selected. The final performance scores were calculated from the average of the results of the ten trials and the standard deviation. The trials were carried out on the initial imbalanced dataset and the balanced version of the dataset due to under sampling.

III. Results and Discussion

The dataset was collected manually from September 2021 to April 2022. The dataset contained a total of 460 memes, with 155 memes (34%) labeled as a negative meme, 216 memes (47%) labeled as a neutral meme, and 89 memes (19%) labeled as a positive meme. This section details the text meme sentiment analysis results using the DT, SVM, NB, and CNN algorithms.

The experiment was conducted based on nine different scenarios of data splitting. Table II shows the number of testing data and the percentage of negative, neutral, and positive data for each scenario. The objective is to determine the model's performance based on varying training and testing data. The data splitting was done randomly by the system, leading to different number of negative, neutral, and positive data for each trial. This ramdom data splitting then causes the varying performance of each model in each trial. Each model was tested ten times for each scenario of data splitting. Table II shows that in each scenario, the percentage of negative, neutral, and positive data of the testing data is almost the same as the percentage of the whole dataset.

TABLE II. Number of Testing Data and Percentage of Each Label

Testing Data	Negative (%)	Neutral (%)	Positive (%)
46	32.4	50.2	17.4
92	34.9	48.0	17.1
138	36.4	46.1	17.5
184	35.9	45.8	18.3
230	35.4	46.3	18.2
276	35.9	45.7	18.4
322	35.2	46.4	18.4
368	34.6	46.7	18.7
414	34.0	47.1	18.9

TABLE III. Confusion Matrix Naïve Bayes Method (Data Testing 20%)

		Predicted class			Total
Trial 1		**Negative**	**Neutral**	**Positive**	
True class	Negative	12	18	0	30
	Neutral	2	47	0	49
	Positive	2	11	0	13
Trial 2					
True class	Negative	5	23	0	28
	Neutral	5	41	0	46
	Positive	4	14	0	18

Table III shows the confusion matrix results for two different trials using the NB method with the same amount of testing data, namely 20% testing data which is 92 memes. To test the consistency of the models, each model was tested more than once for each scenario. It can be seen from Table III that in the first trial, the NB method succeeded in correctly predicting the labels of the text memes more than in the second trial. This is because each test has different amounts of data per label. The neutral class has a high dominance within the dataset. When using machine learning for classification, the highly unbalanced classes in the dataset will affect the classification results because the classification model tends to be more biased towards the majority class. This results in more training samples from the majority class being considered of

higher importance by the model. In contrast, samples from the minority might even be regarded as noisy data [23, 24].

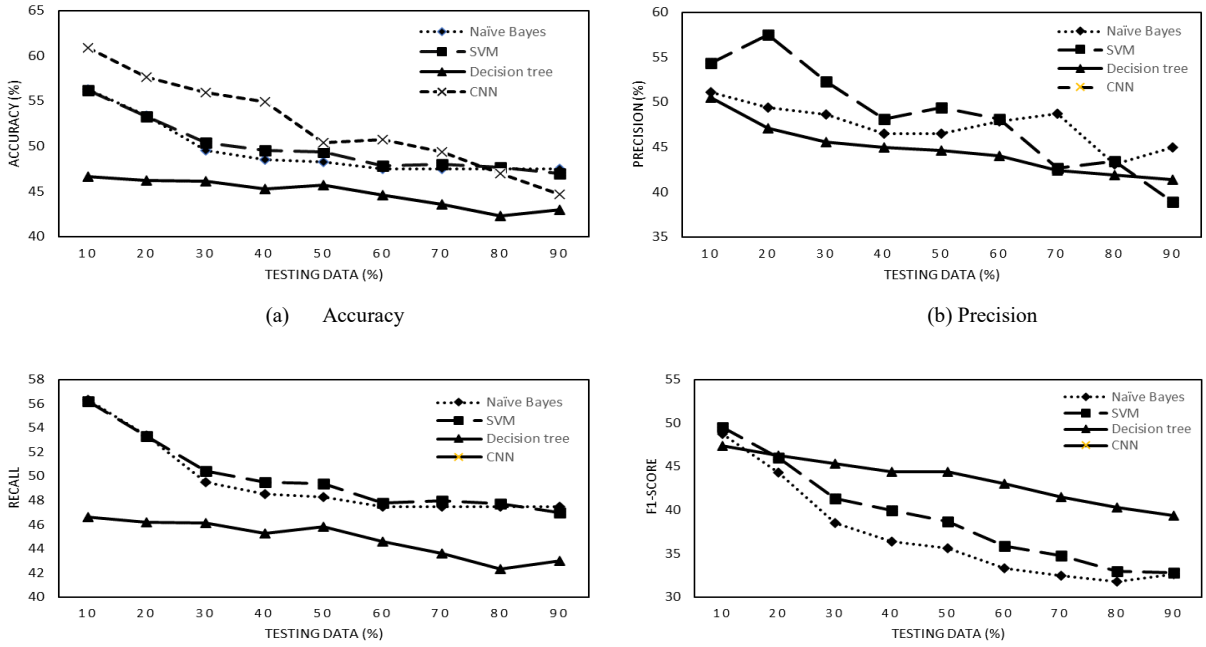
Based on the resulting confusion matrix of the experiment, for all scenarios of data splitting, it can be concluded that: (1) the testing samples tend to be classified into the majority class, namely the neutral class, and so the more testing samples that are labeled neutral, the higher the probability that the prediction will be correct. This, in turn, affects the models' performance and results in an increased value of accuracy, precision, recall, and F-1 score. (2) With the NB method, the positive class samples were never correctly predicted in all the scenarios of data splitting. (3) The testing samples of the neutral class were correctly predicted with an accuracy between 92%-100% by the NB and SVM models. Meanwhile, CNN can correctly predict testing samples of the neutral class with an accuracy between %-%, and DT managed to correctly predict testing samples of the neutral class with an accuracy between 60%-80%.

In the experiment conducted in this study, ten trials were carried out for each of the nine scenarios of data splitting. The performance evaluation scores were calculated by calculating the average and standard deviation of the scores in each scenario of data splitting. The average value is used to compare the performance of all the models.

Figure 4 shows the experimental results of all four models in each scenario of data splitting, starting from 10% testing data to 90% testing data. The obtained accuracy, precision, recall, and F1-score values of all four models are also presented in Figure 4. The x-axis represents the percentage of testing data, while the y-axis represents the classification results based on accuracy, precision, recall, and F1-score values.

Based on Figure 4, it can be seen that the best performance for all four models was obtained when the number of testing data was 10% of the overall data. The performance of all four models decreased when the number of testing data was increased to 20% from the overall data, which also meant that the training data decreased. The performance of all four models tends to be stable when the number of testing data is above 50%. In this study, when the number of testing data was 10% of the overall data, the CNN model obtained the highest average accuracy value with an accuracy of 61.1% ± 1.9%. This means that from 10 trials, the accuracy value obtained by the CNN method ranged from 59.2%-63%. The NB and SVM models produced very similar accuracy values. The average accuracy obtained by the NB model, SVM model, and DT model was 56.3% ± 13.0%, 56.2% ± 12.3%, and 46.4% ± 7.6%, respectively. Overall, in all nine scenarios of data splitting, based on standard deviation, the DT model was the most stable in performance compared to the other models, even though its performance was not very good.

Overall, all four models produced a similar performance for sentiment analysis on Indonesian text meme data, in which the highest average accuracy value of all four models was less than 65%. When the number of testing data was more than 50% of the overall data, the average accuracy of all four models was less than 50%. Compared to the hybrid methods proposed in previous studies [15, 16, 17], the performance of the four models could not outperform the hybrid methods. The method proposed by [16], a hybrid of SentiCircle and ConvNet (convolution neural network), even achieved an accuracy of 87.8%. The data used in the sentiment analysis conducted by the study was in the form of multimodal text data (text, image, infographic). The classification was divided into five classes: highly positive, highly negative, positive, negative, and neutral.



(a) Accuracy (b) Precision

(c) Recall (d) F1-Score

Fig. 4. Classification Performance of the Models

The unsatisfactory performance of all four models was partly due to the highly imbalanced collected text meme dataset. The number of data labeled as neutral was significantly more significant than the number of data labeled as positive and negative. Improved classification performance for this type of data can be achieved by balancing the amount of data in each class or modifying the architecture of the classification model. For example, the amount of data can be balanced by means of under-sampling or over-sampling [25].

Furthermore, in this study, under-sampling was carried out to see if there was an increase in the performance of the models and if the number of classes within the dataset was more balanced. From the dataset, 116 samples labeled as neutral, 95 samples labeled as unfavorable, and 89 samples labeled as positive were taken randomly, so the total number of data used for further testing was 300. The classification was done by dividing the dataset into 90% training data and 10% testing data based on the best performance achieved by all four models in the previous test. The steps carried out in this test were the same as the previous test, namely conducting ten trials and calculating the average performance of each method. In addition, the standard deviation was also calculated to determine the variation in results.

The classification results of all four machine learning models on the new dataset shown are presented in Table IV. All the models on the new dataset exhibited an increase in performance, except for the CNN model. The NB model achieved the highest accuracy value. From the ten trials, the highest accuracy value was 72%, the lowest accuracy value was 62%, and the highest average accuracy value was 65.4%, with a standard deviation of 4.8%. By using almost, the same amount of data per class, the DT model produced an average accuracy that was better than before, namely 63.8% ± 9.5%, while SVM achieved an average accuracy of 62.8% ± 5.3%. Other performance metric values (precision, recall, and F1-score) were similar to the previous tests, which were above 60%.

TABLE IV. MACHINE LEARNING ALGORITHM PERFORMANCE

	Performance (%)			
Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
NB	65.4	69.6	65.4	62.8
SVM	62.8	65.0	62.8	60.0
DT	63.8	63.2	62.6	62.6
CNN	60.8	61.3	60.8	60.0

IV. CONCLUSION

In this study, we compared the performance of several sentiment analysis methods that have been trained to detect sentiment in text memes. The dataset used in this study consists of memes in the form of text written in Bahasa. Sentiment classification was carried out using four supervised machine learning algorithms: NB, SVM, DT, and CNN. These four algorithms were chosen because they have achieved good performance for text sentiment analysis. The experimental results show that using an imbalanced dataset, the CNN method, one of the deep learning methods, successfully predicted the sentiment of text memes with an average accuracy value of 61.1% ± 1.9%.

Meanwhile, the other three methods achieved an average accuracy value below 60%. These three methods, on average, fail to predict the classification correctly for the minor number of datasets. The second test was then carried out on a balanced dataset obtained by means of the under-sampling technique. The second test's results showed increased accuracy of the NB, SVM, and DT methods. The experimental results show that the NB method succeeded in predicting the sentiment of the text memes with an average accuracy value of 65.4% ± 4.8%.

Research to improve the performance of the models in predicting sentiment in text memes can be carried out by modifying the architecture of the classification models, for example, by adding pre-trained word embedding to the classification of the deep learning method. Therefore, in future studies, the effect of the modification to the architecture of the classification models on the performance of the original model on an imbalanced dataset will be investigated.

REFERENCES

[1] B. Liu, “Sentiment analysis and opinion mining,” Chicago: Morgan & Claypool, 2012

[2] F.A. Pozzi, E. Fersini, V. Messina, B. Liu B, “Challenges of sentiment analysis in social networks: an overview. In: Pozzi FA, Fe\][=rsini E, Messina E, Liu B (eds) Sentiment analysis in social networks,” Morgan Kaufmann, Burlington, 2017, pp 1–11

[3] A.D’Andrea, F. Ferri, P. Grifoni, T. Guzzo, "Approaches, tools, and applications for sentiment analysis implementation," International Journal of Computer Applications 125(3): 26-33, 2015.

[4] R. Hasyim, B.Omar, N.S.A. Ba-Anqud, H. Al-Samarraie, “The Application of sentiment analysis in tourism research: A Brief review," International Journal of Business Tourism and Applied Sciences, Vol. 8, No.1, 2020

[5] C. N Devi, R. R. Devi, “Literature review on sentiment analysis in tourism," Test Engineering and Management, Volume 83, Page Number: 2466 - 2474 Publication Issue: March - April 2020

[6] A. Kumar, A. Sharma, “Systematic literature review on opinion mining of big data for government intelligence,” Webology Volume 14, Number 2, pp: 6–47, 2017

[7] A. Avvaru, S. Vobilisetty, "BERT at SemEval-2020 Task 8: Using BERT to analyze meme emotions,” Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1094–1099 Barcelona, Spain (Online), December 12, 2020

[8] https://kbbi.kemdikbud.go.id/entri/meme

[9] M. Mertiya, A. Singh, "Combining Naive Bayes and adjective analysis for sentiment detection on Twitter," Proc. Int. Conf. on Inventive Computation Technologies (ICICT) 2016, vol. 2, 2016

[10] A. Hassan, A. Mahmood, “Deep learning approach for sentiment analysis of short texts,” In Proceedings of the Third International Conference on Control, Automation and Robotics (ICCAR), Nagoya, Japan, 24–26 April 2017; pp. 705–710.

[11] J. Qian, Z. Niu, C. Shi, “Sentiment analysis model on weather-related tweets with a deep neural network,” In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018, pp. 31–35

[12] L.K. Ramasamy, S. Kadry, Y. Nam, M.N. Meqdad, “Performance analysis of sentiments in Twitter dataset using SVM models,” International Journal of Electrical and Computer Engineering (IJECE). Vol. 11, No. 3, June 2021, pp. 2275~2284. doi: 10.11591/ijece.v11i3.pp2275-2284

[13] C. Villavicencio, J.J. Macrohon, X.A. Inbaraj, J. Jeng, J. Hsieh, “Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve Bayes,” Information (Switzerland), 2021, 12(5), 204, https://doi.org/10.3390/info12050204

[14] A. Ligthart, C. Catal, B. Tekinerdogan, “Systematic reviews in sentiment analysis: a tertiary study,” Artificial Intelligence Review, 2021, 54:4997–5053, https://doi.org/10.1007/s10462-021-09973-3

[15] T.N Prakash, A. Aloysius, “Hybrid approaches based emotion detection in memes sentiment analysis,” International Journal of Engineering Research and Technology, vol. 14, issue 2, 2021, pp. 151-155

[16] A. Kumar, K Srinivasan, C Wen-Huang, A.Y. Zomaya, “Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data,” Information Processing and management, 57, 2020, 102141. https://doi.org/10.1016/j.ipm.2019.102141

[17] A. Kumar, G. Garg, “Sentiment analysis of multimodal twitter data,” Multimedia Tools and Applications, 78, pages 24103–24119, 2019, https://doi.org/10.1007/s11042-019-7390-1

[18] G. Zhao, Y. Liu, W. Zhang, Y. Wang, “TFIDF based feature words extraction and topic modeling for short text,” Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences, January 2018, Pages 188–191, https://doi.org/10.1145/3180374.3181354

[19] K.M.M. Baidal, C.D. Vera, E.T.S. Aviles, A.P.H. Espinoza, “Sentiment Analysis in Education Domain: A Systematic Literature Review," 4th International Conference, CITI 2018, Guayaquil, Ecuador, November 6-9, 2018

[20] W. Medhat, A. Hassan, H. Korashy, “Sentiment analysis algorithms, and applications: A survey," Ain Shams Eng. J. 2014, 5, 1093–1113

[21] B. Bhavitha, A.P. Rodrigues, N.N. Chiplunkar, “Comparative study of machine learning techniques in sentimental analysis," In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11 March 2017; pp. 216–221

[22] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, “Sentiment analysis based on deep learning: A Comparative study,” *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483. [Online]. Available: http://dx.doi.org/10.3390/electronics9030483

[23] H.F. Soon, A. Amir, S.N.Azemi, “An Analysis of Multiclass Imbalanced Data Problem in Machine Learning for Network Attack Detections”, 5th International Conference on Electronic Design (ICED), Journal of Physics: Conference Series 2020, doi:10.1088/1742-6596/1755/1/012030

[24] Y.E. Ardiningtyas, P.H.P Rosa, “Analisis Balancing Data Untuk Meningkatkan Akurasi Dalam Klasifikasi”, Seminar Nasional Aplikasi Sains & Teknologi (SNAST), Yogyakarta, 20 Maret 2022

[25] F. Thabtah, S.Hammound, F. Kamalov, A. Gonsalves, "Data imbalance in classification: Experimental evaluation," Information Sciences: an International Journal, vol. 513, Maret 2020, pp 429-441