The Performance of Machine Learning Model Bernoulli Naïve Bayes, Support Vector Machine, and Logistic Regression on COVID-19 in Indonesia using Sentiment Analysis

Wahyu Dirgantara[1], Fairuz Iqbal Maulana[2], Subairi Subairi[3], Rahman Arifuddin[4]

[1,3,4]Department of Electrical Engineering,
Faculty of Engineering,
Universitas Merdeka Malang, Malang
[1]wahyu.dirgantara@unmer.ac.id, [3]subai@unmer.ac.id, [4]rahman.arifudddin@unmer.ac.id

[2]Department Computer Science Department,
School of Computer Science,
Bina Nusantara University, Jakarta
[2]fairuz.maulana@binus.edu

Abstract

The COVID-19 pandemic has significantly impacted Indonesia, necessitating a deeper understanding of public sentiment towards the crisis. This study investigates the performance of three prominent machine learning models: Bernoulli Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression, in analyzing sentiments related to COVID-19 in Indonesia. Utilizing a dataset comprising social media posts, the research aims to classify sentiments into positive, and negative categories, providing insights into the public's perception of the pandemic and associated measures. Sentiment analysis serves as a powerful tool to capture the collective emotions and opinions of the populace, which are pivotal in shaping public health responses and policies. The accuracy of LR and SVM is 99%, whereas Bayesian has an accuracy of 98%. We conclude that Logistic Regression and Support Vector Machine are the best model for the above dataset. This research evaluates these models' accuracy and reliability in the context of the Indonesian language, which influence sentiment interpretation. The findings of this study will contribute to the fields of natural language processing and public health by highlighting the efficacy of machine learning models in sentiment analysis during a health crisis. Moreover, the results will assist policymakers and health officials in understanding public sentiment, enabling them to tailor communication and interventions more effectively.

Keywords: COVID-19, Machine Learning, Bernoulli Naïve Bayes, Support Vector Machine, Logistic Regression, Sentiment Analysis

1. Introduction

The COVID-19 pandemic has had a profound impact on countries worldwide, including Indonesia, where the virus has caused significant infections and fatalities. As a response to the pandemic, various measures, such as social distancing and online learning, have been implemented to mitigate the spread of the virus and its associated risks. As the

Received: 17 Januari 2024 Revised: 19 Maret 2024 Accepted: 19 Maret 2024

world grapples with the effects of the virus, the scientific community has turned to innovative approaches to understand and mitigate its impact. One such approach is the application of machine learning models to analyze sentiments expressed in digital media, providing valuable insights into public perception and response to the pandemic.

In this context, the sentiment of the population towards these measures, has become a subject of interest. Among the many branches of NLP, sentiment analysis offers a valuable approach to understanding public opinion by classifying text data into positive, negative, or neutral sentiments [1], [2]. Among the several branches of NLP known as sentiment analysis, involves the computational study of opinions, emotions, and attitudes expressed in text. It has become an essential tool for gauging public sentiment on various topics, including the ongoing health crisis. In Indonesia, sentiment analysis has been employed to understand public reactions to COVID-19 vaccines [3], government policies [4], and the overall sentiment towards the pandemic's handling [5]. These studies have utilized platforms like Twitter to collect and analyze data, offering a snapshot of the collective mindset during these challenging times.

In the field of computer science, machine learning models, including NB, SVM, and LR, have been widely used for sentiment analysis. Bernoulli Naïve Bayes [6], a probabilistic classifier based on the Bayes theorem, is particularly suited for binary or dichotomous outcomes. It assumes feature independence and is commonly used in text classification tasks where binary features are predominant. SVM, on the other hand, constructs a hyperplane in a high-dimensional space to separate different classes [2]. It is renowned for its efficacy in managing non-linear connections and data with a high number of dimensions. Logistic Regression, a statistical model that predicts the probability of a binary outcome, because of how efficient and easy it is to understand, it is also utilized extensively in sentiment analysis [7]. Movie reviews are only one of several domains that have made use of these models [8] and COVID-19 vaccination sentiment analysis [9], to assess public opinion and sentiment towards specific topics.

In the context of COVID-19 in Indonesia, these machine learning models have been applied to various datasets to predict sentiments [10]. For instance, studies have shown that SVM achieved high accuracy in classifying sentiments towards Sinovac and Pfizer vaccines in Indonesia [3]. Similarly, Naïve Bayes outperformed other models in analyzing sentiments of commuter line passengers regarding the transmission of COVID-19. These findings underscore the potential of machine learning models in providing actionable insights during a health crisis [11]. The objective of this study is to assess the efficacy of NB, SVM, and LR in conducting sentiment analysis of COVID-19 in Indonesia. We will utilize an extensive collection of social media postings to evaluate the models' capacity to precisely categorize attitudes as positive, and negative [12]. The study will enhance comprehension of public attitude during the epidemic and facilitate the formulation of focused communication strategies to tackle public worries and misconceptions.

Furthermore, the research will explore the implications of model choice on the accuracy and reliability of sentiment analysis. The investigation will assess the capabilities and constraints of each model within the specific setting of Indonesian language and cultural subtleties, which might impact the understanding of emotions. The study will also consider the impact of data preprocessing techniques, such as tokenization, stemming, and lemmatization, on model performance. The research will provide a comprehensive evaluation of machine learning models in the context of COVID-19 sentiment analysis in Indonesia. It will offer insights into the effectiveness of these models in capturing the

complexities of human emotions and opinions during a global health crisis. The findings will not only contribute to the field of NLP and sentiment analysis but also support public health officials and policymakers in making informed decisions based on public sentiment.

2. Research Method

A literature review conducted on sentiment analysis and opinion research pertaining to social concerns, utilizing data extracted from newspaper websites. This study asserts that the combination of several categorization algorithms might yield superior outcomes. Furthermore, three criteria for evaluating performance in information retrieval include precision, recall, and F-score. Social media sentiment analysis has been extensively utilized in investigated themes [13]–[15]. The comprehension of a product's impression by the public or customers is highly advantageous for company marketing tactics.

The aim of this research is to develop a robust model for predicting the mood expressed in tweets related to COVID-19 [16]. There are individuals who have the belief that COVID-19 is an actual and perilous infection, but others maintain the perspective that it is only a rumor. The procedural phases of our study are depicted in Figure 1, and elaborations on these processes are provided in the subsequent subsections. Figure 1 illustrates the initial stage of implementing the suggested methodology, which involves gathering Indonesian language tweets on COVID-19 over the Twitter API, using predetermined search keywords. Subsequently, the tweets are stored in a CSV file. During this phase, our objective is to comprehend the available data and detect any possible issues present within the dataset. The subsequent phase involves data preparation. Currently, we execute many sequential actions. Before conducting sentiment polarization, we carry out a series of data preprocessing steps including cleaning, case-folding, tokenizing, filtering, and stemming. This procedure yields what we often refer to as preprocessed data. Subsequently, after an understanding of sentiment polarization, we proceeded to conduct modelling utilizing the three aforementioned models: Naïve Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR). We assessed the outcomes by employing a confusion matrix and ROC curve to choose the model with the highest level of performance. The illustrates the sequence of steps followed in Figure 1.

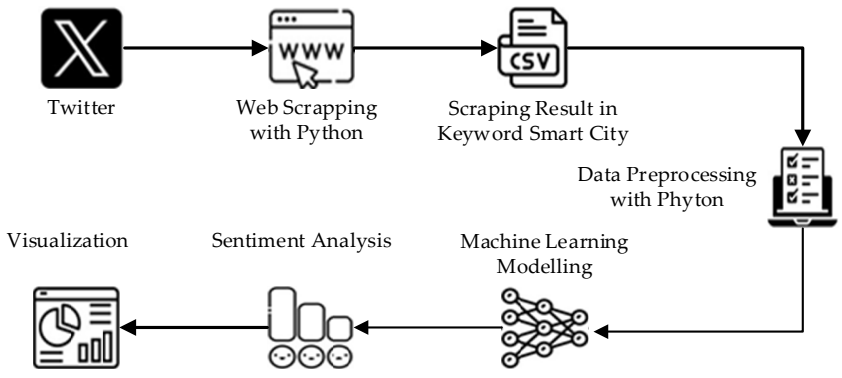


Figure 1. Wordcloud Positive

2.1. Data Collection

The collection of Indonesian language tweets was conducted using Tweet Harvest 2.2.8 on Google Colab, utilizing the Python computer language. The tweets that have been gathered are written in either Indonesian or the standard dialect. We have acquired a total

of 794 tweets in the Indonesian language that are relevant to COVID-19. Search queries are determined by analyzing the phrases that are most often used by users on social media sites while discussing COVID-19. The acquired data is stored in a CSV file for the preprocessing phase. The scraping method imposes limitations on the quantity of tweets that may be retrieved due to the existence of a daily restriction set by Twitter or X for specific account types.

	id_str	tweet	lang	conversation_id_str
0	1.730000e+18	waspada kenaikan covid di indonesia	NaN	NaN
1	1.730000e+18	tak payahlah nak tunggu sampai ada diskaun bar...	NaN	NaN
7	1.730000e+18	waspada kenaikan covid di indonesia virus prok...	in	1.73E+18
8	1.730000e+18	news mutasi ratusan perwira polri	NaN	NaN
9	1.730000e+18	menjadi gereja selama pandemi covid berlangsung	NaN	NaN

Figure 2. Sample of Raw Data

Once the tweets have been gathered, the subsequent task is comprehending the data in order to gain a comprehensive picture of the data gathering process. Gaining comprehension of data is the initial stage in the process of data analysis. The data is scrutinized to identify any existing data anomalies. In addition to that, it is possible to create a summary and identify potential issues. Precision is crucial at this stage as it will directly impact the outcomes in the subsequent step. As seen in Figure 2, the file was first imported into the Jupyter Notebook workspace and then read. Following that, we used the pandas package to investigate the dataset's complexities. This is the initial data set that we have collected, often known as raw data.

2.2. Data Processing

This step is conducted to rectify the issues that were present in the preceding stage. This step also assesses the appropriateness of the data for the algorithm to be employed, as it is preferably revisited several times over the modelling process to identify and resolve any issues until an acceptable solution is achieved. To guarantee that the data is ready for modelling, the tasks include data selection, transformation, and purification. Hence, the process of cleansing the unprocessed data is a crucial step in producing our dataset. Nevertheless, the task of cleaning textual material is not a straightforward endeavor, since it frequently involves the presence of superfluous and/or duplicated terms.

	id_str	tweet	lang	conversation_id_str	Cleaned_Tweets	target	tokenized_tweets	tokenized_tweets_stemmed
0	1.730000e+18	waspada kenaikan covid di indonesia	NaN	NaN	Waspada kenaikan covid-19 di Indonesia	1	[waspada, kenaikan, covid, di, indonesia]	waspada kenaikan covid di indonesia
1	1.730000e+18	tak payahlah nak tunggu sampai ada diskaun bar...	NaN	NaN	Tak payahlah nak tunggu sampai ada diskaun bar...	1	[tak, payahlah, nak, tunggu, sampai, ada, disk...	tak payahlah nak tunggu sampai ada diskaun bar...

Figure 3. Data Processing

The gathered tweets are unfiltered. This tweet contains extraneous elements such as stop words and special characters, necessitating the implementation of preprocessing procedures to ready the data for the categorization process. Initially, we eliminate stop words, emojis, mentions, numbers, white spaces, duplicate lines, and useless columns. Additionally, we convert text documents to lowercase to enhance generalization. Furthermore, we cleanse punctuation to reduce unnecessary noise in the dataset. We also remove repeated characters from words and eliminate URLs/hyperlinks, as they lack significant importance. Subsequently, we will do stemming, which involves reducing the words to their respective base forms. Next, continue with lemmatization, which involves reducing the derived term to its basic form, known as a lemma, in order to get improved outcomes.

2.3. Evaluation Metric

Assessing the effectiveness of machine learning models on unbalanced datasets just based on accuracy is inadequate for evaluating the model's quality due to the accuracy paradox. The evaluation of the classifier output in this study involved assessing its quality using many measures, including accuracy, recall, F1-score, GM, and AUC. This approach aimed to ensure a dependable and impartial review. One way to quantify the ML model's accuracy in identifying positives is by looking at its recall, which is the ratio of true positive predictions to the total number of these cases. The proportion of correct predictions to the total number of correct predictions made by the model is called precision. The measures of accuracy and recall are combined by GM and F1-score. The accuracy of GM is evaluated on both positive and negative class samples, F1 is obtained by taking the harmonic mean of accuracy and recall. The Receiver Operator Characteristic (ROC) curve is a graph that provides a clear representation of the performance of a binary classifier. It summarizes the performance using a single metric known as the area under the curve (AUC), which measures the area under the ROC curve. Numbers ranging from 4 to 10 have been utilized to create several performance indicators expressed as percentages. These indicators are derived from the confusion matrix generated by the algorithm.

$$accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number\ of\ Sentiment} \times 100\% \quad (1)$$

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\% \quad (2)$$

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negative}{False\ Positive + True\ Negative} \times 100\% \quad (3)$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{False\ Positive + True\ Negative} \times 100\% \quad (4)$$

$$False\ Negative\ Rate\ (FNR) = \frac{False\ Negative}{False\ Negative + True\ Positive} \times 100\% \quad (5)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\% \quad (6)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (7)$$

3. Results and Discussion

Once the data has been processed and the relevant columns have been identified, the last stage involves visualizing the data using a mathematical plot. The rationale behind employing a plot is that it enhances the clarity of the data, hence facilitating comprehension. Figure 4 of the data visualization shows that there are more positive tweets than negative

tweets, and this difference is rather substantial. Each color corresponds to a certain column in the data distribution, namely ID, date, lang, user, and text.

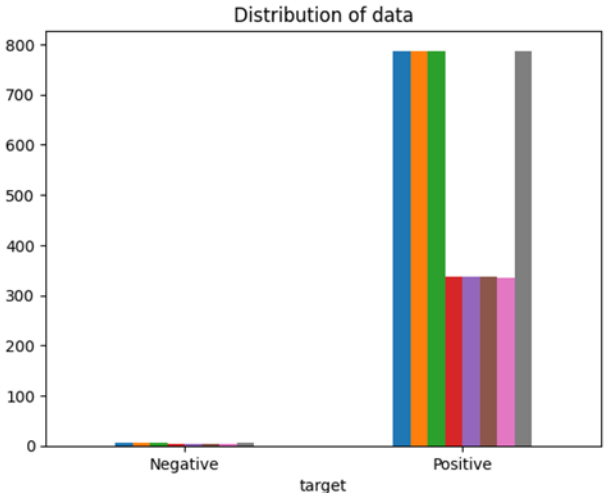


Figure 4. Distribution of Data in this Topic

Prior to doing machine learning, it is advisable to have a comprehensive understanding of the correctness of our data. To accomplish this, we will employ a word cloud, which is a compilation of words depicted in varying sizes. A word's prominence in a text is proportionate to how often and how meaningful it is. In our scenario, we anticipate that the word cloud will include a selection of terms that accurately reflect the category we are visualizing. We will create a word cloud for positive tweets (Figure 5 (b)) and another for negative tweets (Figure 5 (a)) in order to identify the most frequently utilized terms in each tweet category.

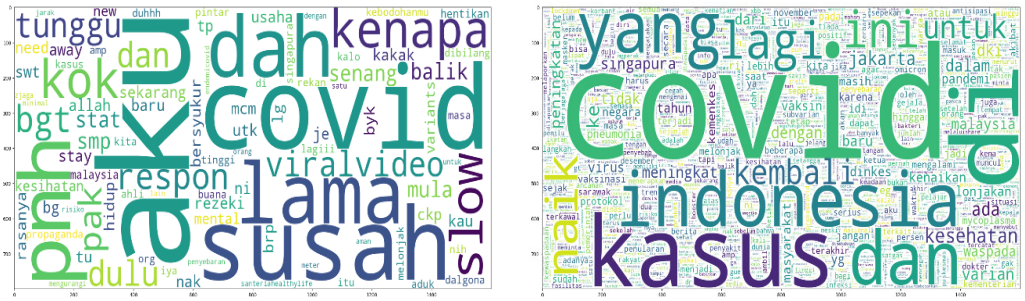


Figure 5. a) Wordcloud Negative, b) Wordcloud Positive

3.1. Modelling

In order to optimize efficiency, we will utilize the entire dataset for certain models, but for computationally intensive models, we will just employ 10% of the original dataset. Generate a novel variable, df_reduced, which comprises a randomized subset of the dataset.

	id_str	tweet	lang	conversation_id_str	Cleaned_Tweets	target	tokenized_tweets	tokenized_tweets_stemmed	tokenized_tweets_stemmed_lemmatized	text_length
163	1.730000e+18	meski semakin ringan	NaN	NaN	Meski Semakin Ringan	1	[meski, semakin, ringan]	meski semakin ringan	meski semakin ringan	20
181	1.730000e+18	sarawak ambil langkah berjagajaga jika jumlah ...	in	1.73E+18	Sarawak ambil langkah berjaga-jaga jika jumlah...	1	[sarawak, ambil, langkah, berjagajaga, jika, j...	sarawak ambil langkah berjagajaga jika jumlah ...	sarawak ambil langkah berjagajaga jika jumlah ...	168
623	1.730000e+18	ntar dibilang konspirasi lagi bang	NaN	NaN	Ntar dibilang KONSPIRASI lagi bang	1	[ntar, dibilang, konspirasi, lagi, bang]	ntar dibilang konspirasi lagi bang	ntar dibilang konspirasi lagi bang	34
591	1.730000e+18	dl majalah tempat drosten menjabat sbg editor ...	in	1.73E+18	... dl majalah tempat Drosten menjabat sbg edi...	1	[dl, majalah, tempat, drosten, menjabat, sbg, ...	dl majalah tempat drosten menjabat sbg editor ...	dl majalah tempat drosten menjabat sbg editor ...	170
142	1.730000e+18	dipenghujung tahun	NaN	NaN	Dipenghujung tahun	1	[dipenghujung, tahun]	dipenghujung tahun	dipenghujung tahun	18

Figure 6. Distribution of Data in this Topic

Once the model has been trained, we proceed to use assessment methods to assess its performance. Consequently, we employ the below assessment criteria to assess the performance of the models:

a. **Accuracy Score**: Generally, the prediction model has a high level of accuracy, typically exceeding 90%;
b. **The ROC-AUC curve** The ROC curve is a simple depiction of the ability of a classifier to discriminate across classes. As the AUC (Area Under the Curve) of a model grows, so does its ability to distinguish between positive and negative classifications;
c. **Plot of the confusion matrix**: A Confusion matrix is a N × N square matrix used to evaluate the performance of a classification model. Here, N represents the total number of classes that the model is attempting to predict. The matrix juxtaposes the factual goal values with the ones anticipated by the machine learning model.

In the problem statement, we have employed three distinct models, each with its own unique characteristics:

a. Model 1: Bernoulli Naïve Bayes (Figure 7 (a));
b. Model 2: Support Vector Machine (Figure 7 (b));
c. Model 3: Logistic Regression (Figure 7 (c)).

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.99	0.99	0.99	104
accuracy			0.98	105
macro avg	0.50	0.50	0.50	105
weighted avg	0.98	0.98	0.98	105

(a)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.99	1.00	1.00	104
accuracy			0.99	105
macro avg	0.50	0.50	0.50	105
weighted avg	0.98	0.99	0.99	105

(b)

0	0.00	0.00	0.00	1
1	0.99	1.00	1.00	104
accuracy			0.99	105
macro avg	0.50	0.50	0.50	105
weighted avg	0.98	0.99	0.99	105

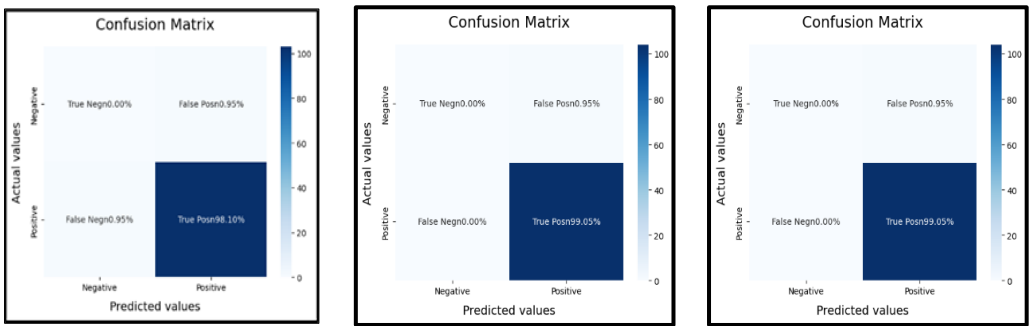
(c)

Figure 7. a) Wordcload Negative, b) Wordcload Positive

The rationale for selecting these models is to evaluate all the classifiers on the dataset, spanning from basic to intricate models, in order to identify the one that exhibits optimal performance.

3.2. Evaluation

In order to assess the performance of the used model, we employ the confusion matrix obtained for each model in conjunction with its ROC curve. Figure 8 (a) displays a confusion matrix generated by the Bernoulli Naïve Bayes model. Figure 8 (b) represents a Support Vector Machine. Lastly, Figure 8 (c) exhibits a confusion matrix derived by Logistic Regression. The confusion matrix in Figure 11 displays the prediction results for the true positive (98.10%), true negative (0.00%), false positive (0.95%), and false negative (0.95%) for the NB model. The SVM confusion matrix in Figure 12 shows that the prediction result for true positive is 99.05%, true negative is 0.00%, false positive is 0.95%, and false negative is 0.00%. The LR confusion matrix in Figure 8 shows that the prediction result is identical to that of SVM.



(a) (b) (c)

Figure 8. a) Bernoulli Naïve Bayes, b) Support Vector Machine, c) Logistic Regression

After evaluating all models, we can conclude the following details:

Table 1. Result from Every Model

Model id	Model Name	Accuracy %	F1-score (class 0) %	F1-score (class 1) %	Training execution time in seconds	Testing execution time in seconds
1	Bernoulli Naïve Bayes	98	0	99	0.00	0.08
2	Support Vector Machine	99	0	100	0.00	0.09
3	Logistic Regression	99	0	100	0.84	0.09

We conclude that Logistic Regression & Support Vector Machine are the best model for the above dataset.

4. Conclusion

The results of testing three models, namely LR, SVM, and NB, allow one to draw the conclusion that LR and SVM demonstrate greater accuracy in comparison to Bernoulli NB in terms of model performance. This conclusion may be reached based on the findings of the testing. In contrast to Bayesian, which has an accuracy of 98%, LR and SVM both have an accuracy of precisely 99%. Following the completion of our investigation, we came to the conclusion that Logistic Regression (LR) and Support Vector Machines (SVM) are the models that are the best suitable for the dataset that was submitted to us. Occam's Razor, which argues that, for a given issue, if the data lacks assumptions, the simplest straightforward model will offer the optimum results, is adhered to by Logistic Regression (LR) in our circumstance. This is because Logistic Regression allows us to avoid making

assumptions. In light of the fact that our dataset does not contain any assumptions that are underlying it and that linear regression is a basic model, the ideas that have been presented are relevant to the dataset that was described before.

References

[1] S. H. Sahir, R. S. A. Ramadhana, M. F. R. Marpaung, S. R. Munthe, and R. Watrianthos, "Online learning sentiment analysis during the covid-19 Indonesia pandemic using twitter data," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1156, no. 1, p. 12011.

[2] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, 2022.

[3] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment analysis on Covid19 vaccines in Indonesia: from the perspective of Sinovac and Pfizer," in *2021 3rd East Indonesia conference on computer and information technology (EIConCIT)*, 2021, pp. 122–127.

[4] B. Sujiwo, A. Wibowo, and D. R. S. Saputro, "Sentiment Analysis of Indonesian Government Policies In Handling Covid 19 Through Twitter Data," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2021, pp. 453–457.

[5] I. C. Sari and Y. Ruldeviyani, "Sentiment analysis of the covid-19 virus infection in indonesian public transportation on twitter data: A case study of commuter line passengers," in *2020 International Workshop on Big Data and Information Security (IWBIS)*, 2020, pp. 23–28.

[6] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019.

[7] F. Fazrin, O. N. Pratiwi, and R. Andreswari, "Comparison of K-Nearest Neighbor and Logistic Regression Algorithms on Sentiment Analysis of Covid-19 Vaccination on Twitter with Vader And Textblob Labeling," in *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 2022, pp. 39–44.

[8] B. Sangeetha, S. Sangeetha, and D. T. Goutham, "Sentiment Analysis on Movie Reviews: A Comparative Analysis," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 2023, pp. 218–223.

[9] F. I. Maulana, P. D. P. Adi, D. Lestari, A. Purnomo, and S. Y. Prihatin, "Twitter Data Sentiment Analysis of COVID-19 Vaccination using Machine Learning," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2022, pp. 582–587.

[10] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert Syst. Appl.*, vol. 212, p. 118715, 2023.

[11] P. Monika, C. Kulkarni, N. H. Kumar, S. Shruthi, and V. Vani, "Machine learning approaches for sentiment analysis: A survey," *Int. J. Health Sci. (Qassim).*, vol. 6, no. S4, pp. 1286–1300, 2022.

[12] F. I. Maulana, Y. Heryadi, W. Suparta, and Y. Arifin, "Social Media Analysis using Sentiment Analysis on COVID-19 from Twitter," in *2022 6th International Conference on*

Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022, pp. 286–290.

[13]T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 70–77.

[14]B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, Springer, 2012, pp. 415–463.

[15]C. Chew and G. Eysenbach, "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak," *PLoS One*, vol. 5, no. 11, p. e14118, 2010.

[16]F. P. E. Putra, F. I. Maulana, N. M. Akbar, and W. Febriantoro, "Twitter sentiment analysis about economic recession in indonesia," *Bull. Soc. Informatics Theory Appl.*, vol. 7, no. 1, pp. 1–7, 2023.