Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 25 NO 03 2024, pp 362-371
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol25-iss03/384

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Classification of Covid-19 Vaccine Sentiment Using K-Nearest Neighbor and Fasttext on Twitter

Article Info

Article history :

Received December 25, 2022
Revised September 20, 2024
Accepted September 28,2024
Published September 30, 2024

Keywords :
K-Nearest Neighbor,
Fasttext,
Sentiment Classification,
Covid-19 Vaccine

Afri Naldi[1], Surya Agustian[1*]

[1]Department of Informatic Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Riau, Indonesia

Abstract. In late 2019 came a flu-like illness that infected the lungs in the city of Wuhan. It is suspected that the disease is suspected to have originated in bats. WHO named this disease Covid-19 and the virus spread throughout the world, causing a pandemic. The government took a vaccination drive to overcome this virus, but received a response of pros and cons from the public. There are many studies that discuss people's sentiments towards vaccination, one of which is the classification of sentiments. This study discusses the classification of sentiment towards covid-19 vaccines using the K-Nearest Neighbor and Fasttext algorithms on twitter. Data is obtained by crawling using the pyton programming language and Twitter API. Data labeling is carried out by crowdsourcing and majority voting techniques. The data used after the balancing process are 6000 training data, 778 development data and 400 test data. The test results after various experiments and feature engineering got the best results with an accuracy value of 69% and an f1-score of 60%. This result is the best result compared to previous studies with the same dataset.

Corresponding Author :
Afri Naldi
Department of Informatic Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Riau, Indonesia
Email : 11651103444@students.uin-suska.ac.id

1. Introduction

At the end of 2019 on December 31, there were several reports of a disease of unknown etiology in China precisely in Wuhan City, Hubei province. This disease has symptoms such as fever, dry cough, dyspnea, and infection of the lungs, then all cases related to this disease are found in the city of Wuhan precisely in the seafood market that sells various types of live animals such as bats, snakes, and poultry [1-2].

Officially, the World Health Organization (WHO) called this disease Covid-19 on February 11, 2020. Then until January 30, 2020, this disease continued to develop rapidly until many countries were affected or infected with this disease, so WHO declared that Covid-19 was a threat to the world. The COVID-19 virus or can be called the corona virus first entered Indonesia on March 2, 2020, which was infected by two West Java residents precisely in the city of Depok. Since then, there have been many reports that people have been confirmed to have the coronavirus that has hit the world. This condition causes side effects not only to the health sector but the economic, educational, and other fields affected by the coronavirus that has hit the world [3-4].

After the pandemic, the government took a policy to reduce the spread of this corona virus by using vaccines [5]. In accordance with the decree of the Minister of Health number HK.01.07 / Menkes / 12758 / 2020 which has been inaugurated precisely on December 28, 2020, it has determined several types of vaccines that are allowed to be used, namely Bio Farma, Sinovac, Moderna, Sinopharm, Novavax, and Pfizer [6]. The use of vaccines in Indonesia has reaped public opinions that contain positive and negative opinions about the vaccineization that the government will do to reduce the level of I-7 spread. Social media became a place to express freely his opinion about this vaccineization, one of the social media used was Twitter [4].

Twitter is a place for people to give their opinions about the vaccineization that has been put forward, there are some tweets that have positive, negative, and neutral values such as tweets from @filoshopiee accounts "GSHSGSHGSYSG I'M EXCITED TOMORROW I GET VACCINATED you guys stay safe and healthy yaa!!", and then there are also those who argue like a tweet from the account @widjaja_harta "Why can such a sebego person sit in the House seat? Refusing the vaccine? Willing to pay 5 million per person for his family members. SELFISH, STUPID and IGNORANT.....". Soit can be concluded that the response from the application of this vaccineization not only received a good response but an unfavorable response was also contained in it.

APJII or also known as the Indonesian Internet Service Providers Association has conducted a survey of internet use in Indonesia which is around 196.7 million people in active use in the second quarter of 2020 [7]. This received an increase in users by 8.9% or around 25.5 million people compared to 2018. Twitter, Instagram and Facebook became the social media of choice of the most used people.

There are several previous studies that are used as references in this study as follows. The classification of sentiment towards the COVID-19 vaccine using the Naïve Bayes method carried out by selecting a combination of text [8], dataset balancing, setra parameter tunnig resulted in the best accuracy of 61% and an f1-score of 57.15%. Furthermore, the classification of COVID-19 sentiment using the Support Vector Machine with the same dataset got the best model with an accuracy of 65% and [9] an f1-score of 56.81%. Sentiment classification using deep learning techniques using Long-short term memory with the same dataset [10] gets a 54% f1-score result with 66% accuracy.

Tendency of community responses to the sinovac vaccine based on lexicon based senteiment analysis [11]. The purpose of this study is to see opinions from the public about the application of the vaccine, the results of this study have the highest percentage is neutral (37.6%) and 100% of the data used.

The Analysis of Public Opinion Sentiment about Vaccination in Indonesia Using Naïve Bayes and Decission Tree [12]. From the research that has been carried out, the results were obtained that the Naïve Bayes and Decission Tree methods can be done with the accuracy obtained for naïve bayes of 100% and the accuracy value of Decission Tree of 50.39%.

Regarding the title of the research that will be carried out using the K-Nearest Neighbor method, there are several related references including, such as the research conducted by Ernawati, et al regarding sentiment analysis in travel agents using the [13] K-Nearest Neighbor method. The data used in the research they conducted was 200 reviews, after that the data will be divided into 100 positive data and 100 negative data, then the data that has been previously divided will be used as several labels, namely as many as 6 sentiments such as Wait, Fast, Good, Bad, Great, Cancel and the

results obtained in the study conducted by testing the value of k = 8 were 87% with an AUC indigo of 0.916 entered the Excellent Classification group.

Based on the background that has been conveyed above, in this study, the author will analyze the K-Nearest Neighbor method in classifying sentiment analysis of vaccineization that occurs in Indonesian tweets by applying the use of word embedding Fasttext and the use of features to improve the quality of the model so as to increase the accuracy of the model to be obtained. This research was conducted on the dataset used in the study, and to evaluate and compare the performance of the Decision tree and XGBoost against the three methods [10][14].

2. Experimental Section

2.1. Materials

Method is the stage of solving a problem in research so that the implementation of research is in accordance with the objectives. Broadly speaking, this study has several stages, such as: problem identification, data analysis, method analysis, tuning parameters, and testing.

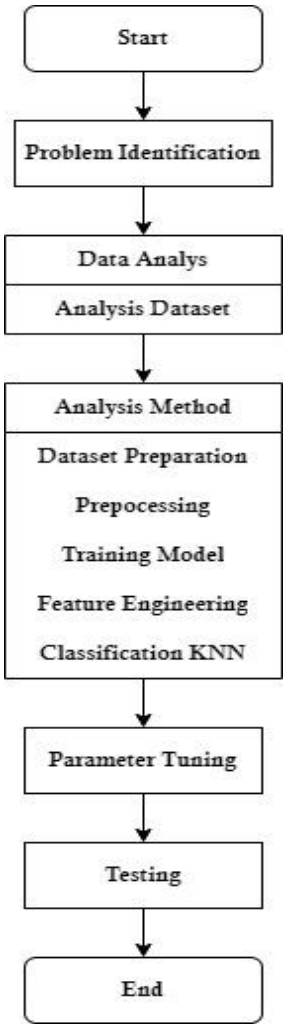


Figure 1. Research Methodology

2.2. Tips

2.2.1 Dataset

The data collected is data on Indonesian-language tweets from March 2021 to April 2021. Data is collected using python programming by crawling using the Twitter API. The keywords used are filtered in such a way that the tweets obtained are related to the COVID-19 vaccine, keywords are: "Vaccination Indonesia", "Vaccine", "Sinovac Vaccine", "Nusantara Vaccine", "Gotong Royong Vaccine", "Free Covid Vaccine", Voluntary Vaccine", "Covid Vaccine", "Corona Vaccine" and keywords referring to the covid vaccine, so that the total *tweets* obtained are 13,115 tweets.

2.2.2 Data Analys

At this stage, the data labeling process is carried out, namely positive, neutral, and negative labels. Positive labels contain positive feelings such as fun, joy, praise, support, and positive suggestions. While negative labels usually contain complaints, disappointments, dissatisfaction, insults, to hate speech. The neutral label contains data that does not fall under the criteria for positive or negative data.

Table 1. Data Labeling Example

Sentiment	Tweet
Positive	Indonesia means that it has received 59.5 million doses of vaccine raw materials from Sinovac with the arrival of againâ€¦ https://t.co/vreS13Cv3y
Neutral	More than a million doses of the Pfizer Covid-19 vaccine will be received by the kingdom next month, said the Minister of Science, Technology and Inoâ€¦ https://t.co/iJtldtFWsa
Negative	Then if you have all been vaccinated, guarantee it is safe? it's also troublesome, if the superior2 in the center has underestimated, even downwards iâ€¦ https://t.co/7to7zjDxv7

After the data labeling process is carried out, the data sharing process is then carried out. The data division carried out is 3 parts, namely train with a total of 8000 data, validation data or development data as many as 778 data, and test data or test data as many as 400 data. Training data is data that will be used as training data for the model to be built. Development data is data that will be used to validate the model to be built. Test data is a type of tweet data designed to test a model.

2.2.3 Prepocessing

In this step, carry out the data process so that the data is ready for processing and analysis aims to make the data to be processed more regularly. The steps taken in this process, namely case folding, tokenizing, punctuation, stopwordremoval and also using several additional steps aimed at adjusting the combination to be done.

2.2.4 Training Model

Training Model is a stage to carry out the process of weighting a word into a vector, the weighting used in this study is using fasttext. In the research that has been carried out, obtaining the results of [15] fasttext analysis has the best performance compared to using 2 other algorithms (glove, word2vec) which can train models from large datasets, fast and can provide representations of words that do not appear in the training data. If the word does not appear then it can be broken down into n-grams to obtain a vector. Fasttext feature extraction that can only convert words into a vector that has dimensions 128 long.

2.2.5 Min Max Scaler

MinMaxScaler is a stage of prepocessing where it transforms features by scaling features in an individual way gradually within a certain range, aiming so that the range of each feature is not too large. This prepocessing is used so that there is a reduction in the sample with the smallest sample value in the feature and a division is carried out between the largest sample value reduced by the smallest sample in the feature [16].

$$X_{sc} = \frac{\mathrm{X} - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

2.2.6 Robust Scaler

RobustScaler is a stage of optimization technique to transform a value by using the median stages and quartiles to get better modeling results, this stage is included in the process of improving through data to eliminate awkward values so that the results of modeling do not enter into oversight [17].

$$\|x\|_q \triangleq \left(\sum_{i=1}^{n} |x_i|^q\right)^{\frac{1}{q}} \qquad (2)$$

2.2.7 K-Nearest Neighbor (K-NN)

The working principle of the K-Nearest Neighbor algorithm is to find the closest distance between the data to be evaluated and the nearest neighbor k in the training data [18]. Here is the working process of the K-Nearest Neighbor algorithm:

Specify the parameter k the number of nearest neighbors then calculate the *Euclidean Distance* of each object against the sample existing data [18].

$$d_i = \sum_{i=1}^{p} (x_{2i} - x_{li})^2 \qquad (3)$$

Then sort the objects into groups that have small Euclidean distances. After being sorted according to the smallest Euclidean distance, then adjustments were made to category Y (Nearest Neighbor Classification). By μ is the mean of the sample, and σ is the standard deviation of the sample [20].

2.2.8 Confusion Matrix

Confusion Matrix is a stage to see the performance of the algorithm or model that has been carried out. The result of the confusion matrix presents the results of the actual class of each data that has been processed and the column presents the prediction class of the data that has been processed [21].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

$$Precision = \frac{TP}{TP + FP} \qquad (5)$$

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

$$F_1 - Score = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP\ TP + FN\ TN + FP\ TN\ + FN)}} \qquad (7)$$

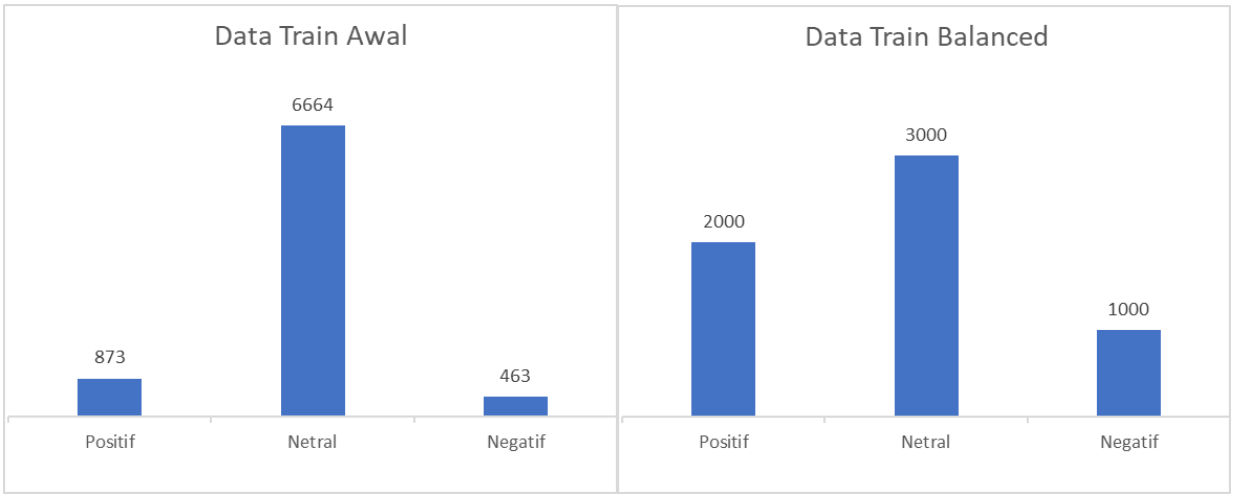
3. Results and Discussion

3.1 Data Train Balancing

Figure 2 shows the difference between the original (imbalanced) data and the balanced data. In the original data, the class distribution is very imbalanced, with the neutral class dominant (6664 data), followed by the negative class (873 data), and the positive class (463 data). This chaos is often a problem in data classification, because the model tends to ignore the minority class and focuses more on the majority class, which in this case is the neutral class.

After balancing is done through the RandomOverSampler technique, each class is equalized to 3000 data. This process is important to ensure that the model has enough data from each class to learn fairly. However, it should be noted that the oversampling process also has the risk of overfitting, especially when the minority data is artificially inflated. While oversampling is effective in dealing with data synchronization, it can also cause the model to overfit the training data, reducing generalization to new data [22-23].

Data that has gone through the oversampling stage has a ratio of 1:1:1, which is potentially overfitting because the initial ratio of data is 1:14:2. Positive and negative label data will be trimmed to form a 1:3:2 data ratio. Figures 2 below are visualizations between the initial train data and the train data that has gone through the balancing process.



(a) (b)

Figure 2. (a) Initial Train Data, (b) Balanced Train Data

3.2 Preprocessing Text Combinations

At this stage will be carried out the search for the best combination of preprocessing texts. In addition to selecting data, variations in preprocessing text are carried out to find the best combination because the preprocessing text depends on how the dataset is conditioned, it can cause high accuracy or even decrease the accuracy value. The experiments to be carried out can be seen in the following table.

Table 2. Text Processing Experiments

Case Folding	Stopword	Punctuation	ID Eksperimen
No	No	No	C1
Yes	No	No	C2
No	Yes	No	C3
No	No	Yes	C4
No	Yes	Yes	C5
Yes	No	Yes	C6
Yes	Yes	No	C7
Yes	Yes	Yes	C8

Table 2 shows the results of various text processing combination experiments, involving Case Folding, Stopword Removal, and Punctuation Removal techniques. Of the eight combinations tested, experiment C2 (Case Folding without Stopword and Punctuation Removal) produced the highest accuracy of 62% with an F1-score of 56%. This shows that not all text processing techniques need to be used together to improve model performance.

Case Folding, which converts all text to lowercase, has been shown to be important for matching word forms. However, removing stopwords and punctuation does not always improve model performance. Stopword removal can sometimes remove important words that have context in sentiment analysis, which can reduce model accuracy [24-25].

Model validation against dev data gets the best combination in C2 experiments with 62% accuracy and 56% f1-score. Furthermore, the testing phase will be carried out using the MinMaxScaler and RobustScaler feature engineering. At this stage, the C2 experimental model will be tested against the test data. Test data is data that is not yet known by the model when conducting training. The tests that will be carried out are testing without scaling, using MinMaxScaler, and RobustScaler while looking for the best K value parameters.

Table 3. Test Results

Feature	Value K	Accuracy	F1-Score
NonScaler	13	0.68	0.59
Minmax	19	0.68	0.59
Robust	17	0.69	0.60

Table 3 shows the test results using various Scaler techniques in feature processing, such as Non-Scaler, MinMaxScaler, and RobustScaler. The test results show that RobustScaler gives the best results with an accuracy of 69% and an F1-score of 60%, compared to MinMaxScaler and Non-Scaler, which each give an accuracy of 68% and an F1-score of 59%.

RobustScaler is a data normalization technique that uses the median and interquartile range (IQR) to eliminate the influence of outliers, in contrast to MinMaxScaler which stretches data values on a certain scale. RobustScaler is proven to be more effective in dealing with outliers in the data, which may occur in tweets containing very extreme sentiments, both positive and negative [26-27]. The results of comparison with previous research methods are in the following table:

Table 4. Comparison Results

Method	Accuracy	F1-score
Naïve Bayes(1	61%	57%
SVM (1	65%	56%
LSTM (1	66%	54%
KNN	69%	60%

Table 4 shows a comparison of the results between the K-Nearest Neighbor (KNN) method used in this study with other methods that have been used previously, such as Naïve Bayes, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). The best results were obtained from the KNN method, with an accuracy of 69% and an F1-score of 60%, which outperformed the other methods: Naïve Bayes with an accuracy of 61%, SVM with 65%, and LSTM with 66%.

The superiority of KNN compared to other methods in this study can be attributed to its simplicity in calculating the distance between data points and its flexibility in handling text data after being processed with FastText. KNN is effective for text classification because this model works well with vector-based text representations such as FastText, which have been proven effective in solving word representation problems in lower dimensions [28-30].

4. Conclusion

In conclusion from the results of the tests that have been carried out, it can be concluded that the best KNN model obtained is the KNN model of the C2 experiment using RobustScaler and a K value parameter of 17. The model obtained the best accuracy and f1-score results compared to previous research methods with an accuracy value of 69% and an f1-score of 60%.

The result of this study is a KNN model that can classify sentiment towards the COVID-19 vaccine. This model can also be used as a comparison with other methods and can also be developed by creating an application-based system to be able to predict sentiment through manual sentence input. Advice that can be given to subsequent researchers is to increase the portion of positive and negative label data so that the data is more balanced and does not need to undersampling and oversampling, so that it is hoped that it will improve the quality of the model and improve classification results

References

[1] Makmun, A., & Hazhiyah, S. F. (2020). Tinjauan Terkait Pengembangan Vaksin Covid 19. *Molucca Medica*, 52-59.

[2] Haque, A., & Pant, A. B. (2020). Efforts at COVID-19 vaccine development: challenges and successes. *Vaccines*, *8*(4), 739.

[3] Sohrabi, C., Alsafi, Z., O'neill, N., Khan, M., Kerwan, A., Al-Jabir, A., ... & Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International journal of surgery*, *76*, 71-76.

[4] Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta bio medica: Atenei parmensis*, *91*(1), 157.

[5] Laurensz, B., & Sediyono, E. (2021). Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, *10*(2).

[6] Keputusan Menteri, Keputusan Menteri Kesehatan Republik Indonesia Nomor Hk.01.07/Menkes/12757/2020 Tentang Penetapan Sasaran Pelaksanaan Vaksinasi Corona Virus Disease 2019 (Covid-19), *Keputusan Menteri Kesehatan Republik Indonesia Nomor Hk.01.07/Menkes/12757/2020 Tentang Penetapan Sasaran Pelaksanaan Vaksinasi Corona Virus Disease 2019 (Covid-19)*, vol. 284, pp. 99–119, 2020.

[7] Mutikasari, A. D., & Susila, I. (2023). Analysis Factors Affecting Customer Loyalty of Indihome Provider During The Covid-19 Pandemic In Surakarta. *Jurnal Pamator: Jurnal Ilmiah Universitas Trunojoyo*, *16*(4), 727-744.

[8] Pristiyono, Ritonga, M., Ihsan, M. A. A., Anjar, A., & Rambe, F. H. (2021, February). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012045). IOP Publishing.

[9] Putraa, F. M., & Santiyasaa, I. W. Sentiment Analysis of the Indonesian Health Ministry Performance in Covid-19 Crisis using Support Vector Machine (SVM). *Jurnal Elektronik Ilmu Komputer Udayana p-ISSN*, *2301*, 5373.
[10] Ihsan, M., Negara, B. S., & Agustian, S. (2022). Metode LSTM (Long short term memory) untuk Klasifikasi Sentimen Vaksin Covid-19 pada Twitter. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(1), 1-13.
[11] Kahraman, E., Demirel, S., & Gündüz, U. (2023). COVID-19 vaccines in twitter ecosystem: Analyzing perceptions and attitudes by sentiment and text analysis method. *Journal of Public Health*, 1-15.
[12] Harun, A., & Ananda, D. P. (2021). Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve bayes dan Decission Tree: Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 58-64.
[13] Ernawati, S., & Wati, R. (2018). Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel. *jurnal khatulistiwa informatika*, *6*(1).
[14] Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1308.
[15] Fibrianda, M. F., & Bhawiyuga, A. (2018). Analisis Perbandingan Akurasi Deteksi Serangan Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(9), 3112-3123.
[16] Prasetyo, V. R., Mercifia, M., Averina, A., Sunyoto, L., & Budiarjo, B. (2022). Prediksi Rating Film Pada Website Imdb Menggunakan Metode Neural Network. *Nero (Networking Engineering Research Operation)*, *7*(1), 1-8.
[17] Bertsimas, D., Dunn, J., Pawlowski, C., & Zhuo, Y. D. (2019). Robust classification. *INFORMS Journal on Optimization*, *1*(1), 2-34.
[18] Ernawati, S., & Wati, R. (2018). Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel. *jurnal khatulistiwa informatika*, *6*(1).
[19] Taufiqurrahman, T., Nababan, E. B., & Efendi, S. (2021). Analysis of dimensional reduction effect on K-Nearest Neighbor classification method. *Sinkron: jurnal dan penelitian teknik informatika*, *5*(2B), 222-230.
[20] Arsi, P., Hidayati, L. N., & Nurhakim, A. (2022). Komparasi model klasifikasi sentimen issue vaksin COVID-19 berbasis platform instagram. *Jurnal Media Informatika Budidarma*, *6*(1), 459-466.
[21] Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creat. Inf. Technol. J*, *6*(1), 1.
[22] Kerwin, K. R., & Bastian, N. D. (2021). Stacked generalizations in imbalanced fraud data sets using resampling methods. *The Journal of Defense Modeling and Simulation*, *18*(3), 175-192.
[23] Kumar, A., Saxena, N., Jung, S., & Choi, B. J. (2021). Improving detection of false data injection attacks using machine learning with feature selection and oversampling. *Energies*, *15*(1), 212.
[24] Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). Effective text data preprocessing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering (KSE)* (pp. 1-8). IEEE.
[25] Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*, 319-335.
[26] Gkaimanis, D. (2024). Stock Market Prediction using Double-DQN and Sentiment Analysis.

[27] Alkurdi, A., & Abdulazeez, A. M. (2024). Comprehensive Classification of Fetal Health Using Cardiotocogram Data Based on Machine Learning. *Indonesian Journal of Computer Science*, *13*(1).
[28] Putra, S. J., Gunawan, M. N., & Hidayat, A. A. (2022, September). Feature engineering with Word2vec on text classification using the K-nearest neighbor algorithm. In *2022 10th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-6). IEEE.
[29] Syaputra, R. A., & Ali, R. (2022). Improving mental health surveillance over Twitter text classification using word embedding techniques. In *Artificial intelligence, machine learning, and mental health in pandemics* (pp. 235-258). Academic Press.
[30] Soleimani, B. H., & Matwin, S. (2019, July). Fast PMI-based word embedding with efficient use of unobserved patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 7031-7038).