p-ISSN 2301–4156 | e-ISSN 2460–5719

Indonesian Society's Sentiment Analysis Against the COVID-19 Booster Vaccine

Dionisia Bhisetya Rarasati[1], Angelina Pramana Thenata[2], Afiyah Salsabila Arief[3]

[1,2,3] Informatics Study Program, Faculty of Technology and Design, Universitas Bunda Mulia, Tangerang 15143 INDONESIA (tel. : 021 80821428; email: [1]dionisia.rarasati@gmail.com, [2]angelina.pramana@outlook.com, [3]afiiyahsarief@outlook.com)

[Received: 10 April 2023, Revised: 4 August 2023]
Corresponding Author: Dionisia Bhisetya Rarasati

ABSTRACT — The COVID-19 pandemic is still occurring in various countries, including Indonesia. This pandemic is caused by the coronavirus, which has mutated into multiple virus variants, such as Delta and Omicron. As of 9 February 2022, 4,626,936 people were confirmed positive for COVID-19 in Indonesia. This number continues to rise. The Indonesian government has prevented the spread of these virus variants by introducing booster vaccines to the public. However, this vaccination program has caused various sentiments among Indonesians. To optimize efforts to combat COVID-19, the government needs to know these sentiments immediately. Based on these problems, the researcher proposes the application of machine learning technology to develop a system that can analyze the sentiments of the Indonesians toward the booster vaccine. This research has several stages: data collection, data labeling, text preprocessing, feature extraction, and application of the support vector machine (SVM) algorithm using various kernels, namely the linear kernel, Gaussian radial basis function (RBF) kernel, and polynomial kernel. Furthermore, the results of the system were tested for accuracy using a 10-fold cross validation and confusion matrix. The dataset used was 681 tweets with the hashtag "vaksinbooster." The dataset consists of two classes: negative (0) and positive (1). The results showed that the data were positive for the booster vaccine, as evidenced by the higher number of positive tweets, with 554 data, compared to 127 negative tweets. In addition, the dataset was divided into training data of 545 and testing data of 136. In addition, the test results of this study revealed that the SVM algorithm with the polynomial kernel, which was evaluated with 10-fold cross validation, yielded the highest level of accuracy, namely 79.22%.

KEYWORDS — COVID-19, Booster Vaccine, Sentiment Analysis, Support Vector Machine.

I. INTRODUCTION

The coronavirus, commonly known as COVID-19, is still spreading almost worldwide, including in Indonesia. Various variants of the COVID-19 spreading in Indonesia include Alpha, Delta, Beta, Kappa, and Omicron. As of 9 February 2022, a total of 4,626,936 Indonesians were tested positive for COVID-19 and this number is increasing [1]. It has then prompted the government to put a greater effort to suppress the spread of the COVID-19's various variants by conducting massive vaccination for Indonesians.

In Indonesia, the COVID-19 vaccination itself consists of two mandatory doses, and the most recent of which is the booster vaccination. The Indonesian government is striving to socialize and invite the Indonesians to have booster vaccinations in order to suppress the surge in COVID-19 cases. However, these government's efforts have induced various sentiments among the Indonesians. These sentiments must be immediately known by the government to optimize efforts to combat the numerous variants of the COVID-19. These various sentiments are usually poured into social media such as Twitter, where users can exchange information through text, images, and video [2].

Because of these issues, this research aims to develop a system that can analyze the sentiments of Indonesians on Twitter toward the booster vaccine. This analysis will classify positive and negative groups using the support vector machine (SVM) algorithm. The SVM algorithm will characterize it by developing an N-dimensional hyperplane that isolates information into two types of classification (positive and negative) [3]. Furthermore, the results of the system will be tested for accuracy using a 10-fold cross validation [4] and confusion matrix [5]. The sentiment analysis results are expected to be input for the government or interested parties to decide on the next steps to suppress the spread of the COVID-19.

This paper is organized as such. Section II discusses research related to this study and introduce the methods used. Section III introduce the methodology that is formulated to achieve the research goal. Examples of some methodology processes are also presented in this section. Then, Section IV presents the result of the research, along with the implementation of the formulated methodology. Finally, Section V is the conclusion of the research conducted.

II. SENTIMENT ANALYSIS

Sentiment analysis is conducted to know public/user opinions on a topic or platform. The sentiment analysis process generally uses data from the internet and social media/platforms. A study whose data was sourced from Facebook examines public sentiment analysis towards Indonesia's presidential and vice-presidential candidates in 2019. The research resulted in the presidential and vice-presidential candidates, Jokowi-Ma'ruf received 56.76% positive sentiment and 43.24% negative sentiment, while Prabowo-Sandi received 24.21% positive sentiment and 75.79% negative sentiment. The said results were obtained using the naïve Bayes algorithm [6].

A sentiment analysis study on COVID-19 was done in 2019. With data sourced from Twitter, the researcher used the k-nearest neighbors (KNN) and naïve Bayes algorithm to attain results. The results of the sentiment test obtained an accuracy rate of 63.21% for the naïve Bayes and 58.10% for the KNN. Then, the precision obtained was 59.11% for the naïve Bayes and 53.10% for the KNN. In addition, it was found that the

tendency of public sentiment on Twitter was positive. It is evidenced by the fact that there were 610 positive sentiments and 488 negative sentiments [7].

Research done in 2020 compared the SVM, random forest (RF), and stochastic gradient descent (SGD) algorithms to classify the performance (good or bad) of programmers during social media activities. It obtained accurate results with cross-validation of SVM (81.3%), RF (74.4%), and SGD (80.1%). These results indicate that the SVM algorithm performs better than the other two algorithms in classifying programmer performance (good or bad) during social media activities [8]. Furthermore, most related research descriptions only examine sentiment analysis of public opinion regarding the presidential election and the COVID-19 outbreak using the naïve Bayes algorithm.

However, public opinion on the COVID-19 booster vaccination, which is an effort from the government to suppress the surge in COVID-19 cases, has yet to be investigated. In addition, the SVM algorithm has good performance in grouping datasets. Thus, using the SVM algorithm, this study applies machine learning technology to analyze public sentiment on the COVID-19 booster vaccination from Twitter.

III. MACHINE LEARNING

Machine learning is a part of artificial intelligence that is widely used to solve various problems through learning through data in various forms, such as texts, numbers, images, and videos. The learning process from these data are obtained through two stages: training and testing [9]. The discovery of interesting knowledge is obtained from data in the form of texts. Machine learning works with various algorithms such as decision trees, naïve Bayes, KNN, RF, and SVM [10].

IV. SUPPORT VECTOR MACHINE

The SVM algorithm aims to find the best hyperplane that can perfectly separate two classes with the widest margins. Margins are described as the distance between the said hyperplane with the nearest support vectors of each class, while support vectors can be described as the furthest data point of each class in the hyperplane [5]. The SVM theory has been evolving since the 1960s, but it was not introduced until 1992 by Vapnik, Boser, and Guyon. The SVM functions as a method for linearly generating a hyperplane from a dataset into two classes. In Figure 1, the hyperplane is a general term for all dimensions [11]. For example, for a one-dimensional data set, the hyperplane can manifest as a point; if the set is in the form of two dimensions, the hyperplane is a straight line [12].

Figure 1 shows that a pair of parallel hyperplanes can separate two classes. The first boundary plane becomes the boundary of the first class. In contrast, the second boundary plane is the boundary of the second class, so (1) and (2) are obtained where w is the normal plane and b is the position of the plane relative to the coordinate center [11].

$$x_i.w + b \geq +1, \text{if } y_i = +1 \quad (1)$$

$$x_i.w + b \leq -1, \text{if } y_i = -1. \quad (2)$$

As for this algorithm, the most widely used kernel learning includes linear kernels, Gaussian RBF, and polynomials. The SVM algorithm with this kernel finds hyperplanes by data mapping from feature space to higher dimensional kernel space. This way leads to achieving nonlinear separation in kernel space. In addition, the linear kernel can be expressed as (3) [13].

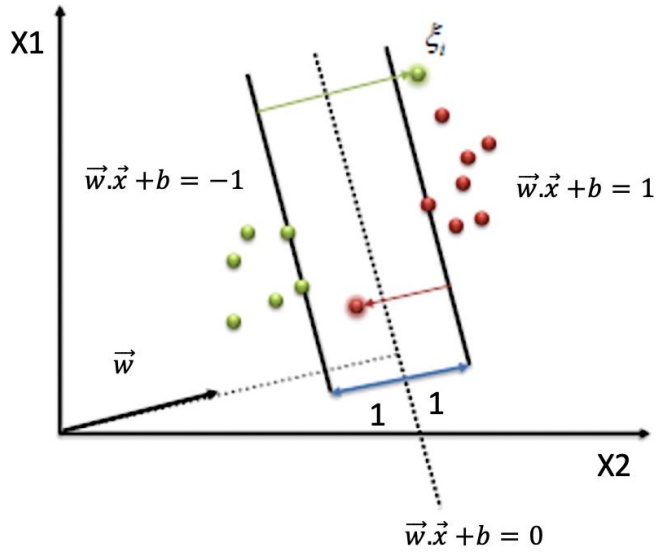


Figure 1. SVM hyperplane.

$$\kappa(x_i, x_j) = x_i^T x_j. \quad (3)$$

An appropriate kernel function, such as the nonlinear Gaussian RBF, can enable the SVM if there is an occurrence of problems which cannot be separated linearly. The kernel equation is shown in (4), where σ signifies the kernels width [13].

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i x_j\|^2}{2\sigma^2}\right). \quad (4)$$

However, in the Gaussian RBF kernel with *the* σ parameter as the kernel width, the SVM will be overfitting in all training instances when the parameter is close to zero. Setting a more significant value to σ may result in underfitting, in which all instances are classified into one class. Because of that, the correct value must be selected for the kernel. The polynomial kernel degrees control higher degrees, allowing more flexible decision limits than linear limits and the flexibility of the classifier width. The kernel equation is shown in (5), where p signifies the degree of the polynomial kernel [13].

$$\kappa(x_i, x_j) = \left(1 + x_i^T x_j\right)^p. \quad (5)$$

Meanwhile, when training data using the SVM and selecting kernel functions, several decisions must be made during the data preparation, among others, by labeling and setting SVM parameters to provide optimal results.

V. ACCURACY

Below is the implementation of the testing stages.

A. CROSS VALIDATION

Cross validation is one of the methods that could be used to search for the validity of machine learning models. This method works by partitioning learning set into k-subsets and thus, have *knn*s. Where, in each fold, as much as (k-1)-subsets are used as training set and the rest is used as the validation set. This procedure is repeated until the entire subset becomes a validation set. It is recommended to use 10 as the best value for k when trying to validate a model, or it is widely known as 10-fold cross validation [4].

B. CONFUSION MATRIX

The confusion matrix test is a size $N \times N$ matrix, where N is the number of classes used in the classification. This matrix

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABLE I
CONFUSION MATRIX 2 × 2

Predicted Values	Actual Values	
	Positive	***Negative***
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True Negative (TN)

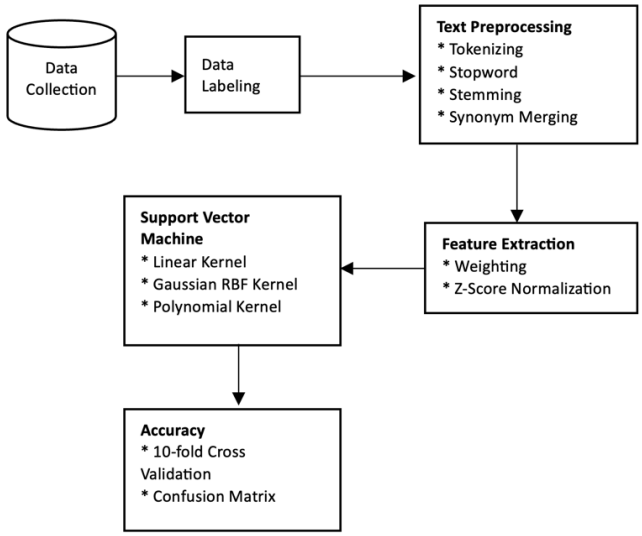


Figure 2. Research flow.

compares the value predicted by the model with the actual value. The confusion matrix used in this study is a 2 × 2 matrix, as shown in Table I [5].

Table I shows that the true positive (TP) value is obtained when the predicted and actual values are positive. On the other hand, the true negative (TN) value is obtained when the predicted value and actual value are negative. Furthermore, the false positive (FP) value is obtained if the predicted value is positive, but the actual value is negative. The false negative (FN) value is obtained if the predicted value is negative, but the actual value is positive. Meanwhile, the accuracy formula in this test can be seen in (6). Based on the equation, the greater the value of TN and TP, the higher the level of accuracy [5].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (6)$$

VI. METHODOLOGY

The research is conducted with the research flow depicted in Figure 2.

A. DATA COLLECTION STAGE AND LABELING STAGE

The data used in this research are Indonesian tweets sourced from Twitter. The use of these data aligns with the research's goal of developing a system that can analyze the sentiments of Indonesian Twitter users regarding the booster vaccine. The aforementioned data were collected from the data source using Python script with "vaksinbooster" as the keyword. The data comprised 681 tweet data with the hashtag "vaccination booster," gathered from 12 January 2022 until 12 April 2022. Then, the step proceeded to the labeling stage, during which the data were assigned a class label based on two classes (positive and negative).

B. TEXT PREPROCESSING STAGE

This stage prepares data so the machine can analyze them easily [10]. There are several stages: tokenizing, removing stopword, and stemming. These steps are carried out to remove noise in tweets, such as URLs and usernames. In addition, the system will also convert nonformal Indonesian words or abbreviation into words that adhere to the Great Dictionary of Indonesian Language (Kamus Besar Bahasa Indonesia, KBBI) and extract words that start with hashtags.

TABLE II
TOKENIZING STAGE PROCESS

Stages	Results
Initial Tweet	DPP Projo mengadakan vaksin booster Covid- gratis untuk rakyat selama lima hari untuk masyarakat DKI Jakarta
Lowercasing	dpp projo mengadakan vaksin booster covid- gratis untuk rakyat selama lima hari untuk masyarakat dki jakarta
Punctuation removal	dpp projo mengadakan vaksin booster covid gratis untuk rakyat selama lima hari untuk masyarakat dki jakarta
Words separation as individual token	"dpp" "projo" "mengadakan" "vaksin" "booster" "covid" "gratis" "untuk" "rakyat" "selama" "lima" "hari" "untuk" "masyarakat" "dki" "jakarta"

TABLE III
STOPWORD STAGE PROCESS

Stages	Results
Before stopword removal	"dpp" "projo" "mengadakan" "vaksin" "booster" "covid" "gratis" "untuk" "rakyat" "selama" "lima" "hari" "untuk" "masyarakat" "dki" "jakarta"
After stopword removal	"dpp" "projo" "mengadakan" "vaksin" "booster" "covid" "gratis" "rakyat" "lima" "masyarakat" "dki" "jakarta"

TABLE IV
STEMMING STAGE PROCESS

Stages	Examples
Token	mengadakan
Affixes	meng–ada-kan
Stemming Results	ada

1) TOKENIZING

The first stage in preprocessing is tokenizing [12]. Several processes are carried out at this stage: lowercasing, punctuation removal, and breaking a sentence into separate words [14]. The process can be seen in Table II.

2) STOPWORD

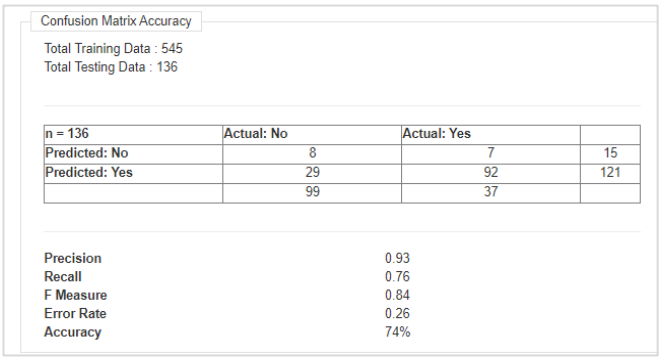
The second stage is the stopword. This stage is carried by removing high-frequency words in the texts that have no special or insignificant meaning. Examples of stopword in Indonesian are "dan," "atau," "sebuah," and "adalah" [15]. This process can be seen in Table III. As seen in Table III, stopword will be deleted once it is identified.

3) STEMMING

The third stage is stemming, a process that eliminates affixes in order to find the root words. This stage aims to reduce the number of words processed in text mining with the purpose of reducing the processing time and minimizing the memory used [16]. The results of the stemming process can be seen in Table IV.

4) SYNONYM MERGING

The final stage in preprocessing is combining synonyms or language forms with similar meanings. At this stage, different words with similar meanings will be merged. Examples of



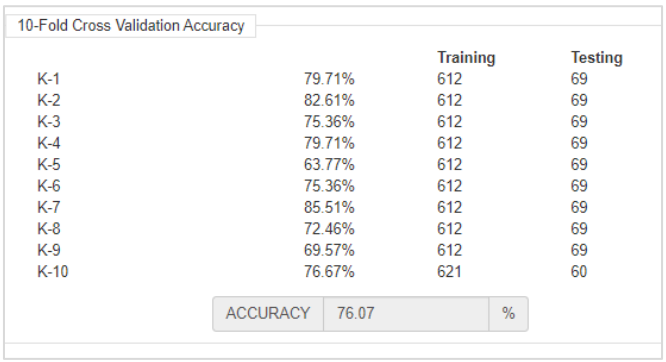
Confusion Matrix Accuracy

Total Training Data : 545
Total Testing Data : 136

n = 136	Actual: No	Actual: Yes	
Predicted: No	8	7	15
Predicted: Yes	29	92	121
	99	37	

Precision	0.93
Recall	0.76
F Measure	0.84
Error Rate	0.26
Accuracy	74%

Figure 3. confusion matrix accuracy with linear kernel.



10-Fold Cross Validation Accuracy

		Training	Testing
K-1	79.71%	612	69
K-2	82.61%	612	69
K-3	75.36%	612	69
K-4	79.71%	612	69
K-5	63.77%	612	69
K-6	75.36%	612	69
K-7	85.51%	612	69
K-8	72.46%	612	69
K-9	69.57%	612	69
K-10	76.67%	621	60
ACCURACY	76.07	%	

Figure 4. Ten-fold cross validation accuracy with linear kernel.

words with the same meaning are "saya" and "aku." This stage aims to minimize the number of words in the system while maintaining the number of frequencies [3].

C. FEATURE EXTRACTION STAGE

Feature extraction is the subsequent stage after the preprocessing stage. Since the textual data (in this case is tweets) are public opinion, the writing structure is often not structured. Therefore, it is necessary to transform textual data into structured data so that the machine learning algorithm can work immediately [17]. There are two processes in this stage, namely weighting and z-score normalization.

1) WEIGHTING

Weighting is a stage that reflects how important a word is in the document. The method used to do weighting in text mining is the term frequency-inverse document frequency (TF-IDF), where the formula can be seen in (7). The weight of the terms in the text will increase as their occurrences rise. However, this increase is also in line with the frequency of terms' occurrences that are included within the research domain of interest [18].

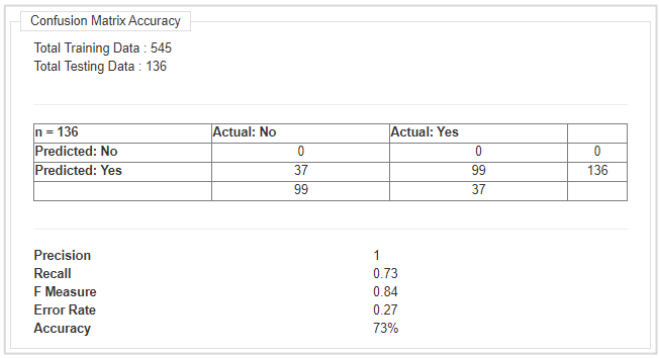
$$W_{t,d} = tf_{t,d} \; x \; idf_t. \tag{7}$$

where $W_{t,d}$ represent the weight, $tf_{t,d}$ represent the term frequency (TF) of the word, and idf_t represent the inverse document frequency (IDF).

2) Z-SCORE NORMALIZATION

The z-score normalization process is a step that is carried out after obtaining the weight value. Due to the significant difference in the range value that will impact the classification problems later, normalization must be carried out [19]. In addition, (8) can be used to perform z-score normalization [20].

$$Z = \frac{x - \bar{x}}{s}. \tag{8}$$



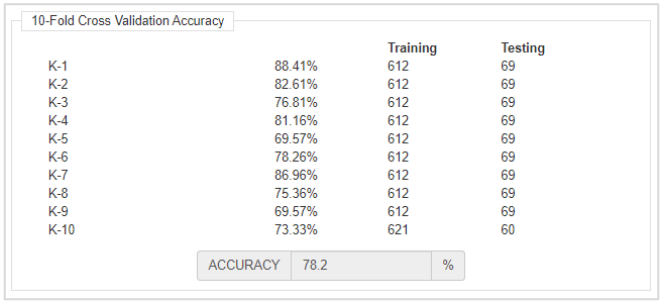
Confusion Matrix Accuracy

Total Training Data : 545
Total Testing Data : 136

n = 136	Actual: No	Actual: Yes	
Predicted: No	0	0	0
Predicted: Yes	37	99	136
	99	37	

Precision	1
Recall	0.73
F Measure	0.84
Error Rate	0.27
Accuracy	73%

Figure 5. Confusion matrix accuracy with the Gaussian RBF kernel.



10-Fold Cross Validation Accuracy

		Training	Testing
K-1	88.41%	612	69
K-2	82.61%	612	69
K-3	76.81%	612	69
K-4	81.16%	612	69
K-5	69.57%	612	69
K-6	78.26%	612	69
K-7	86.96%	612	69
K-8	75.36%	612	69
K-9	69.57%	612	69
K-10	73.33%	621	60
ACCURACY	78.2	%	

Figure 6. 10-fold cross validation accuracy with the Gaussian RBF kernel.

On (8), Z represents the z-score normalization, x represents the value from data, $\bar{x}$ is the mean of the data, and s is the standard deviation.

D. SUPPORT VECTOR MACHINE (SVM) STAGE

The system groups tweets into two clusters, namely positive and negative. A hyperplane groups each tweet. When using SVM algorithm, there are three kernels that could be used to find the best hyperplane, namely the Gaussian RBF kernel, linear kernel, and polynomial kernel [13]. Thus, during the research the three kernels' accuracies were compared with each other, and the best kernel found was used as a result of sentiment analysis based on their accuracy results.

E. ACCURACY STAGE

The system results were tested for accuracy using the 10-fold cross validation and confusion matrix. The test results of the two methods were then compared, and the best results was utilized to test the accuracy of sentiment analysis.

VII. RESULTS AND DISCUSSION

The dataset collected on Twitter with the hashtag "vaksinbooster" is 681 tweets. These data consisted of two classes: negative (0) and positive (1). In addition, the data showed that public sentiment towards the booster vaccine tended to be positive, as indicated by the 554 positive and 127 negative tweets. The dataset was divided into 545 training data and 136 testing data. Furthermore, text preprocessing was applied to the dataset, resulting in an array of significant words from a tweet. Meanwhile, the weight values of each word from the text preprocessing stage were obtained from conducting the feature extraction stage. The weight value from extraction results were afterwards processed as inputs using the SVM algorithm. The SVM algorithm was then applied using three kernels: the linear kernel (see Figure 3 and Figure 4), Gaussian RBF kernel (see Figure 4), and polynomial kernel (see Table V and Table VI).

Figure 3 describes the results of testing the SVM algorithm with a linear kernel on the sentiment analysis of booster vaccine.

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABLE V
CONFUSION MATRIX RESULTS

***n* = 136**	**Actual: No**	**Actual: Yes**	**Total**
Predicted: No	0	0	0
Predicted: Yes	37	99	136
	37	99	

TABLE VI
10-FOLD CROSS VALIDATION RESULTS

Fold	**Accuracy**	**Training**	**Testing**
k-1	88.41%	612	69
k -2	82.61%	612	69
k -3	81.16%	612	69
k -4	81.16%	612	69
k -5	75.36%	612	69
k -6	78.26%	612	69
k -7	86.96%	612	69
k -8	75.36%	612	69
k -9	69.57%	612	69
k -10	73.33%	621	60

Testing with the confusion matrix method yielded positive predictive data following positive facts in as many as 92 and negative predictive data following negative reality in as many as eight positive predictive data. Still, negative reality found as many as 29 data and negative predictive data, but positive reality found as many as 7 data. Therefore, the results of the accuracy level with the confusion matrix test method were 74%. On the other hand, as seen in Figure 4, testing with the 10-fold cross validation method yielded a higher level of accuracy than the confusion matrix, which was 76.07%.

Figure 5 shows the results of testing the SVM algorithm with the Gaussian RBF kernel for sentiment analysis of booster vaccines. Testing with the confusion matrix method yielded positive predictive data corresponding to positive reality as much as 99 and negative predictive data corresponding to negative reality as much as 0. Meanwhile, positive predictive data with negative reality were found in 37 data, and negative predictive data with positive reality were found in 0 data. Therefore, the results of the confusion matrix testing yielded an accuracy level of 73%. On the other hand, the utilization of the 10-fold cross validation method yielded a higher level of accuracy than the confusion matrix, which was 78.2% (see Figure 6).

In 2019, research on the human development index (HDI) classification using the SVM with the Gaussian RBF kernel achieved an accuracy of 98.1%, which was higher than the application of the linear kernel, which yielded an accuracy of 95.1% [21].

Table V and Table VI describes the results of testing the SVM algorithm with polynomial kernel on the sentiment of booster vaccine analysis. As shown on Table V, testing with the confusion matrix method obtained the same results as those achieved by the SVM algorithm with Gaussian RBF kernel, which had an accuracy rate of 73%. However, as depicted on Table VI, testing using the 10-fold cross validation method yielded a higher level of accuracy than the confusion matrix, which was averaged at 79.22%.

Thus, the application of the SVM algorithm with the best kernel was found in the polynomial kernel, exhibiting the highest level of accuracy through the implementation of using the 10-fold cross validation test method, which is 79.22%.

In 2020, research was conducted to investigate the sentiment analysis of OVO services on Twitter using the SVM algorithm. The research found that using a polynomial kernel in the SVM process had a high accuracy rate of 89.09% [22]. In addition, research on sentiment analysis of the large-scale social restrictions (p*embatasan sosial berskala besar*, PSBB*)* policies on Twitter found that the SVM algorithm with a polynomial kernel had an accuracy rate of 94.55% [23]. Based on the results of this study, the SVM algorithm with a polynomial kernel is the best method for analyzing booster vaccine sentiment on Twitter social media.

VIII. CONCLUSION

In Indonesia, the COVID-19 vaccination entails the administration of two compulsory doses, with the most recent dosage being the booster vaccination. However, the booster vaccines have rose various sentiments among the Indonesians. Therefore, the researcher proposes the application of machine learning technology to develop a system that can analyze the sentiments of the Indonesian towards the booster vaccine using the SVM algorithm. This study gathered 681 Indonesian Twitter data, specifically focusing on the hashtag "vaksinbooster." The data were subsequently categorized into negative (0) and positive (1). In addition, the dataset was divided into training data consisting of 545 data and a testing data consisting of 136 data. The data shows that public sentiment toward the booster vaccine tends to be positive, as evidenced by the positive tweets of 554 data and the negative tweets of 127.

The system testing using the 10-fold cross validation found that the SVM algorithm using the linear kernel achieved an accuracy rate of 76.07%. At the same time, the Gaussian RBF kernel obtained an accuracy rate of 78.2%. Then, the polynomial kernel yielded an accuracy rate of 79.22%. The system testing using the confusion matrix method found that the SVM algorithm using the linear kernel resulted in an accuracy rate of 74%. Meanwhile, the Gaussian RBF kernel obtained an accuracy rate of 73%. The polynomial kernel achieved an accuracy rate of 73%. Hence, the best kernel for the application of the SVM algorithm is the polynomial kernel, which achieves the highest level of accuracy when tested with the 10-fold cross validation.

CONFLICT OF INTEREST

All authors state that there is no conflict of interest in this study.

AUTHOR CONTRIBUTION

Conceptualization, Dionisia Bhisetya Rarasati; methodology, Dionisia Bhisetya Rarasati; writing—original draft preparation, Angelina Pramana Thenata and Afiyah Salsabila Arief; writing—review and editing, Dionisia Bhisetya Rarasati, Angelina Pramana Thenata, and Afiyah Salsabila Arief.

ACKNOWLEDGMENT

Thanks to the Ministry of Research, Technology, and Higher Education who has supported this research through the Beginner Lecturer Research Grant for the 2022 fiscal year with the contract number of 155/E5/PG.02.00.PT/2022.

REFERENCES

[1] (2022) "Peta Sebaran COVID-19," [Online], https://covid19.go.id/peta-sebaran#, access date: 14-Mar-2022.

[2] D.A. Agustina, S. Subanti, and E. Zukhronah, "Implementasi Text Mining pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine," *Indones. J. Appl. Stat.*, Vol. 3, No. 2, pp. 109–122, Nov. 2020, doi: 10.13057/ijas.v3i2.44337.

[3] D.B. Rarasati, "A Grouping of Song-Lyric Themes Using K-Means Clustering," *JISA (J. Inform., Sains)*, Vol. 3, No. 2, pp. 38–41, Dec. 2020, doi: 10.31326/jisa.v3i2.658.

[4] D. Berrar, "Cross-Validation," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1, S. Ranganathan, K. Nakai, C. Schönbach, Eds., Amsterdam, Netherlands: Elsevier, 2019, pp. 542–545, doi: 10.1016/B978-0-12-809633-8.20349-X.

[5] Y. Sari, P.B. Prakoso, and A.R. Baskara, "Road Crack Detection using Support Vector Machine (SVM) and OTSU Algorithm," *2019 6th Int. Conf. Electr. Veh. Technol. (ICEVT)*, 2019, pp. 349–354, doi: 10.1109/ICEVT48285.2019.8993969.

[6] B. Haryanto *et al*., "Facebook Analysis of Community Sentiment on 2019 Indonesian Presidential Candidates From Facebook Opinion Data," *Procedia Comput. Sci.*, Vol. 161, pp. 715–722, 2019, doi: 10.1016/j.procs.2019.11.175.

[7] M. Syarifuddin, "Analisis Sentimen Opini Publik Mengenai COVID-19 pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," *Inti Nusa Mandiri*, Vol. 15, No. 1, pp. 23–28, Aug. 2020, doi: 10.33480/inti.v15i1.1347.

[8] R. Umar, I. Riadi, and Purwono, "Perbandingan Metode SVM, RF dan SGD untuk Penentuan Model Klasifikasi Kinerja Programmer pada Aktivitas Media Sosial," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 4, No. 2, pp. 329–335, Apr. 2020, doi: 10.29207/resti.v4i2.1770.

[9] A. Roihan, P.A. Sunarya, and A.S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review Paper," *IJCIT (Indones. J. Comput. Inf. Technol.*, Vol. 5, No. 1, pp. 75–82, May 2020, doi:10.31294/ijcit.v5i1.7951.

[10] A.P. Thenata, "Text Mining Literature Review on Indonesian Social Media," *J. Eduk., Penelit. Inform.*, Vol. 7, No. 2, pp. 226–232, Aug. 2021, doi: 10.26418/jp.v7i2.47975.

[11] F.S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist., Teknol. Inf.)*, Vol. 1, No. 1, pp. 19–25, Apr. 2017, doi: 10.29207/resti.v1i1.11.

[12] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A Comparative Evaluation of Pre-Processing Techniques and Their Interactions for Twitter Sentiment Analysis," *Expert Syst. Appl.*, Vol. 110, pp. 298–310, Nov. 2018, doi: 10.1016/j.eswa.2018.06.022.

[13] C. Savas and F. Dovis, "The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines," *Sens.*, Vol. 19, No. 23, pp. 1–16, Nov. 2019, doi: 10.3390/s19235219.

[14] V.Y. Radygin *et al*., "Application of Text Mining Technologies in Russian language for Solving the Problems of Primary Financial Monitoring for Solving the Problems of Primary Financial Monitoring," *Procedia Comput. Sci.*, Vol. 190, pp. 678–683, 2021, doi: 10.1016/j.procs.2021.06.078.

[15] H. Najjichah, A. Syukur, and H. Subagyo, "Pengaruh Text Preprocessing dan Kombinasinya pada Peringkas Dokumen Otomatis Teks Berbahasa Indonesia," *J. Teknol. Inf. Cyberku*, Vol. 15, No. 1, pp. 1–11, Jan. 2019.

[16] D. Sebastian, "Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari Beberapa E-marketplace," *J. Tek. Inform., Sist. Inf.*, Vol. 5, No. 1, pp. 51–61, Apr. 2019, doi: 10.28932/jutisi.v5i1.1581.

[17] H. Liu, P. Burnap, W. Alorainy, and M. L. Williams, "A Fuzzy Approach to Text Classification with Two-Stage Training for Ambiguous Instances," *IEEE Trans. Comput. Soc. Syst.*, Vol. 6, No. 2, pp. 227–240, Apr. 2019, doi: 10.1109/TCSS.2019.2892037.

[18] I. Yahav, O. Shehory, and D. Schwartz, "Comments Mining With TF-IDF: The Inherent Bias and Its Removal," *IEEE Trans. Knowl., Data Eng.*, Vol. 31, No. 3, pp. 437–450, Mar. 2019, doi: 10.1109/TKDE.2018.2840127.

[19] Henderi, W. Tri, and R. Efana, "Comparison of Min-Max Normalization and Z-Score Normalization in the K-Nearest Neighbor (KNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS Int. J. Inform., Inf. Syst.*, Vol. 4, No. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[20] M.A. Imron and B. Prasetiyo, "Improving Algorithm Accuracy K-Nearest Neighbor Using Z-Score Normalization and Particle Swarm Optimization to Predict Customer Churn," *J. Soft Comput. Explor.*, Vol. 1, No. 1, pp. 56–62, Sep. 2020, doi: 10.52465/joscex.v1i1.7.

[21] H. Al Azies, D. Trishnanti, and E. Mustikawati P.H, "Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI)," *IPTEK J. Proc. Ser.*, No. 6, pp. 53–57, Nov. 2019, doi: 10.12962/j23546026.y2019i6.6394.

[22] F. Romadoni, Y. Umaidah, and B.N. Sari, "Text Mining untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine," *J. Sisfokom (Sist. Inf., Komput.)*, Vol. 9, No. 2, pp. 247–253, Jul. 2020, doi: 10.32736/sisfokom.v9i2.903.

[23] H.P.P. Zuriel and A. Fahrurozi, "Implementasi Algoritma Klasifikasi Support Vector Machine untuk Analisa Sentimen Pengguna Twitter Terhadap Kebijakan PABB," *J. Ilm. Inform. Komput.*, Vol. 26, No. 2, pp. 149–162, Aug. 2021, doi: 10.35760/ik.2021.v26i2.4289.