JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 85 – 92 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3364
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

THE ROLE OF FEATURE SELECTION IN ENHANCING THE ACCURACY OF AI ASSISTANT AUTO-LABELING

Indri Tri Julianto[1*], Dede Kurniadi[1], Benedicto B. Balilo Jr[2], Fauza Rohman[1]
[1] Computer Science, Institut Teknologi Garut
[2]CS/IT, College of Science, Bicol University
email: *indritrijulianto@itg.ac.id

Abstract: The development of AI assistants such as Gemini and ChatGPT can significantly assist in daily human tasks. In the field of Sentiment Analysis, AI assistants can be utilized as an automated labeling alternative to provide positive, negative, or neutral sentiments within a dataset. This research aims to enhance the performance of AI assistants in automated labeling processes by employing the Feature Selection algorithm, specifically Forward Selection. The methodology involves utilizing the Naïve Bayes and K-NN algorithms, and subsequently improving accuracy through the Feature Selection algorithm. The evaluation is conducted using K-Fold Cross Validation. Research findings indicate an improvement in the accuracy of the best model, which is ChatGPT, when using the Naïve Bayes algorithm and Shuffled Sampling technique. The initial accuracy of 79.09% increased to 87.18% after Feature Selection was applied. This demonstrates the effectiveness of Feature Selection, particularly Forward Selection, in enhancing the accuracy performance of the model.

Keywords: ai; assistant; chat gpt; feature selection; gemini.

Abstrak: Pekembangan Asisten AI seperti Gemini dan Chat GPT dapat membantu pekerjaan manusia sehari-hari. Dalam bidang Analisis Sentimen, Asisten AI dapat digunakan sebagai alternatif pelabelan otomatis untuk memberikan sentimen positif, negatif atau netral dalam suatu dataset. Penlitian ini bertujuan untuk meningkatkan performa yang dihasilkan oleh Asisten AI dalam proses pelabelan otomatis menggunakan Algortima Feature Selection yaitu Forward Selection. Metode yang digunakan adalah dengan menggunakan Algoritma Naïve Bayes dan K-NN kemudian hasil akurasi akan ditingkatkan menggunkan Algoritma Feature Selection. Evaluasi yang digunakan adalah K-Fold Cross Validation. Hasil penelitian menunjukkan peningkatan akurasi model terbaik berada pada Chat GPT dengan menggunakan Algoritma Naïve Bayes dan Teknik Shuffled Sampling, dari nilai akurasi awal sebesar 79.09%, setelah ditingkatkan menggunakan Feature Selection, maka nilai akurasinya meningkat menjadi 87.18%. Hal ini membuktikan peran Feature Selection, dimana yang digunakan adalah Forward Selection dalam meningkatkan akurasi ternyata memang efektif dalam meningkatkan performa akurasi model.

Kata kunci: ai; assisten; chat gpt; feature selection; gemini

INTRODUCTION

The development of AI assistants such as Gemini and ChatGPT could have assisted humans with their daily tasks. In the field of sentiment analysis, AI assistants could have been used as an automated labeling alternative to provide positive, negative, or neutral sentiment within a dataset [1]. Data labeling involves the classification of a dataset into predefined categories. This fundamental process is the precursor to any machine learning endeavor, and its significance lies in the direct correlation between label quality and model efficacy [2]. Data labeling is the cornerstone of tasks including data clustering, object recognition, and machine learning model construction [3]. Historically, data labeling has been a manual, human-driven process. In contrast, automated data labeling provides a more efficient and scalable approach. Automation enables expedited and precise data labeling, thereby expediting the development of machine learning models. [4]. Artificial intelligence (AI) assistants, powered by natural language processing (NLP) technology, are designed to interact naturally with users and perform a variety of tasks. These AI assistants can generate contextually relevant responses to user queries. Prominent examples of such AI assistants include Gemini and ChatGPT. Gemini, a large language model developed by Google, is designed to engage in natural and informative conversations. It leverages the LaMDA architecture and Google's extensive knowledge base to generate relevant and high-quality text outputs [5].

ChatGPT, a state-of-the-art AI language model, was introduced by OpenAI in November 2022. It was developed through Reinforcement Learning and trained on a dataset comprising over 150 billion parameters [5]–[8]. ChatGPT serves as a versatile text-based virtual assistant designed to engage in human-like conversations. Its applications span a wide range, including chatbot development, content generation, and machine translation, with accuracy levels varying across tasks [9], [10]. OpenAI, the organization behind ChatGPT, is an American artificial intelligence research laboratory

Feature selection, a fundamental process in machine learning, involves identifying the most pertinent and informative subset of features from a given dataset [11]–[13]. This technique enhances model accuracy by excluding irrelevant or redundant features, enabling AI models to concentrate on the most critical attributes for prediction [14].

Prior research has investigated the application of automatic data labeling using the VADER Lexicon in sentiment analysis [15]–[17]. The first study employed automatic labeling using VADER for sentiment analysis on a sentiment dataset related to the Nusantara Capital City relocation [15]. Results showed that SVM achieved an accuracy of 76.70%. The second study conducted sentiment analysis on a dataset related to the Facebook outage, also using automatic labeling with VADER [16]. Naive Bayes yielded an accuracy of 73.69% in this study. The third study focused on automatic labeling using VADER on a PLN Mobile dataset [17]. Naive Bayes achieved an accuracy of 70% in this study.

These studies reported accuracy scores within the range of 70-76%, classified as fair. To further improve model performance, feature selection techniques like Forward Selection can be implemented [11], [14]. Previous studies have demonstrated that feature selection can boost accuracy by 4-9%. A concise overview of these studies is provided in the Research Gap Table1.

This research aims to address a gap in the existing literature by exploring the use of AI assistants like Gemini and ChatGPT for text labeling, an underexplored area. To enhance model performance, feature selection using the Forward Selection method will be implemented. The study will compare the per-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 85 – 92 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3364
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

formance of Naive Bayes and K-Nearest Neighbor algorithms, employing K-Fold Cross Validation and Stratified and Shuffled Sampling techniques. The objective is to identify the model that yields the highest accuracy.

Naïve Bayes has proven to be a robust algorithm for classification tasks. It excels in handling smaller datasets, effectively identifying key features, and rapidly completing the classification process [5], [18]. Research on Brimo app sentiment analysis [19], demonstrates the success of Naïve Bayes, achieving 'Good Classification' with this method. K-Nearest Neighbors (KNN) was deemed the ideal algorithm for this task given its ability to effectively manage noisy training data, execute rapid training, offer simplicity, and handle extensive datasets with proficiency [7], [20]. The effectiveness of KNN is well-established. A study analyzing Twitter sentiment on the G20 Summit in Indonesia achieved 'Excellent Classification' results using this method [21].

METHOD

The research framework is illustrated, as shown in Image 1.

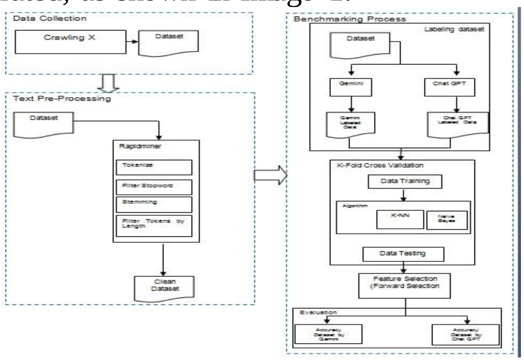


Image 1. Reseach Framework

The initial phase of this research framework is data collection. In this stage, data was gathered from the social media platform X using the keyword "Artificial Intelligence." The objective was to ascertain the sentiment of X's users regarding the presence of AI in everyday life. A dataset of 100 tweets was compiled. The rationale behind using 100 data points was to experiment with Gemini and ChatGPT's ability to label each data point and to expedite the modeling process.

The second phase involved text pre-processing, which encompassed tokenization, stop word removal, stemming, and token filtering (by length). Tokenization is the process of breaking down a sentence into individual words, known as tokens. Filter Stopword, is a process of eliminating words that are considered irrelevant and carry no significant meaning in a sentence, based on a predefined stop word list. Stemming is a technique used to reduce words to their root or base form. Filter Token By Length, is a process of restricting the inclusion of words based on a minimum and maximum character limit.

The third phase involves labeling. In this stage, the pre-processed dataset will be assigned positive, negative, or neutral labels. This phase represents a gap in previous research, as it will involve Gemini and ChatGPT in the data labeling process. Data labeling will be conducted directly on the Gemini and ChatGPT websites.

The fourth phase is modeling using Rapidminer. In this phase, the dataset labeled by Gemini and ChatGPT will be used to build models. The accuracy of each model using Naive Bayes and K-NN algorithms will be evaluated. Model validation will be performed using K-Fold Cross Validation in combination with stratified and shuffled sampling techniques. Naïve Bayes Algorithm is a classification method derived from the

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 85 – 92 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3364
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Bayes theorem, which can predict future opportunities based on opportunities that existed in the past [22]. The equation is as follows:

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

Explanation:
X = A data sample with an unknown class (label).
C = The hypothesis that X belongs to class (label) C.
P(C) = The probability of hypothesis C being true.
P(X) = The probability of observing the data sample (regardless of the class).
P(X|C) = The probability of observing the data sample given that the hypothesis C is true.

K-Nearest Neighbour Algorithm is often used for classification. The way this algorithm works is grouping data into a class that has been determined based on the closest distance or similarity to the existing data set or training data. The stages of this algorithm are as follows:

1. Determine the value of k;
2. Calculate the distance between the data that will be classified against the label data;
3. Determine the smallest value of k;
4. Classify data based on a distance metric.

Calculation of proximity using a distance matrix can use the following formula:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{21})} \quad (2)$$

Explanation:
d(x,y) = distance between data points x and y
x_i = i-th training data point
y_i = i-th testing data point
i = index of a data variable
n = dimensionality of the data

The fifth phase is Evaluation. In this phase, the accuracy of each constructed model will be assessed. Initially, the direct modeling results using the algorithms will be examined. Subsequently.

RESULT AND DISCUSSION

The result of the first stage is the crawling of a dataset from social media platform X. The dataset comprises eight attributes, of which not all are necessary for this research. Therefore, text preprocessing is required to prepare the data for use.

In the second stage, text preprocessing was performed using RapidMiner. This involved selecting relevant attributes and cleaning the text to optimize model performance. RapidMiner, a versatile data mining tool, facilitated data preparation, integration, modeling, analysis, and deployment. The specific steps involved in text preprocessing using Rapidminer are visually depicted in Image 2.

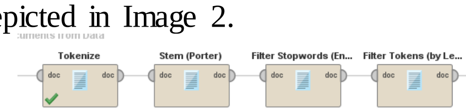
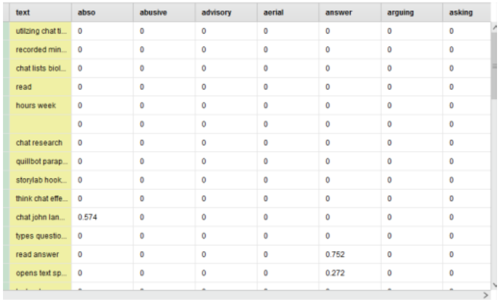


Image 2. Text Pre-Processing in Rapidminer

All necessary operators for text pre-processing are available in Rapidminer, ranging from tokenization to token filtering (by length). The results of the text pre-processing are presented visually in Image 3.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 85 – 92 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3364
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi



text	abso	abusive	advisory	aerial	answer	arguing	asking
utilizing chat ti...	0	0	0	0	0	0	0
recorded min...	0	0	0	0	0	0	0
chat lists biol...	0	0	0	0	0	0	0
read	0	0	0	0	0	0	0
hours week	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
chat research	0	0	0	0	0	0	0
quillbot parap...	0	0	0	0	0	0	0
storylab hook...	0	0	0	0	0	0	0
think chat effe...	0	0	0	0	0	0	0
chat john lan...	0.574	0	0	0	0	0	0
types questio...	0	0	0	0	0	0	0
read answer	0	0	0	0	0.752	0	0
opens text sp...	0	0	0	0	0.272	0	0

Image 3. Text Pre-Processing Result

As depicted in Figure 3, each data point will be broken down into individual words, or tokens, with each token becoming a new attribute. This text pre-processing process results in 154 new attributes.

In the third stage, each data point will be assigned a label. Both Gemini and ChatGPT will be involved in the labeling process. The labeling process using Gemini is illustrated in Image 4 and 5, while the labeling process using ChatGPT is shown in Image 6 and 7.

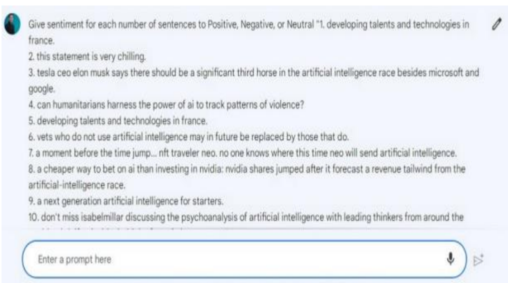


Image 4. Instruction Labeling for Gemini

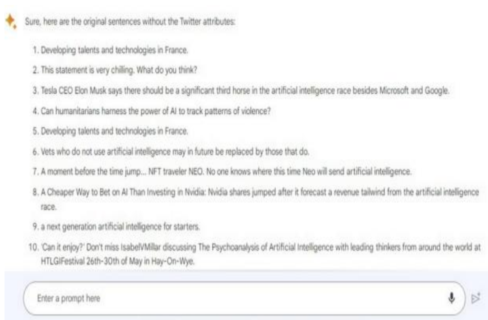


Image 5. Labeled by Gemini

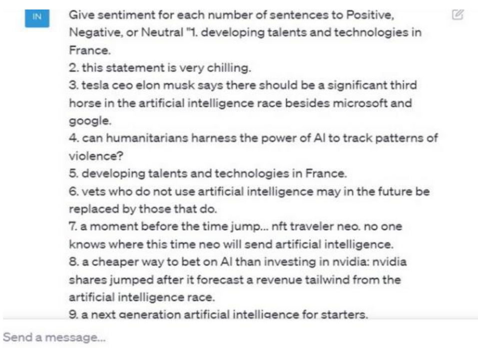


Image 6. Instruction Labeling for Chat GPT

Image 7. Labeled by Chat GPT

All labeling processes were conducted directly on the Gemini and ChatGPT websites. The labeling process was found to be quite efficient, and the results were satisfactory. Subsequently, the labeled data was passed on to the next stage for accuracy evaluation.

The results of the fourth stage present the accuracy performance of each data point labeled by both Gemini and ChatGPT. The modeling process is visually depicted in Image 8.

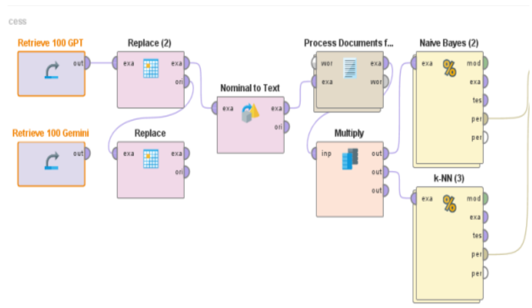


Image 8. Modelling

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 1, Desember 2024, hlm. 85 – 92 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i1.3364
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

The labeled dataset from Gemini and ChatGPT will be evaluated using the Naive Bayes and K-NN algorithms. The resulting model accuracy performance is presented in tabular form, as shown in Tabel 3.

Table 3. Performance Model Result.

Dataset	Naïve Bayes		K-NN	
	Stratified	Shuffled	Stratified	Shuffled
Gemini	75.27%	74.45%	76.27%	76.36%
Chat GPT	71.27%	79.09%	76.27%	**76.45%**

On Table 3, the best model performance is achieved with a combination of the ChatGPT dataset, the K-NN algorithm, and the shuffled sampling technique, resulting in an accuracy of 76.45%. Overall, the accuracy performance of all models falls into the fair classification category. The next step involves improving model accuracy by using feature selection (forward selection). This process is illustrated in Image 9.

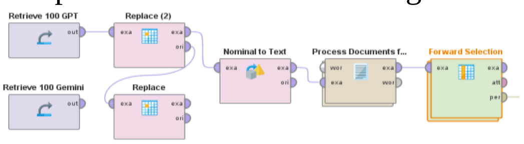


Image 9. Modelling Feature Selection

Each dataset will undergo a feature selection process (forward selection) to compare with the previous results. This comparison will determine whether there is an improvement in accuracy performance. The results are presented in tabular form, as shown in Table 4.

Table 3. Feature Selection Performance Model Result.

Dataset	Naïve Bayes		K-NN	
	Stratified	Shuffled	Stratified	Shuffled
Gemini	76.45%	80.36%	79.45%	81.45%
Chat GPT	81.36%	**87.18%**	84.36%	81.36%

It can be observed that after applying feature selection (forward selection), the accuracy of all models has improved. The best model performance is demonstrated by the combination of the ChatGPT dataset, the Naive Bayes algorithm, and the shuffled sampling technique, achieving an accuracy of 87.18%, which falls into the good classification category.

This research successfully demonstrates that AI assistants such as Gemini and ChatGPT can be used as an alternative for sentiment analysis data labeling. The results obtained are comparable to previous studies that employed VADER for labeling [15]–[17]. Furthermore, this research also proves that using Feature Selection (Forward Selection) can improve accuracy performance, aligning with findings in [11], [14]. In this study, the accuracy improvement ranged from 1-8 percent.

Naive Bayes emerged as the best algorithm compared to K-NN. Its effectiveness and speed are key factors contributing to its superior accuracy performance. This aligns with previous research [5], [18], [19].

CONCLUSION

The research results demonstrate an increase in accuracy from the initial values for Gemini using the Naive Bayes algorithm with shuffled sampling, reaching 74.45%, and for K-NN with shuffled sampling, reaching 76.36%. After enhancing accuracy using Feature Selection, the results improved to 80.36% for Naive Bayes and 81.45% for K-NN. Similarly, for ChatGPT, using the Naive Bayes algorithm with shuffled sampling yielded an initial accuracy of 79.09%, and for K-NN with shuffled sampling, 76.45%. After applying Feature Selection, the values increased to 87.18% for Naive Bayes and 81.36% for K-NN. These results confirm the effectiveness of Feature Selection in improving accuracy.

This research focuses on the role of Feature Selection in enhancing model accuracy. Future research could explore other methods to further improve performance and achieve more optimal accuracy values

BIBLIOGRAPHY

[1] N. Fitriyah, B. Warsito, and D. A. I. Maruddani, "Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM)," *J. Gaussian*, vol. 9, no. 3, pp. 376–390, 2020, doi: 10.14710/j.gauss.v9i3.28932.

[2] P. A. Telnoni, Suryatiningsih, and E. Rosely, "Pelabelan Data Deng an Latent Dirichlet Allocation dan K-Means Clustering pada Data Twitter Menggunakan Bahasa Indonesia Data Labeling using Latent Dirichlet Allocation and K-Means Clustering on Indonesian-Based Twitter," *J. Elektro Telekomun. Terap.*, vol. 7, no. 2, pp. 885–892, 2020.

[3] Aditya Quantano Surbakti, Regio lina Hayami, and Januar Al Ami en, "Analisa Tanggapan Terhadap PSBB Di Indonesia Dengan Algoritma Decision Tree Pada Twitter," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 2, no. 2, pp. 91–97, 2021, doi: 10.37859/coscit ech.v2i2.2851.

[4] A. Ahmad and W. Gata, "Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Meta verse dengan Algoritma Support Vector Machine," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 4, pp. 548–555, 2022, doi: 10.35870/jtik.v6i4.569.

[5] I. T. Julianto, D. Kurniadi, and B. B. B. Jr, "Enhancing Sentiment Analysis With Chatbots : A Comparative Study Of Text Pre-Processing," *JUTIF*, vol. 4, no. 6, pp. 1419–1430, 2023.

[6] M. Dowling and B. Lucey, "ChatGPT for (Finance) research: The Bananarama Conjecture," *Financ. Res. Lett.*, no. 103662, pp. 1–20, 2023, doi: 10.1016/j.frl. 2023.103662.

[7] I. T. Julianto, D. Kurniadi, Y. Septiana, and A. Sutedi, "Alterna tive Text Pre-Processing using Chat GPT Open AI," *Janapati*, vol. 12, no. 1, pp. 67–77, 2023, [Online]. Available: https://w jaets.com/content/artificial-intelli gence-ai-based-chatbot-study-chat gpt-google-ai-bard-and-baidu-ai.

[8] S. Singh, S. Tiwari, and P. K. Yadav, "Chat GPT : Exploring The Capabilities And Limitations Of A Large Language," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 12, pp. 111–115, 2023.

[9] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," *openai.com*, 2022. https://op enai.com/blog/chatgpt/.

[10] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare," *SSRN Electron. J.*, vol. 2, pp. 22–31, 2022, doi: 10.2139/ssrn.4308687.

[11] I. Ubaedi and Y. M. Djaksana, "Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit,"

JSiI (Jurnal Sist. Informasi), vol. 9, no. 1, pp. 17–26, 2022, doi: 10.30656/jsii.v9i1.3505.

[12] D. Kurniadi, F. Nuraeni, and S. M. Lestari, "Implementasi Algoritma Naïve Bayes Menggunakan Featu re Forward Selection dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana," *J. Sist. Cerdas*, vol. 05, no. 02, pp. 63–82, 2022.

[13] I. T. Julianto, D. Kurniadi, F. A. Fauziah, and R. Rohmanto, "Impro vement of Data Mining Models using Forward Selection and Backward Elimination with Crypto currency Datasets," *J. Appl. Intell. Syst.*, vol. 8, no. 1, pp. 100–109, 2023.

[14] Junadhi, Agustin, M. Rifqi, and M. K. Anam, "Sentiment Analysis Of Online Lectures Using K-Nearest Neighbors Based On Feature Selection," *Janapati*, vol. 11, no. 3, pp. 216–225, 2022.

[15] J. Watori, R. Aryanti, and A. Juna idi, "Penggunaan Algoritma Kla sifikasi Terhadap Analisa Sentim en Pemindahan Ibukota Dengan Pelabelan Otomatis," *J. Inform.*, vol. 7, no. 1, pp. 85–90, 2020.

[16] I. F. Ashari, "Analysis Sentiments In Facebook Down Case Using Vader And Naive Bayes Classification Method," *Multitek Indones. J. Ilm.*, vol. 16, no. 2, pp. 75–89, 2023.

[17] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pela belan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR J. Pengkaj. dan Penerapan Tek. Inform.*, vol. 15, no. 2, pp. 264–275, 2022.

[18] A. Pebdika, R. Herdiana, and D. Solihudin, "Klasifikasi Mengguna kan Metode Naive Bayes Untuk Menentukan Calon Penerima PIP," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 452–458, 2023.

[19] M. K. Insan, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggu nakan Algoritma Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 478–483, 2023.

[20] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Imple mentasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Bea siswa," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 2, pp. 118–127, 2021.

[21] H. Andriana, S. S. Hilabi, and A. Hananto, "Penerapan Metode K-Nearest Neighbor pada Sentimen Analisis Pengguna Twitter Terha dap KTT G20 di Indonesia," *JUR IKOM (Jurnal Ris. Komputer)*, vol. 10, no. 1, pp. 60–67, 2023, doi: 10.30865/jurikom.v10i1.5427.