Recursive Journal of Informatics
Vol. 3, No. 1, March 2025
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

METHOD Sentiment Analysis of Presidential Candidates in 2024: A Comparison of the Performance of Support Vector Machine and Random Forest with N-Gram Method

Muhamad Rizki Ramadhan[1], Kholiq Budiman[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. This paper conducts a sentiment analysis of presidential candidates in Indonesia's 2024 election using Twitter data. Utilizing the "Indonesia Presidential Candidate's Dataset, 2024" from Kaggle, containing 8555 Twitter entries, sentiment was categorized as positive or negative. Preprocessing techniques cleaned and normalized the data, followed by labeling with the VADER lexicon. This study contributes insights into public sentiment towards presidential candidates and the effectiveness of machine learning algorithms for political sentiment analysis.
Purpose: This study aims to analyze public sentiment towards presidential candidates in Indonesia's 2024 election using the N-Gram method. By employing Support Vector Machine and Random Forest algorithms, we compare their performance in sentiment analysis. Utilizing the "Indonesia Presidential Candidate's Dataset, 2024" from Kaggle, containing 8555 Twitter data entries, we seek to provide insights into the electorate's perceptions and preferences, contributing to a deeper understanding of the political landscape during this crucial period.
Methods/Study design/approach: The study uses Support Vector Machine (SVM) and Random Forest algorithms for sentiment analysis on a dataset of 8555 tweets about Indonesia's 2024 presidential candidates. SVM, paired with TF-IDF, and Random Forest, paired with N-Gram, are used for feature extraction. The data is labeled using the Vader lexicon.
Result/Findings: The study compared Support Vector Machine (SVM) with TF-IDF and Random Forest with N-Gram methods in analyzing public sentiment towards Indonesia's 2024 presidential candidates. Results showed Random Forest with N-Gram achieved 85% accuracy, outperforming SVM with TF-IDF at 82%.
Novelty/Originality/Value: This study provides insights into sentiment analysis applied to the 2024 Indonesian presidential election, enhancing understanding of public sentiment dynamics. Comparing SVM with TF-IDF and Random Forest with N-Gram contributes to the field, suggesting avenues for future research such as integrating contextual information or social network analysis for deeper insights into political opinion trends.

Keywords: Sentiment Analysis, Presidential Election, Twitter Data, Support Vector Machine, Random Forest, N-Gram Method.
Received June 28, 2024 / **Revised** July 07, 2024 / **Accepted** March 27, 2025

INTRODUCTION

The presidential election is a crucial moment where citizens cast their votes to determine the political direction and policies of the country. The impact of technological advancements on presidential elections is significant, affecting political campaigns, communication between candidates and voters, and the way public opinion is shaped. The selection of a president is a cornerstone of democratic processes worldwide, symbolizing citizens direct participation in governance[1]. Social media has become a primary platform for individuals to express themselves and share opinions on various matters, including presidential elections. Platforms such as Twitter, Facebook, and Instagram provide spaces for discussion, information sharing, and opinion expression about presidential candidates. Social media allows the public to actively engage in shaping public opinion and influencing political narratives. With the advent of social media platforms like Twitter, public opinions on presidential candidates have become increasingly accessible and influential[2]. Analyzing these opinions provides valuable insights into voter sentiments and preferences, contributing to a deeper understanding of political landscapes[3].

[1]*Corresponding author.
Email addresses: rizkirmdn@students.unnes.ac.id (Ramadhan)
DOI: 10.15294/rji.v3i1.8385

Twitter has become a platform for public freedom of expression. The sentiments they express can be directed at various objects, including public views on the presidential election[4]. Public sentiments written on Twitter can be analyzed in more detail using sentiment analysis techniques[5]. Sentiment analysis is the process of determining sentiments and categorizing the polarity of text in documents or sentences so that it can be classified as positive, negative, or neutral sentiment[6]. The sentiment analysis process requires a lexicon-based method to identify whether the polarity of an opinion tends to be positive or negative in a sentiment based on a dictionary, such as the Vader sentiment library[7]. Artificial intelligence has another branch, text mining, which aims to extract useful information from unstructured textual data through the identification and implicit patterns[8].

Support Vector Machine (SVM) is one of the classification methods using supervised learning that can predict classes based on patterns from the data training process[9]. The classification technique involves processing data and classifying it into two classes: positive and negative[10]. Random Forest is a classification method based on the aggregation of decision trees[11]. This method is renowned for its accuracy and its ability to handle small samples and high-dimensional feature spaces. Random Forest is considered to work effectively on large datasets and has higher accuracy compared to other classification algorithms[12].

METHODS

The method of analyzing public opinion is carried out in several stages. The first is inputting the presidential candidate dataset, followed by the preprocessing stage, which includes data cleaning, case folding, tokenization, normalization, stop word removal, and stemming. The next step is the labeling process using the Vader lexicon, followed by feature representation using unigram, and finally, the classification process using support vector machine and random forest algorithms. These steps are illustrated in Figure 1.

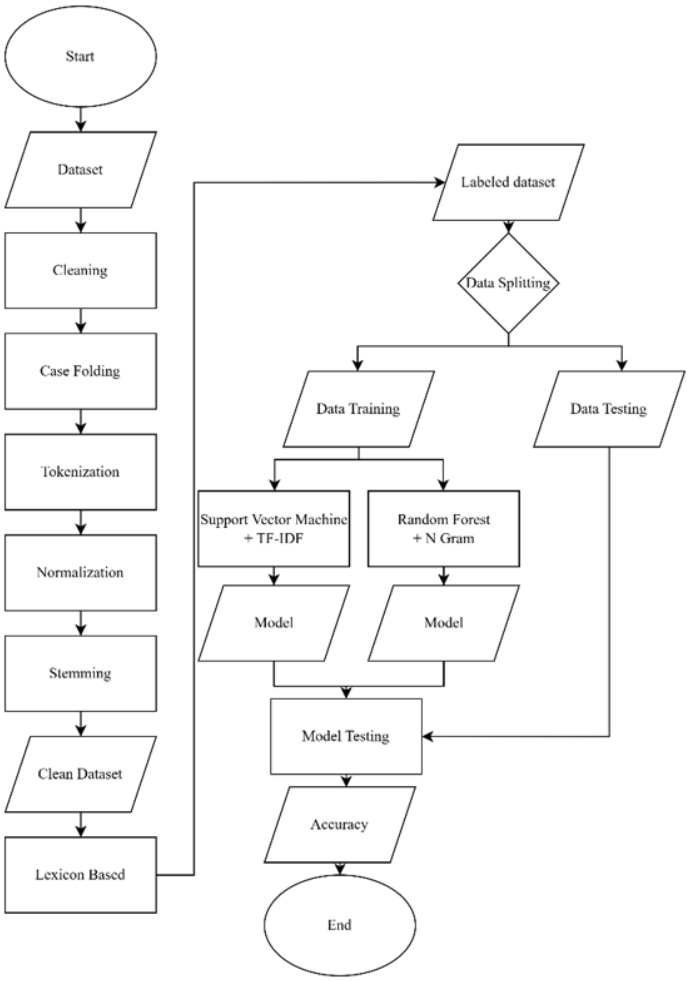

Figure 1. Research Method Flowchart

Data Collection
The data was obtained from the Indonesia Presidential Candidate's Dataset, 2024, downloaded from the Kaggle platform. The attributes of the dataset and their descriptions can be seen in Table 1.

Table 1. Dataset Input

created_at	text
16/04/2023 17:00	Gerindra Party politician Sandiaga Uno responded to the question about being paired again with former Jakarta Governor Anies Baswedan in the upcoming presidential election.
16/04/2023 16:14	Go ahead, Mr. Anies, we will support you until you become president.
16/04/2023 14:03	May Allah SWT protect the nation and the Republic of Indonesia from traitors to the people's mandate. O Allah, the Most Gracious and Most Merciful, elevate the status of Brother Anies Rasyid Baswedan to become the President of the Republic of Indonesia in the coming years and beyond. Aamiin, Ya Rabbal 'Alamin.

Preprocessing
The preprocessing stages consist of data cleaning, case folding, tokenization, normalization, stop word removal, and stemming. These stages are intended to clean the tweet data that contains numbers, characters, or meaningless words, such as symbols, punctuation marks, retweets, mentions, hashtags, and others, in order to assist the algorithm in classifying the text data[13].

Data Labelling
In the data labeling stage, tweets that have undergone preprocessing are then labeled using Valence Aware Dictionary and Sentiment Reasoner (Vader). Vader will categorize the data into negative, positive, or neutral based on their polarity by looking at the processed polarity score. If the polarity score is greater than or equal to 0.05, the text will be labeled 1, which means positive. If the polarity score is less than or equal to -0.05, the text will be labeled -1, which means negative. For neutral, with a label of 0, the polarity score ranges between -0.05 and 0.05[14].

Data Splitting
To ensure a robust evaluation of the model's performance, this study employs the K-Fold Cross-Validation method. This technique divides the dataset into multiple subsets, where each subset takes turns being the validation set while the remaining data is used for training. This approach helps in minimizing overfitting and provides a more generalized performance assessment of the model. Figure 2 illustrates the K-Fold Validation process, where the dataset is split into 10 equal parts. In each iteration, one part is designated as the validation set (orange) while the rest are used for training (green).

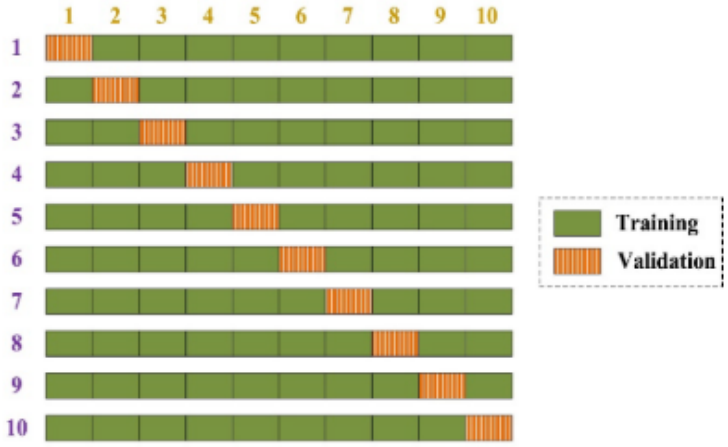

Figure 2. K-Fold Validation Illustration

The method used for data splitting in this document is K-Fold Validation. This method works by dividing the dataset into 'K' number of folds or subsets. The process involves the following steps[15]:
1. Partitioning: The entire dataset is randomly partitioned into 'K' equal-sized folds.
2. Validation: In each iteration, one fold is used as the validation set (or test set), and the remaining 'K-1' folds are used as the training set.

3. Training and Testing: The model is trained on the ‘K-1’ folds and then tested on the validation fold to evaluate its performance.
4. Iteration: This process is repeated ‘K’ times, with each of the ‘K’ folds used exactly once as the validation set.
5. Averaging: The ‘K’ results from the folds can then be averaged (or otherwise combined) to produce a single estimation.

Model Evaluation

After splitting the dataset into 'K' folds and obtaining the classification model from training, the model is tested with the testing data. The test results are then evaluated to determine the error rate and accuracy.

The following is the formula for K-Fold Validation in calculating accuracy results, as shown in the formula below:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

where :

TP : Number of data points that are actually Positive and correctly predicted as Positive.
FP : Number of data points that are actually Negative but predicted as Positive.
TN : Number of data points that are actually Negative and correctly predicted as Negative.
FN : Number of data points that are actually Positive but predicted as Negative.

The following is the formula for K-Fold Validation in calculating the overall accuracy results, as shown in the formula below:

$$Measure\ Accuracy = \frac{(K_1\ Accuracy+K_2\ Accuracy+\cdots+K_n\ Accuracy)}{K} \quad (2)$$

where :

Measure Accuracy : The average of all accuracies.
$K_n\ Accuracy$: The accuracy result from each iteration.
K : The number of folds used.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification, regression, and outlier detection tasks. The task of the Support Vector Machine algorithm is to divide these two groups as effectively as possible by determining the best hyperplane. This involves finding a boundary line that can separate the two groups with the maximum distance between the outermost points in each group and the boundary line. Figure 3 below illustrates how the Support Vector Machine (SVM) algorithm separates two different classes using a hyperplane.

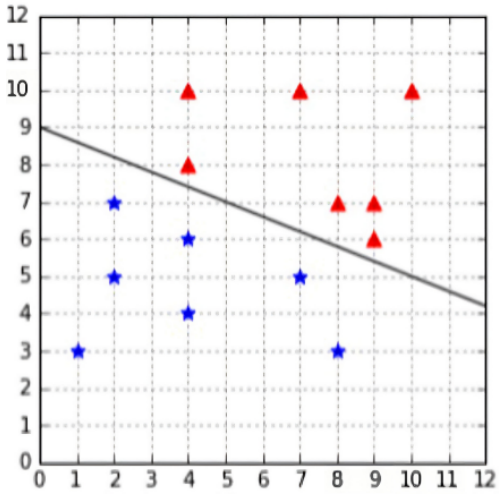

Figure 3. Support Vector Machine Illustration

To separate two classifications using the optimal hyperplane, the following equation can be used:

$$f(x) = w \bullet x + b \quad (3)$$

where:

f(x) : Decision function to predict the class of input data x.
x : Feature vector.
w : Weight vector.
b : Bias (constant).
• : Dot product operation.

In the context of this research, SVM is utilized for its effectiveness in text classification, particularly for sentiment analysis[13]. The SVM flowchart is shown in Figure 2.

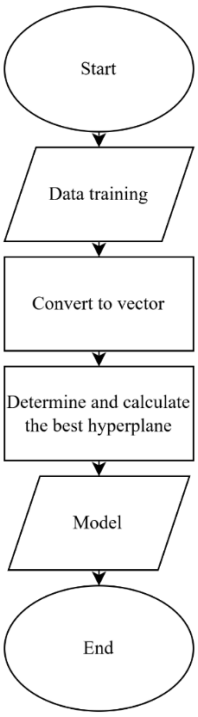

Figure 4. Flowchart Support Vector Machine

Random Forest

Random Forest is a machine learning algorithm used for classification, regression, and other tasks. It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [13].

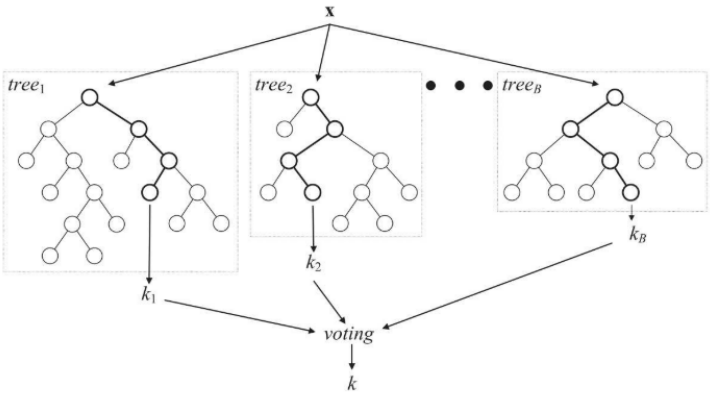

Figure 5. Random Forest Illustration

The Random Forest algorithm creates multiple decision trees, known as a forest. When classifying new data, each decision tree produces a predicted category, and the forest selects the most frequent category. The more decision trees in the Random Forest algorithm, the higher the accuracy.

Here are the formulas used in Random Forest:

$$Entropy\ (Y) = -\Sigma_i p(Y) \log_2 p(Y) \tag{4}$$

where:
Y : The dataset under evaluation.
$p(Y)$: The proportion of each class in the dataset Y.
$\log_2 p(Y)$: $\log_2$ from $p(Y)$.

$$Information\ Gain\ (Y,a) = Entropy(Y) - \Sigma_{v \varepsilon Values(a)} \frac{|Y_V|}{|Y_a|} Entropy(Y_V) \tag{5}$$

where :
$Entropy\ (Y)$: The entropy value.
Values(a) : Values in set a..
$|Y_V|$: Number of samples in subset Y_V.
$|Y_a|$: Total number of samples in dataset *Y*.
$Entropy(Y_V)$: Entropy of each subset Y_V after separation.

In the context of this research, Random Forest is used to analyze public sentiment towards presidential candidates in Indonesia for the year 2024. It is chosen for its ability to handle large datasets and its high accuracy in classification tasks. The research aims to compare the performance of Random Forest with Support Vector Machine (SVM) using the N-Gram method[12]. The Random Forest flowchart is shown in Figure 6.

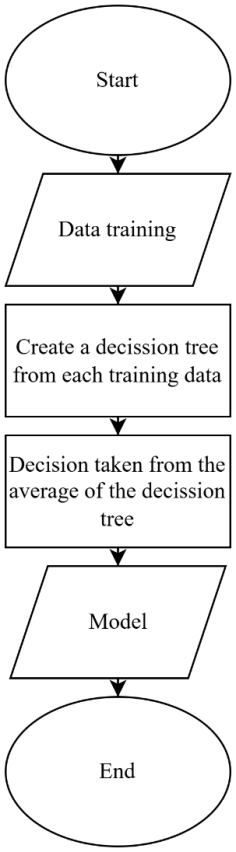

Figure 6. Flowchart Random Forest

RESULT AND DISCUSSION

In this study, the dataset used is data regarding presidential candidates in the 2024 presidential election, downloaded from the Kaggle platform under the title 'Indonesia Presidential Candidate's Dataset, 2024,' consisting of 10,000 data points. After inputting the dataset, the data will undergo preprocessing. This stage functions to clean the data, which still contains numbers, characters, or meaningless words such as symbols, punctuation marks, retweets, mentions, hashtags, and others, as well as to remove duplicate data. The quantity of the database is reduced from the initial 10,000 raw data points to 8,555 clean data points.

In the data labeling stage, tweets that have undergone preprocessing are then labeled using Vader. Vader will categorize the data into negative, positive, or neutral based on their polarity by looking at the processed polarity score. If the polarity score is greater than or equal to 0.05, the text will be labeled 1, which means positive. If the polarity score is less than or equal to -0.05, the text will be labeled -1, which means negative. For neutral, with a label of 0, the polarity score ranges between -0.05 and 0.05[14]. The accumulated sentiment results based on the three categories of positive, neutral, and negative can be seen in the Figure 7.

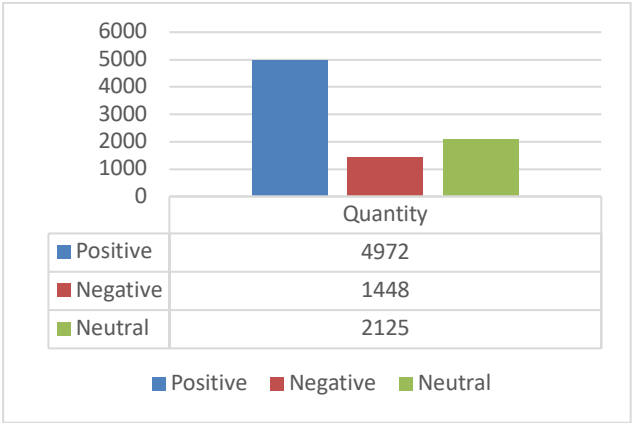

Figure 7. Sentiment accumulation result

In this research, a comparison of the results of two text classification algorithm models was carried out where the higher the accuracy of an algorithm model, the better the algorithm used. The research was conducted using k-fold cross validation, namely by dividing the data into two randomly into training data and testing data by weighting words using TF-IDF. The comparison is made with the ratio on the testing data, which is 10% and iterations are carried out 20 times with a total of 8555 data. After preprocessing and labeling the sentiment analysis on the Indonesia Presidential Candidate's Dataset on Twitter, then classification using a support vector machine and Random forest, the comparison results are obtained as shown in Figure 8.

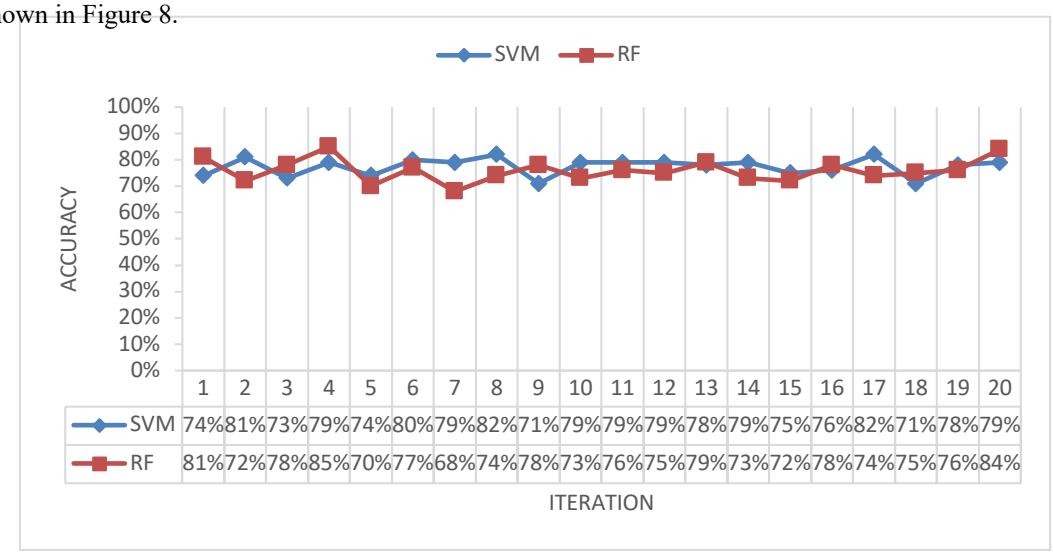

Figure 8. Classification accuracy results

The accuracy results show that the support vector machine get the highest accuracy in the 8th and 17th iteration with results 82% and random forest algorithms get the highest accuracy in the 4th iteration with results 85%.

From the test results, overall random forest is better overall than the support vector machine algorithm in classifying data about 8555 Indonesia Presidential Candidate's Dataset. With the level of accuracy generated by the support vector machine algorithm and random forest algorithm, it is expected to be able to analyze data sentiment on the Indonesia Presidential Candidate's Dataset well.

CONCLUSION

Based on the results of research that has been carried out on the independent campus policy on Twitter, it can be ascertained that the way to analyze the sentiment about the independent campus public policy on Twitter with a vector engine support algorithm and naïve Bayes classifier is carried out in several stages. The first stage is crawling data on Twitter social media, then preprocessing the data with stages of cleaning, case folding, tokenization, normalization, stop word removal, and stemming. The next step is to perform the data labeling process using Vader, then word vectorization of the data is carried out with TF-IDF. The last step is to classify using a support vector machine algorithm and a naïve Bayes classifier.

The study compared the effectiveness of Support Vector Machine (SVM) with TF-IDF and Random Forest with N-Gram in analyzing public sentiment towards presidential candidates in the 2024 election. Through a multi-step process with 20 folds and a 10% testing data ratio, revealed that Random Forest with N-Gram achieved the highest accuracy of 85%, surpassing SVM with TF-IDF which attained 82%. This suggests that Random Forest with N-Gram is superior in sentiment analysis of presidential candidates' public opinion for the 2024 election. With 4972 positive sentiments, 1448 negative encompassing dataset input, preprocessing, labeling with Vader lexicon, feature representation using unigram, and classification, sentiment analysis was conducted. The evaluation, employing K-fold validation sentiments, and 2125 neutral sentiments identified, the majority of the public opinion towards the candidates was positive.

REFERENCES

[1] G. Liva, C. Codagnone, G. Misuraca, V. Gineikyte, and E. Barcevicius, *Exploring digital government transformation*, no. January. 2020.

[2] A. R. Isnain, A. I. Sakti, D. Alita, and N. S. Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm," *J. Data Min. dan Sist. Inf.*, vol. 2, no. 1, p. 31, 2021, doi: 10.33365/jdmsi.v2i1.1021.

[3] A. H. Muzahid, J. Sjaiful, W. Gunawan, and I. I. Muhammad, "Peran Media Sosial Dalam Komunikasi Politik," *J. Indones. Sos. Teknol.*, vol. 2, no. 1, pp. 104–114, 2021, doi: 10.36418/jist.v2i1.61.

[4] S. N. J. Fitriyyah, N. Safriadi, and E. E. Pratama, "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 279, 2019, doi: 10.26418/jp.v5i3.34368.

[5] R. Vindua and A. U. Zailani, "Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python," *JURIKOM (Jurnal Ris. Komputer)*, vol. 10, no. 2, p. 479, 2023, doi: 10.30865/jurikom.v10i2.5945.

[6] W. A. Prabowo and C. Wiguna, "Sistem Informasi UMKM Bengkel Berbasis Web Menggunakan Metode SCRUM," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 149, 2021, doi: 10.30865/mib.v5i1.2604.

[7] P. A. Sumitro, Rasiban, D. I. Mulyana, and W. Saputro, "Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based," *J-ICOM - J. Inform. dan Teknol. Komput.*, vol. 2, no. 2, pp. 50–56, 2021, doi: 10.33059/j-icom.v2i2.4009.

[8] A. Segnini, J. Joyce, and T. Motchoffo, "Random Forests and Text Mining," 2021.

[9] B. W. Sari and F. F. Haranto, "Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom Dan Biznet," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, 2019, doi: 10.33480/pilar.v15i2.699.

[10] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[11] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment

analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no. 1, p. 45, 2021, doi: 10.26555/jifo.v15i1.a20111.
[12] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Comput. Sci.*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.
[13] S. Sudianto, P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "Comparison Of Random Forest And Support Vector Machine Methods On Twitter Sentiment Analysis (Case Study : Internet Selebgram Rachel Vennya Escape From Quarantine) Perbandingan Metode Random Forest Dan Support Vector Machine Pada Analisis Sentimen Twitt," *Jutif*, vol. 3, no. 1, pp. 141–145, 2022.
[14] R. Bose, P. S. Aithal, and S. Roy, "Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries," *Int. J. Manag. Technol. Soc. Sci.*, vol. 6, no. 1, pp. 110–127, 2021, doi: 10.47992/ijmts.2581.6012.0132.
[15] M. Kuhn and K. Johnson, *Applied Predictive Modeling with Applications in R*, vol. 26. 2013.