Vol. 08, No. 02, July 2024, pp. 66-76

INTERNATIONAL JOURNAL OF INNOVATION IN ENTERPRISE SYSTEM

Journal homepage: https://ijies.sie.telkomuniversity.ac.id
e-ISSN: 2580-3050, DOI: https://doi.org/10.25124/ijies.v8i02.236

Analyzing Twitter Sentiments on Booster Vaccination with Support Vector Machine (SVM) Method

Rahmat Fauzi[1]*, Faqih Hamami[2], Fakhri Hassan Maulana[3], Brillian Adhiyaksa Kuswandi[4], Muhammad Ayyub Ramli[5]

[1] School of Industrial and System Engineering
Telkom University, Bandung, Indonesia
rahmatfauzi@telkomuniversity.ac.id

[2] School of Industrial and System Engineering
Telkom University, Bandung, Indonesia
faqihhamami@telkomuniversity.ac.id

[3] School of Industrial and System Engineering
Telkom University, Bandung, Indonesia
fakhrihassan@student.telkomuniversity.ac.id

[4] School of Industrial and System Engineering
Telkom University, Bandung, Indonesia
brillianadhiyaksa@student.telkomuniversity.ac.id

[5] School of Industrial and System Engineering
Telkom University, Bandung, Indonesia
mhyubr@student.telkomuniversity.ac.id

*rahmatfauzi@telkomuniversity.ac.id

ARTICLE INFO

Article history:
Received 29 November 2024
Accepted 30 November 2024
Published 18 December 2024

Keywords:
Sentiment Analysis; Text Pre-Processing; Support Vector Machine; TF-IDF; Twitter

ABSTRACT IN ENGLISH

The Indonesian government has implemented various measures to prevent the spread of the COVID-19 virus, one of which is through a vaccination program with two doses (the first dose and the second dose). However, new variants of the virus have emerged, reducing the effectiveness of the initial vaccinations. To address this, the government introduced a booster vaccination program aimed at enhancing immunity by up to 80%. The government's plan for booster vaccination has received both positive and negative opinions from the public through various media platforms, including Twitter. This study analyzes public opinions on the booster vaccination plan into three classes: positive, negative, and neutral. SVM is a classification method in machine learning categorized as supervised learning, which involves finding an optimal line (hyperplane) as a separator for two different data classes. The stages of this research include data collection, data cleaning, data transformation, and data classification using the Support Vector Machine (SVM) method. The results of this study indicate that the accuracy of the SVM model reaches 80.42%.

1. INTRODUCTION

Since the beginning of March 2020, Indonesia has been dealing with the Covid-19 outbreak [1]. This outbreak quickly turned into a public health crisis, significantly impacting the country's economy and putting a strain on the healthcare systems. The severity of the situation prompted the government to declare COVID-19 a national disaster by issuing Presidential Decree Number 12 of 2020, which gave the COVID-19 task force the authority to manage the country's response. One of the measures implemented by the task force was the introduction of Large-Scale Social Restrictions (PSBB) through Government Regulation Number 21 of 2020, which was a crucial strategy in controlling the spread of the virus. However, while these measures helped slow down transmission, they also had significant socio-economic consequences, emphasizing the need for more sustainable solutions such as vaccination.

Vaccination became a vital intervention in preventing the spread of COVID-19, with research demonstrating that vaccines can trigger a strong immune response, lowering the risk of severe illness and death. By the end of 2021, the Indonesian government had effectively administered the initial and booster doses of the COVID-19 vaccine to a large segment of the population. However, the emergence of new variants, especially the Omicron variant, presented new challenges. Studies indicated that the original vaccines were less effective against Omicron, particularly for individuals with compromised immune systems. This decreased effectiveness highlighted the need for booster doses, which have been proven in multiple studies to restore vaccine efficacy by enhancing the immune response by up to 80%.

Acknowledging the ongoing risk posed by COVID-19 mutations and the potential for future outbreaks, the Indonesian government has scheduled a nationwide booster vaccination campaign targeting individuals who have already completed their second dose [10]. This campaign is not only crucial for maintaining public health but also for ensuring that the population remains protected against evolving variants of the virus.

The government's plan for booster vaccination has garnered both positive and negative responses from the public across various media platforms [2]. One of the commonly used platforms by the Indonesian public to express their opinions is Twitter, which serves as a valuable indicator for research influence [3][4]. Retrieving information from a collection of tweets is challenging to accomplish manually due to the high volume of daily tweets covering various topics, making it essential to have a rapid and precise data analysis model. Utilizing data analysis techniques like sentiment analysis can aid in uncovering concealed insights from a series of tweets [5]. Sentiment analysis, a component of text mining, is aimed at categorizing textual content into opinions, thereby producing sentiment data that could be positive or negative.
Wardhani [6] conducted an analysis of public sentiment regarding the COVID-19 vaccine in Jakarta, focusing on how Twitter users express their views on vaccination.

Twitter was selected as the platform for sentiment analysis in this study for several compelling reasons. Firstly, Twitter produces a substantial volume of real-time data, with around 600 million tweets being shared daily, making it an excellent and immediate source of data for evaluating public sentiment on dynamic topics like COVID-19 vaccination. Moreover, Twitter's data is predominantly public and easily accessible through APIs, which simplifies the collection and analysis of large datasets, providing a significant advantage over other social media platforms [7]. Additionally, Twitter boasts a diverse and engaged user base spanning across different demographics, offering a comprehensive insight into public opinion and sentiment on crucial matters. Furthermore, Twitter serves as a prominent platform for public discourse, where users frequently engage in discussions around social, political, and health-related issues, thereby making it an invaluable resource for understanding public responses and the broader societal implications of health policies, such as vaccination campaigns. These considerations collectively validate the choice of Twitter as the preferred platform for conducting sentiment analysis in this research.

In numerous studies, researchers have delved into text data classification to extract valuable insights. For example, a research study by Hayati et al [8], centered on sentiment analysis to evaluate public perceptions of the Covid-19 vaccine using the Support Vector Machine (SVM) technique. The primary aim was to determine the most effective model by comparing the outcomes of unigram and bigram approaches. The findings indicated that the SVM model using unigram and bigram yielded a comparable accuracy rate, with a variance of around 0.6-0.7, achieving an overall accuracy of 84%. Another investigation by Raharjo [9] examined public feedback about the COVID-19 vaccine, employing both Naïve Bayes and SVM methods, with the SVM model surpassing Naïve Bayes with an accuracy of 87% in contrast to 81%.

In this study, the decision to utilize the Support Vector Machine (SVM) method was driven by its proficiency in managing high-dimensional data and its resilience in various text classification tasks, particularly sentiment analysis. SVM is well-suited for binary classification scenarios, where the objective is to segregate data points into two categories, rendering it a fitting choice for sentiment analysis tasks that typically categorize data into positive and negative sentiments [7]. The choice to employ SVM is further bolstered by its capacity to identify the optimal hyperplane that maximizes the margin between data points of distinct classes, thereby mitigating the risk of misclassification and enhancing the model's

*rahmatfauzi@telkomuniversity.ac.id

generalization capabilities. Moreover, the SVM method has been extensively validated in prior research, demonstrating superior performance in sentiment analysis compared to other machine learning algorithms, particularly in handling imbalanced datasets and high-dimensional feature spaces. These attributes establish SVM as a dependable and effective method for accurately categorizing public sentiments regarding the COVID-19 vaccine.

Based on the background, this research aims to examine the response and opinions of the Indonesian public towards booster vaccines using data sourced from the social media platform, Twitter. This study will conduct sentiment analysis by classifying the public's responses into positive, negative, and neutral sentiments, and categorizing public opinions towards booster vaccines using the Support Vector Machine (SVM) method.

2. METHOD

The flow of research methods is as follows:

1. Problem Identification
   During this initial phase, the research issue is recognized, requiring the development of a framework for application and data collection. To gather insights from a range of sources, such as books, articles, journals, and pertinent references, a thorough review of the literature is carried out. This review assists in shaping the research by pinpointing existing deficiencies and backing the research goals [10].
2. Problem Analysis
   This stage is to comprehensively analyze the identified issues and system requirements. The goal is to gain a deep understanding of the problems and identify the best possible solutions to resolve them. This analysis forms the basis for the later stages of the research[11].
3. Data Collection
   The data collection process is performed using the Twitter API, specifically through the tweepy library in Python3 (https://github.com/tweepy/tweepy). Data is collected using the keyword "booster vaccine" and is restricted to tweets from Indonesia. The collected data is then stored in a .csv file for further processing. Twitter was selected due to its real-time data availability and the rich, diverse opinions expressed on the platform[12].
4. Text Preprocessing
   At this stage, text pre-processing is the process of summarizing data into text data that is ready for further processing. The pre-processing stages used are:
   a. Removing Character: Elimination of unnecessary characters such as URLs, mentions, and punctuation.
   b. Case Folding: Conversion of all text to lowercase to maintain uniformity.
   c. Tokenizing: Splitting text into individual words or tokens
   d. Stopword Removal: Removing common words that do not contribute to the statement, such as “and”, “the” etc.
   e. Stemming: Reducing words to their base or root form.
5. Data Labelling
   The process of data labeling involves a semi-manual method supported by the SentiStrength library, which is designed for assessing sentiment in social media content. Validation of the labeled data is undertaken to guarantee its accuracy before any additional analysis[13].
6. Data Visualization
   At this stage, data visualization also serves to communicate research findings effectively and persuasively, survey results, or data analysis to others.
7. TF-IDF Word Weighting
   At this stage, TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method for measuring the importance of words in a document or a collection of documents [14].
8. Support Vector Machine (SVM) Classifier
   The sentiment analysis utilizes the Support Vector Machine (SVM) classifier. SVM is selected due to its strong and efficient handling of high-dimensional data, which makes it especially suitable for text classification tasks. SVM's capacity to maximize the margin between classes decreases the chances of misclassification, resulting in improved accuracy when predicting sentiments[7].
9. Model Generation
   At this stage, the tweets that have undergone word weighting, specifically TF-IDF and classification using the Support Vector Machine (SVM) classifier, will generate both a classification model and a weighting model. These models are then saved to be used for the weighting and classification of new data that does not have a sentiment label. This allows the data to be directly classified into positive, neutral, or negative sentiment categories [7].
10. Testing Scenario and Result
    In this phase, the implementation of the program using the Python programming language covers various stages, including data collection and the creation of a confusion matrix. Subsequently, the classification model is tested by evaluating the accuracy of the confusion matrix, which allows us to assess the model's performance and its ability to classify data accurately[7].

2.1. Problem Identification

In the early stages, researchers identify issues that arise in the community related to public opinion or response to booster vaccines. Next, collect literature studies related to sentiment analysis on public responses related to booster vaccination to determine which algorithm to use in the case study to solve the problem. Then the purpose of this research will be determined. Finally, determine the boundaries of the problem so that the topic discussed does not expand.

2.2. Data Collection

Data collection was conducted by retrieving Indonesian-language tweets containing various keywords related to Booster Vaccines during the COVID-19 pandemic in Indonesia on Twitter. The data retrieval process is carried out with the Twitter API using Python3 [15]. The data collection process was carried out at three different times. First, on 12 February 2022 when the booster vaccine policy was announced, then on 6 April 2022 when the booster vaccine policy was implemented, and finally on 7 May 2022, after the policy was implemented.

2.3. Data Pre-processing

During the pre-processing phase, data is obtained from the social network Twitter. Data is acquired in the form of Twitter users' comments or thoughts on vaccination boosters. The collected data was then categorized as good, negative, or neutral. The next step in data pre-processing is data filtering, which selects the relevant Twitter data for processing. As to his steps:

a. Cleaning is Remove URLs, mentions with usernames, punctuation, hashtags, and read marks To improve classification,
b. Case folding is the process of transforming the character of the data into the same shape, large and small; tokenization is the division of text data into sections known as tokens.
c. Tokenization may take the form of words, numbers, symbols, or other meaningful elements.
d. Stopwords removal is the process of filtering crucial words from the token result by removing words that are frequently used but have no substantial impact on the phrase. Words that can be eliminated include (or), and (and), and others.
e. Stemming is the process of identifying basic words by eliminating replacements.

2.4. Data Labeling

Data labeling was conducted by semi-manual labeling with the help of the masdevid/sentistrength_id repository on GitHub into three types of labels, namely positive (1), negative (-1), and neutral (0). The Sentistrength sentiment algorithm aims to measure the strength of negative sentiment in a text, with the assumption that positive and negative sentiments can coexist within the text [16]. However, the final score may change if additional rules are met during feature extraction. Then, the final sentiment decision is based on these rules [16]:

a. If positive value > negative value then positive sentiment.
b. If positive value < negative value then negative sentiment.
c. If positive value = negative value then neutral sentiment.

2.5. Data Visualization

The data visualization process is carried out to observe the distribution of the numbers of data for positive, negative, and neutral sentiments, which will be presented in the form of a bar chart, as well as the distribution of words for each class, which will be presented in the form of a word count.

2.6. Data Preprocessing

In the data preprocessing stage, the researcher collected data from the Twitter social network. The data was collected in the form of comments or opinions from Twitter users regarding booster vaccination. The collected data was then labeled based on their respective categories, whether positive, negative, or neutral. The next step in data preprocessing is to perform data filtering to select relevant tweets for processing [17]. The steps involved are as follows:

a. Cleaning, which is the process of removing attributes that are not related to words such as URLs, hashtags, mentions, retweets, punctuation characters, removing excess spaces and emoticons, and case folding.
b. Case folding is the process of converting all characters in the data to a consistent form, either uppercase or lowercase.
c. Tokenization, which involves dividing the text data into smaller units called tokens. Tokens can be words, numbers, symbols, or other meaningful elements.
d. Stopword removal is the process of filtering important words from the tokenized results by removing commonly used words that do not have a significant impact on the sentence. Some examples of words that can be removed are "atau", "dan" and others.
e. Stemming is the process of finding the base form of a word by removing word affixes. “

2.7. TF-IDF Word Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) is a technique used to assign weights to words in a document based on their frequency within the document and across all documents [14]. TF-IDF value is obtained by multiplying the TF result with the IDF calculation result [14], as shown in equation 1.

$$tfidf = tf(t,d) \times idf(t) \tag{1}$$

Where *tf(t,d)* represents the number of occurrences of the word *t* in a document *d*. The next step is to determine the number of documents that contain the specific word and calculate its inverse, known as IDF [14], as depicted in equation 2.

$$idf(t) = 1 + \left(\log \frac{D}{df(t)}\right) \tag{2}$$

Where *D* represents the total number of documents, and *df(t)* represents the frequency of the word *t* occurring in all documents.

2.8. Support Vector Machine Classifier

Support Vector Machine (SVM) is a supervised technique in machine learning that can be used to solve classification or regression problems. SVM is the most used technique for classification processes in the context of machine learning because the final results obtained are relatively better compared to other methods [18]. SVM is a classification technique that aims to find the hyperplane with the largest margin. The best hyperplane is the one that maximizes the margin value. Margin refers to the distance between the hyperplane and the support vectors (points closest to the hyperplane) [9].

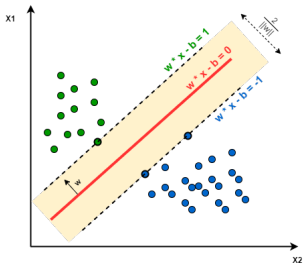


Figure 1 – Formation of Hyperplane in SVM

Figure 1 illustrates the structure of SVM, which consists of two different classes, namely class -1 and class +1. The data of both classes are separated by a hyperplane. The data points that are closest to the hyperplane are called support vectors, with the margin as the distance value between the hyperplane and the support vectors [17]. The value of the hyperplane must be determined first to maximize the margin value in equation 3 [19]:

$$margin = \frac{1}{2}\|w\|^2 \tag{3}$$

Under the condition in equation 4:

$$w.x_i + b = 0 \tag{4}$$

The formula to calculate the maximum margin in equations 5 and 6 is as follows [20]:

$$w.x_i - b = 1 \tag{5}$$

$$w.x_i - b = -1 \tag{6}$$

The value of *w* represents the hyperplane needed to obtain the perpendicular line between the hyperplane line and the support vector points, xi represents the *i* attribute in the data, and *b* is the bias. The hyperplane class is divided into two, namely the positive class (+1) and the negative class (-1). Then, the data is predicted using the equations as 7 and 8 below:

$$w.x_i + b \leq -1 \tag{7}$$

$$w.x_i + b \geq 1 \tag{8}$$

The commonly used types of kernels include polynomial kernel, Gaussian kernel, linear kernel, and Radial Basis Function (RBF) kernel.

*rahmatfauzi@telkomuniversity.ac.id

2.9. Confusion Matrix

The true target values from the test dataset can be compared to the predicted answers, allowing for the calculation of the model's accuracy percentage, which is an effective way to evaluate each model overall [21]. Model evaluation is utilized to determine the performance value of the model, following the classification results using parameters based on the confusion matrix table. Table 1 below represents the confusion matrix for three-class classification.

Table 1 – Confusion Matrix Three Classes

		Predict		
		Negative	Neutral	Positive
Actual	Negative	**True Negative**	False Negative	False Negative
	Neutral	False Neutral	**True Neutral**	False Neutral
	Positive	False Positive	False Positive	**True Positive**

Note:
True Positive: The data is correctly predicted positive.
True Neutral: The data is correctly predicted neutral.
True Negative: The data is correctly predicted negative.
False Positive: The data is incorrectly predicted positive.
False Neutral: The data is incorrectly predicted neutral.
False Negative: The data is correctly predicted negative.

The confusion matrix can be calculated with the following equation [21]:

a. Accuracy is a statistical metric that assesses the model's capability to classify correctly.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (9)$$

b. Precision indicates the proportion of correctly labeled positive data compared to the total data predicted as positive.

$$precision = \frac{TP}{TP+FP} \quad (10)$$

c. Recall denotes the proportion of positively classified data related to the actual positive data.

$$recall = \frac{TP}{TP+FN} \quad (11)$$

d. F-measure is a metric that compares the average precision and recall.

$$f_{measure} = \frac{2*precision*recall}{precison+recall} \quad (12)$$

3. RESULT AND DISCUSSION

3.1. Data Collection

The data collection process was conducted at three different times. First, on February 12, 2022, when the booster vaccine policy was announced, then on April 06, 2022, when the booster vaccine policy was implemented, and finally, on May 07, 2022, after the policy was implemented. The total collected data is 1890 records. Figure 2 illustrates the processing of importing data after crawling from Twitter and labeling data.

*rahmatfauzi@telkomuniversity.ac.id

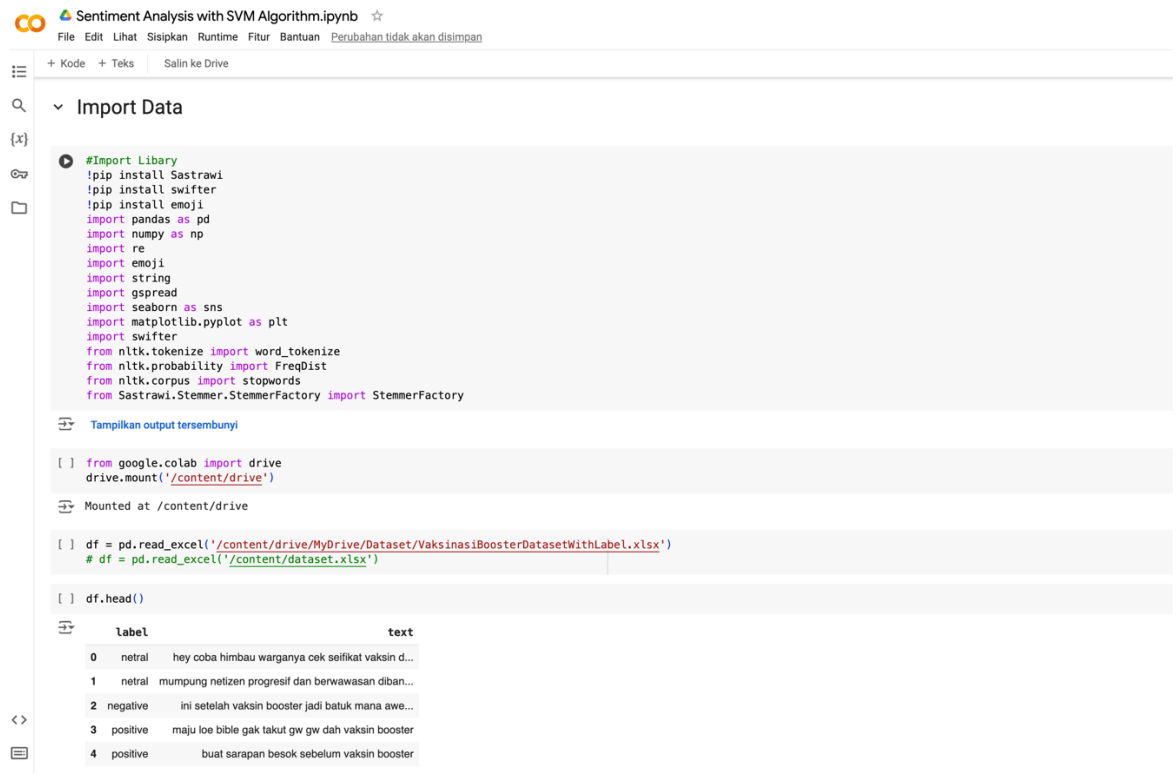


Figure 2 – Importing Collecting Data from Twitter and Labelling Data Processing

3.2. Data Labeling

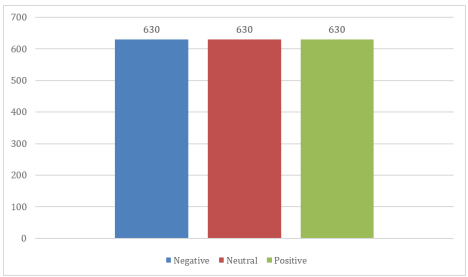
In sentiment analysis, data labeling is necessary for the obtained data. The data labeling process is conducted semi-manually using SentiStrength, followed by label validation to achieve more accurate and easily understandable results. Sentiment labels given to each data point are positive (1), negative (-1), and no sentiment or neutral (0).

Table 2 – Sample Data

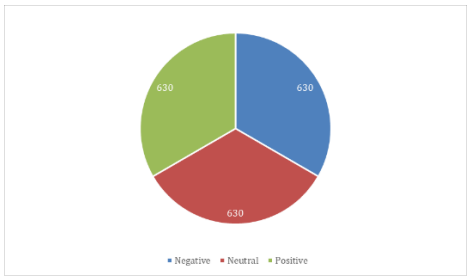
No	Text	Label
1	Stop vaksin dan tambah booster hanya akan memperparah mutasi virus dan merusak kesehatan kita efek jangka panjang	-1
2	Masyarakat hendaknya bisa segera mengikuti vaksin Booster	0
3	Tak perlu ragu untuk vaksin booster, vaksin dijamin aman karena sudah mendapatkan fatwa halal dari MUI	1

3.3. Data Visualization

The purpose of data visualization is to transform complex information and data into visual forms that are easier to understand and quickly interpretable by humans.



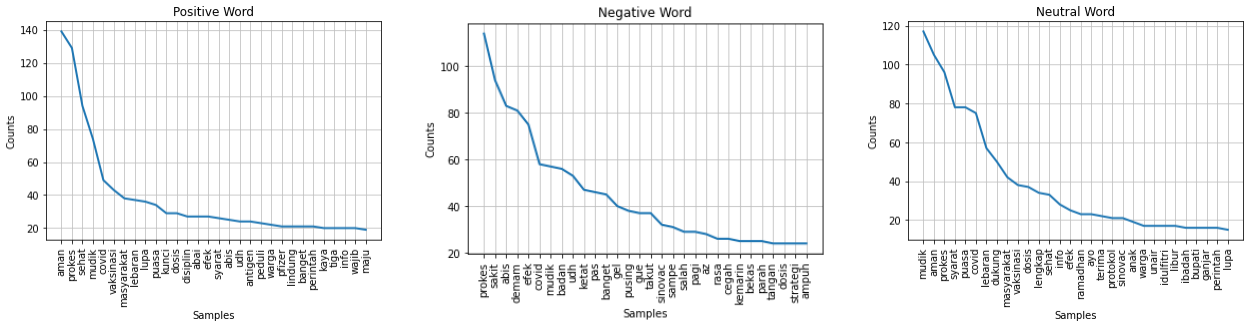
(a)



(b)

Figure 3 – Data Visualization (a) Bar Chart and (b) Pie Chart

In Figure 3, the distribution of data for each sentiment label is presented in the form of a bar chart and a pie chart, with each sentiment label having a total of 630 data.



(a) (b) (c)

Figure 4 – Wordcount for (a) positive sentiment, (b) negative sentiment, and (c) neutral sentiment

In Figure 4, the distribution of word count for each sentiment label is presented using a line chart, where the higher the line, the higher the frequency of word occurrences. In the positive sentiment, it can be observed that certain words such as "aman", "prokes" "sehat", and "mudik" have high frequencies of occurrences. In the negative sentiment, certain words like "prokes", "sakit", "abis", "demam", and "efek" have high frequencies of occurrences. Lastly, in the neutral sentiment, certain words like "mudik", "aman", "prokes", and "syarat" have high frequencies of occurrences.

3.4. Data Preprocessing

Data preprocessing refers to the steps taken to clean, transform, and prepare raw data into a format that can be used for analysis or modeling. The goal of preprocessing is to ensure that the data is of high quality, free from missing or erroneous values, and consistent in format so that accurate insights and predictions can be derived from the data. Here are several stages of data preprocessing in Table 3.

Table 3 – Preprocess Step with Example Result

Preprocessing Step	Result
Raw Data	RT @rockmansick: MAJU LOE BIBLE GAK TAKUT GW, GW DAH VAKSIN BOOSTER!!! #BibleBuildXFANDOMLive https://t.co/OyMLtweub4
Cleaning	MAJU LOE BIBLE GAK TAKUT GW GW DAH VAKSIN BOOSTER
Case folding	maju loe bible gak takut gw gw dah vaksin booster
Tokenizing	[‘maju’, ‘loe’, ‘bible’, ‘gak’, ‘takut’, ‘gw’, ‘gw’, ‘dah’, ‘vaksin’, ‘booster’]
Stopwords Removal	[‘maju’, ‘bible’, ‘gak’, ‘takut’, ‘dah’, ‘vaksin’, ‘booster’]
Stemming	[‘maju’, ‘bible’, ‘tidak’, ‘takut’, ‘sudah’, ‘vaksin’, ‘booster’]

3.5. Term Frequency – Inverse Document Frequency (TF – IDF)

After preprocessing the data by removing URLs, mentions, punctuation, performing case folding, tokenization, and stopword removal, the next stage is Term Frequency - Inverse Document Frequency (TF-IDF). This stage involves word weighting, which determines the frequency of terms within a word dataset. TF-IDF is applied to the independent variables in the sentiment dataset and is used to describe the importance of a term within a sentence in a collection or corpus. Term Frequency (TF) refers to the frequency or number of times a specific term appears in a document. Inverse Document Frequency (IDF) is a method that decreases the significance of terms that frequently occur in several documents by considering the reciprocal of the document frequency. During the initial phase, the frequency of word occurrence in documents (TF) is determined using the following equation:

$$tft = 1 + log\ (\ \boldsymbol{tf_t}\ ) \tag{13}$$

In the equation above, tft is the number of words from t.

Next, the calculation of documents containing certain words is carried out, then the inverse (IDF) will be calculated with the equation:

$$idft = log\ (D\ /\ \boldsymbol{df_t}) \tag{14}$$

Where idf$t$ is inverse document frequency, D is the number of all existing documents, and df$t$ is the number of documents containing the word t.

The last step in this process is to calculate the TF-IDF value by multiplying the TF result with the IDF calculation result with the equation:

$$W_{t,d} = tft \times idft \quad (15)$$

Where Wt,d is term weight (t) in document (d), tft is the number of occurrences of term t, $idft$ is the frequency of inverted documents containing term t.

3.6. Splitting Data

Before creating a machine learning model, the dataset that has been processed by TF-IDF will first be split data. The dataset is divided with a ratio of 80:20. The 80% dataset will be used as a training dataset that will be used to conduct training on the dataset, while the remaining 20% will be used as a testing dataset that will be used to evaluate the performance of the machine learning model. The result of splitting the dataset will produce 4 new variables, namely the X_train variable to perform model training for features data or independent variables, the y_train variable to perform model training for labels data or dependent variables, and the X_test and y_test variables to be used as evaluation data on models on independent and dependent variables.

3.7. SVM Model

Subsequently, the next course of action involves constructing a machine learning model utilizing the Support Vector Machine (SVM) model. The accuracy of the Support Vector Machine model is evaluated by two simulations: one before Hyperparameter Tuning and one after Hyperparameter Tuning. The hyperparameters used are:

a. Parameter C
   Parameter C refers to a specific variable or value that is used in a given context. The parameter C in Support Vector Machines (SVM) is a parametric value used to regulate the occurrence of faults in the SVM model. The value of C must be greater than zero. A smaller C value indicates a lower level of error, while a larger C value corresponds to a higher level of error. The C value utilized in this analysis is 1000.
b. Gamma
   Gamma is employed in conjunction with the rbf, poly, and sigmoid kernels. The gamma parameters dictate the extent to which a training example's influence extends, with lower values indicating a greater distance and higher ones indicating closer proximity. Models with lower gamma values tend to have lesser accuracy, whereas models with higher gamma values tend to have higher accuracy. The gamma value is limited to a range of 0 to 1. The gamma value utilized in this analysis is 0.05.
c. Kernel
   Kernels are utilized to transform input data into the necessary format for data processing. The term "kernel" is often in the context of Vector Machine Support to refer to a set of mathematical functions that enable the manipulation of data. The Support Vector Machine (SVM) offers a variety of kernels, including linear, nonlinear, polynomial, radial basis function (RBF), sigmoid, and precomputed.

3.8. Testing Result

Training data was collected consisting of 1,890 data points, which were then manually labeled. Additionally, there were 378 data for prediction. These data points were referred to as testing data. The sentiment will be predicted using the built model. The testing was conducted by creating two simulations, one before Hyperparameter Tuning and one after Hyperparameter Tuning. The performance measurement of the Support Vector Machine (SVM) model to predict public sentiment towards booster vaccinations through Twitter, and the prediction results using Confusion Matrix are presented. There are two Confusion Matrices used, one before and one after Hyperparameter Tuning in the Support Vector Machine (SVM) model. The prediction results from the model can be seen in Table 4 and Table 5.

Table 4 – Confusion Matrix Result without Hyperparameter Tuning

		Predict		
		Negative	Neutral	Positive
Actual	Negative	**105**	12	8
	Neutral	15	**101**	23
	Positive	6	13	**95**

Based on the confusion matrix results in Table 4 (without parameter tuning), the f1-score accuracy is 79.63%, precision is 79.3%, and recall is 80%.

*rahmatfauzi@telkomuniversity.ac.id

Table 5 – Confusion Matrix Result with Hyperparameter Tuning

		Predict		
		Negative	Neutral	Positive
Actual	Negative	**106**	12	8
	Neutral	11	**100**	20
	Positive	9	14	**98**

After hyperparameter tuning as shown in Table 5, the model performance slightly improved, with an f1-score accuracy of 80.42%, precision of 80.3%, and recall of 80.3%.

The results from Table 4 and Table 5 show that tuning hyperparameters led to a slight improvement in the performance of the SVM model. Specifically, there was a 0.79 increase in the f1-score, as well as consistent improvements in precision and recall across the three sentiment categories. Although the changes in evaluation metrics were not substantial, they highlight the positive influence of fine-tuning model parameters on its ability to accurately classify sentiments.

The marginal performance improvement resulting from hyperparameter tuning indicates that while the original model was already strong, optimizing specific parameters such as the regularization parameter (C) and kernel choice further enhances the model's ability to classify. This study emphasizes the significance of parameter optimization in machine learning models, particularly when working with complex data like public sentiment on social media.

The significance of this research goes beyond just the immediate enhancements in classification metrics. By demonstrating the effectiveness of SVM in sentiment analysis of social media data, this study establishes a useful framework that can be extended to other areas where public opinion holds critical importance. For example, the methodology used here could be adapted to monitor public sentiment on various health policies, enabling policymakers to better gauge public reaction and adjust their communication strategies accordingly.

In summary, the improvements achieved through hyperparameter tuning highlight the potential for further refinement and optimization in machine learning models, paving the way for more effective and precise sentiment analysis tools that can be applied in real-world scenarios. Subsequent research could explore the incorporation of more advanced techniques, such as deep learning or hybrid models, to further enhance classification performance and investigate the impact of these methods on larger and more diverse datasets.

4. CONCLUSION

After conducting experiments, it was determined that the Support Vector Machine (SVM) method achieved an 80.42% accuracy in sentiment prediction with parameter tuning. Incorporating hyperparameter tuning consistently enhanced the performance of the sentiment analysis model, resulting in a marginal but consistent improvement compared to the untuned model. Specifically, the f1-score increased by 0.79, with both precision and recall reaching 80.3%. These results show that while the differences in evaluation metrics may not be substantial, parameter tuning positively contributes to the model's ability to accurately recognize and classify sentiments across positive, neutral, and negative classes.

This study underscores the importance of optimizing machine learning models to achieve even minor performance enhancements, particularly in applications like public sentiment analysis on social media. The research confirms that the tuned SVM model is a dependable tool for analyzing public opinions on health policies, such as booster vaccinations. Additionally, it lays the groundwork for future research that could explore and compare advanced machine learning techniques, like deep learning or ensemble methods, potentially leading to improved accuracy and deeper insights into public sentiment trends.

Moreover, the impact of this research goes beyond the immediate analysis of Twitter data, providing valuable insights for policymakers and public health officials. They can utilize these findings to gain a better understanding of public sentiment and adjust their communication strategies accordingly. Future research should consider expanding the dataset, employing more advanced text processing techniques, and exploring the integration of additional features, such as demographic information, to enrich the analysis and improve the model's performance.

In summary, this study illustrates that even minor improvements in machine learning models, such as those achieved through parameter tuning, can have significant implications in practical applications, especially in areas where understanding public sentiment is critical.

*rahmatfauzi@telkomuniversity.ac.id

REFERENCES

[1] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem dan Informatika (JSI)*, vol. 15, no. 1, pp. 27–29, Nov. 2020, doi: 10.30864/jsi.v15i1.332.

[2] R. Al Habsi, R. A. D. Anggoro, M. A. Valio, Y. Widiastiwi, and N. Chamidah, "Analisis Sentimen Terhadap Vaksin Covid-19 di Jejaring Sosial Twitter Menggunakan Algoritma Naïve Bayes," in *Prosiding Seminar Nasional Informatika, Sistem Informasi Dan Keamanan Siber (SEINASI-KESI) 2021*, Sep. 2021, pp. 239–248.

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank," in *Proceedings of the third ACM international conference on Web search and data mining*, New York, NY, USA: ACM, Feb. 2010, pp. 261–270. doi: 10.1145/1718487.1718520.

[4] N. Hardi, Y. Alkahfi, P. Handayani, W. Gata, and M. R. Firdaus, "Analisis Sentimen Physical Distancing pada Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier," *SISTEMASI*, vol. 10, no. 1, p. 131, Jan. 2021, doi: 10.32520/stmsi.v10i1.1118.

[5] A. Alamsyah, W. Rizkika, D. D. A. Nugroho, F. Renaldi, and S. Saadah, "Dynamic Large Scale Data on Twitter Using Sentiment Analysis and Topic Modeling," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, IEEE, May 2018, pp. 254–258. doi: 10.1109/ICoICT.2018.8528776.

[6] I. P. Wardhani, Y. I. Chandra, and F. Yusra, "Application of the Naïve Bayes Classifier Algorithm to Analyze Sentiment for the Covid-19 Vaccine on Twitter in Jakarta," *International Journal of Innovation in Enterprise System*, vol. 7, no. 01, pp. 1–18, 2023, doi: 10.25124/ijies.v7i01.171.

[7] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Soc Netw Anal Min*, vol. 13, no. 1, pp. 1–14, 2023, doi: 10.1007/s13278-023-01030-x.

[8] H. Hayati and M. R. Alifi, "ANALISIS SENTIMEN PADA TWEET TERKAIT VAKSIN COVID-19 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE," *JTT (Jurnal Teknologi Terapan)*, vol. 7, no. 2, p. 110, Oct. 2021, doi: 10.31884/jtt.v7i2.349.

[9] R. A. Raharjo, I. M. G. Sunarya, and D. G. H. Divayana, "Perbandingan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Kasus Analisis Sentimen Terhadap Data Vaksin Covid-19 Di Twitter," *Elkom : Jurnal Elektronika dan Komputer*, vol. 15, no. 2, pp. 456–464, Dec. 2022, doi: 10.51903/elkom.v15i2.918.

[10] J. W. Creswell and J. D. Creswell, *Mixed Methods Procedures*. 2018.

[11] R. S. Pressman, *Software engineering: a practitioner's approach. Palgrave macmillan.* 2014.

[12] E. Bird, S., Klein, E., & Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc., 2009.

[13] R. McKerlich, C. Ives, and R. McGreal, "Sentiment Strength Detection in Short Informal Text," *International Review of Research in Open and Distance Learning*, vol. 14, no. 4, pp. 90–103, 2013.

[14] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[15] N. D. Putranti and E. Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 8, no. 1, p. 91, Jan. 2014, doi: 10.22146/ijccs.3499.

[16] D. H. Wahid and A. SN, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 10, no. 2, p. 207, Jul. 2016, doi: 10.22146/ijccs.16625.

[17] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 406, Apr. 2021, doi: 10.30865/mib.v5i2.2835.

[18] R. Coughlin, J.-C. Coetsier, and R. Jiamthapthaksin, "Integrating Labeled Latent Dirichlet Allocation into sentiment analysis of movie and general domains," in *2017 9th International Conference on Knowledge and Smart Technology (KST)*, IEEE, Feb. 2017, pp. 18–22. doi: 10.1109/KST.2017.7886071.

[19] F. E. Cahyanti, Adiwijaya, and S. Al Faraby, "On The Feature Extraction For Sentiment Analysis of Movie Reviews Based on SVM," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Jun. 2020, pp. 1–5. doi: 10.1109/ICoICT49345.2020.9166397.

[20] I. Subagyo, L. D. Yulianto, W. Permadi, A. W. Dewantara, and A. D. Hartanto, "Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM," *JURNAL SISTEM INFORMASI DAN TEKNOLOGI INFORMASI* , vol. 7, no. 1, pp. 47–56, 2019.

[21] E. Tyagi and A. K. Sharma, "Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm," *Indian J Sci Technol*, vol. 10, no. 35, pp. 1–9, Jun. 2017, doi: 10.17485/ijst/2017/v10i35/118965.