ISSN 2714-9714

Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter

Khafifah Munawaroh [1,*], Alamsyah [1]

[1] Departement of Computer Science, Faculty of Mathematics and natural Sciences, Semarang State University, Semarang, Indonesia
*Corresponding author: khofi_muna@students.unnes.ac.id

ARTICLE INFO

Article history
Received 4 October 2022
Revised 15 October 2022
Accepted 26 October 2022

Keywords
Text Mining
Sentiment Analysis
Vaccine Covid-19
Classification

ABSTRACT

The emergence of the COVID-19 virus in 2020 has created a new breakthrough in the form of a vaccine as a solution to slow the spread of the virus. However, the COVID-19 vaccine is considered controversial and invites many people to express their views on various media, one of which is social media Twitter. Using Twitter data on the COVID-19 vaccine, sentiment analysis can be performed. Sentiment analysis aims to evaluate whether the tweet contains a positive sentence or sentiment. In this study, the analysis of sentiments on the COVID-19 vaccine on social media Twitter was carried out using the Support Vector Machine (SVM), Naïve Bayes, and k-Nearest Neighbor (KNN) algorithms. SVM has the advantage of being able to identify hyperplanes that maximize the margins between different sentiments. Meanwhile Naïve Bayes is an algorithm that is simple, fast and produces maximum accuracy with training. The KNN algorithm was chosen because it is superior to noise. The performance of the three classification algorithms will be compared, so that it can be seen which algorithm is better in classifying text mining. Sentiment classification results in this study consist of positive sentiment and sentiment classes. The resulting accuracy value will be a benchmark for finding the best test model in the case of sentiment classification. Based on ten tests, the final result of accuracy and best performance using the SVM algorithm with an accuracy value of 96.3% is obtained. Meanwhile, the Naïve Bayes and KNN algorithms have an accuracy of 94% and 91%, respectively. The high accuracy results are supported by the feature extraction TF-IDF the TextBlob library.

1 Introduction

The World Health Organization (WHO) determined on March 11, 2020, that the COVID-19 virus is a global pandemic (Sohrabi et al., 2020). WHO said that more than fifty-two million people were confirmed positive for COVID-19 and in the second week of November 2020 it was reported that 1.2 million people had died (Alamsyah et al., 2021). Given the rapid spread of COVID-19 and the consequences of efforts that arise if the problem is not immediately addressed, one way to slow down the process of spreading the virus is to make a vaccine (Fitriana et al., 2021).

There is a controversy with the emergence of the COVID-19 vaccine, which has led some people to turn to the media to express their opinions and views. According to global digital statistics "Digital, Social & Mobile in 2019" that social media users reached one hundred and fifty million in 2019. Twitter is the social media with the most active users in Indonesia, accounting for fifty-two percent of total social media users (Fitriana et al., 2021). Social media Twitter is a means to obtain information that can be used for sentiment analysis, by dividing public opinion about the COVID-19 vaccine into two classes of sentiment, namely negative and positive. Sentiment analysis is a way to classify people's emotional levels as neutral, positive, or negative (Mubarok et al., 2017). Automatically tweets be retrieved by the system and classify an evaluation of tweets that contain neutral, positive, or negative sentences (Tripathy et al., 2016).

In previous observations using the Support Vector Machine (SVM) method, Windasari, Uzzi, and Satoto conducted research on public sentiment towards Gojek's online transportation on Twitter. In the conference, articles accumulated into 1000 tweets positive and tweets negative. These results are known with the help of the Support Vector Machine (SVM) to classify the existing data. After the analysis, the results obtained are 86% accuracy scores and 14% error prediction scores. For tweets positive tweets negative (Windasari et al., 2017).

By looking at the problems above, the focus of this research is to compare the classification algorithm using Feature Extraction TF-IDF TextBlob Library in sentiment tweet by taking the COVID-19 vaccine limit. SVM, Naïve Bayes (NB), and k-Nearest Neighbor (KNN) are used as classification algorithms in the research to be carried out. SVM has the advantage of being able to identify separate hyperplanes that maximize the margin between the two different classes. Meanwhile, Naïve Bayes is an algorithm that is simple, fast and produces maximum accuracy with little training data. The KNN algorithm was chosen because the algorithm is superior to noise data. The performance of the three classification algorithms will be compared, so that it can be seen which algorithm is better in classifying text mining. The resulting accuracy value will be a benchmark for finding the best test model in the case of sentiment classification (Ashari et al., 2016). The purpose of this research is to help the government and the public to find out the public's responses or concerns about the COVID-19 vaccine, and also as material for the government's evaluation to determine further strategies related to education and socialization about COVID-19 vaccination to the public.

2 The Proposed Algorithm

2.1 Sentiment Analysis

Sentiment analysis is a system of determining sentiment in a document or sentence and classifying the polarity of the text so that it can be categorized as positive, negative, or neutral class sentiment. In sentiment evaluation, statistical mining is carried out to research, process, and extract textual data in an entity such as a service, product, person, phenomenon, or subject. The analytical step includes evaluating texts, forums, tweets, or blogs using pre-processed data including tokenization, stopwords, deletion, root detection, sentiment identity, and the sentiment classification process (Rasenda et al., 2020).

Sentiment evaluation is generally carried out in 3 different stages, such as sentence level, document level, and aspect level. Document level has the goal of grouping all documents into positive or negative sentiment classes. Sentence level is based on the polarity of each sentence (Tripathy et al., 2015). Meanwhile, aspect level is to classify the opinion of a document as an opinion with good or bad sentiments based on several large documents with the same topic. Sentence level in sentiment analysis, classifies the sentiment in each sentence by identifying whether the sentence is a positive or negative opinion (Medhat et al., 2014).

2.2 Support Vector Machine

Support Vector Machine (SVM) is one of several past methods that is still being used by several researchers in big data (Larasati et al., 2019). SVM is a hyperplane that functions to distinguish two classes in the input space (Athoillah, 2018). SVM aims to find out future data by using existing data based on special characteristics and to find a hyperplane to separate data based on two possible categories of variables, namely positive and negative (Cortes & Vapnik, 1995).

2.3 Naïve Bayes

Algorithm Naive Bayes is a classification algorithm with simple probability by applying Bayes' with the assumption of high independence. Algorithm Naïve Bayes is based on the number of data sets used, so we need a method with high classification performance and high accuracy. The advantage of using Naive Bayes is that this algorithm requires not a lot of training data to determine the estimated parameters needed during the classification process (Kao & Poteet, 2007). In implementing classifier for sentiment analysis, multinomial Naïve Bayes due to the general use of this method for the text grouping process and has the aim of creating two simple independent assumptions, namely (Krishnaiah et al., 2013):

- Assumption Bag-of-words, the condition about the word position is not important from an assumption.
- The assumption of conditional independence is the assumption that the probability of a feature is independent in a class.

2.4 k-Nearest Neighbor

k-Nearest Neighbor (KNN) is one of several ways of grouping in machine learning. Algorithm KNN aims to group objects into one of the predefined classes from sample groups that have been created by machine learning. The KNN algorithm is a supervised algorithm that can classify data based on the level of proximity of the data to other data sets. This algorithm includes the lazy learning, meaning that the search process is done by classifying k features from the closest training data with (similar) features from new data or test data (Mustakim & Oktaviani F, 2016).

2.5 Feature Extraction TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is used as a word weighting method to find information in text mining. The number of words that appear in the document is set as the TF-IDF value, of course also balanced by the frequency of words in the word set, used as a determinant of words that occur frequently. The TF-IDF calculation aims to give word weighting values determined from the number of words that appear in a document (Dadgar et al., 2016). Term Frequency or TF means the repetition of the number of words in one sentence. Inverse Document Frequency or IDF is a dimension of the amount of information contained in a word, the intensity of the occurrence of words in all document texts is often or rarely (Hakim et al., 2014).

2.6 Library TextBlob

TextBlob is one of several libraries in Python that are used in textual data processing, such as tokenization, sentiment analysis, and the process of translating a language into many common languages around the world (Pedregosa et al., 2011). This library provides a simple API for diving into Natural Language Processing (NLP) (Bose et al., 2020). This study uses TextBlob for sentiment analysis. The sentiment analysis design in TextBlob is only available in English, so users need to translate it into English to use TextBlob.

3 Method

In this study, by comparing three classification algorithms in the sentiment analysis process for the COVID-19 vaccine on Twitter, the algorithm with the highest accuracy results was used as a reference in the sentiment analysis process. Therefore, prior to the algorithm design process, a research design is made so that research objectives can be achieved and scientifically accounted for. Research design is described in flowchart which can be seen in Figure 1.

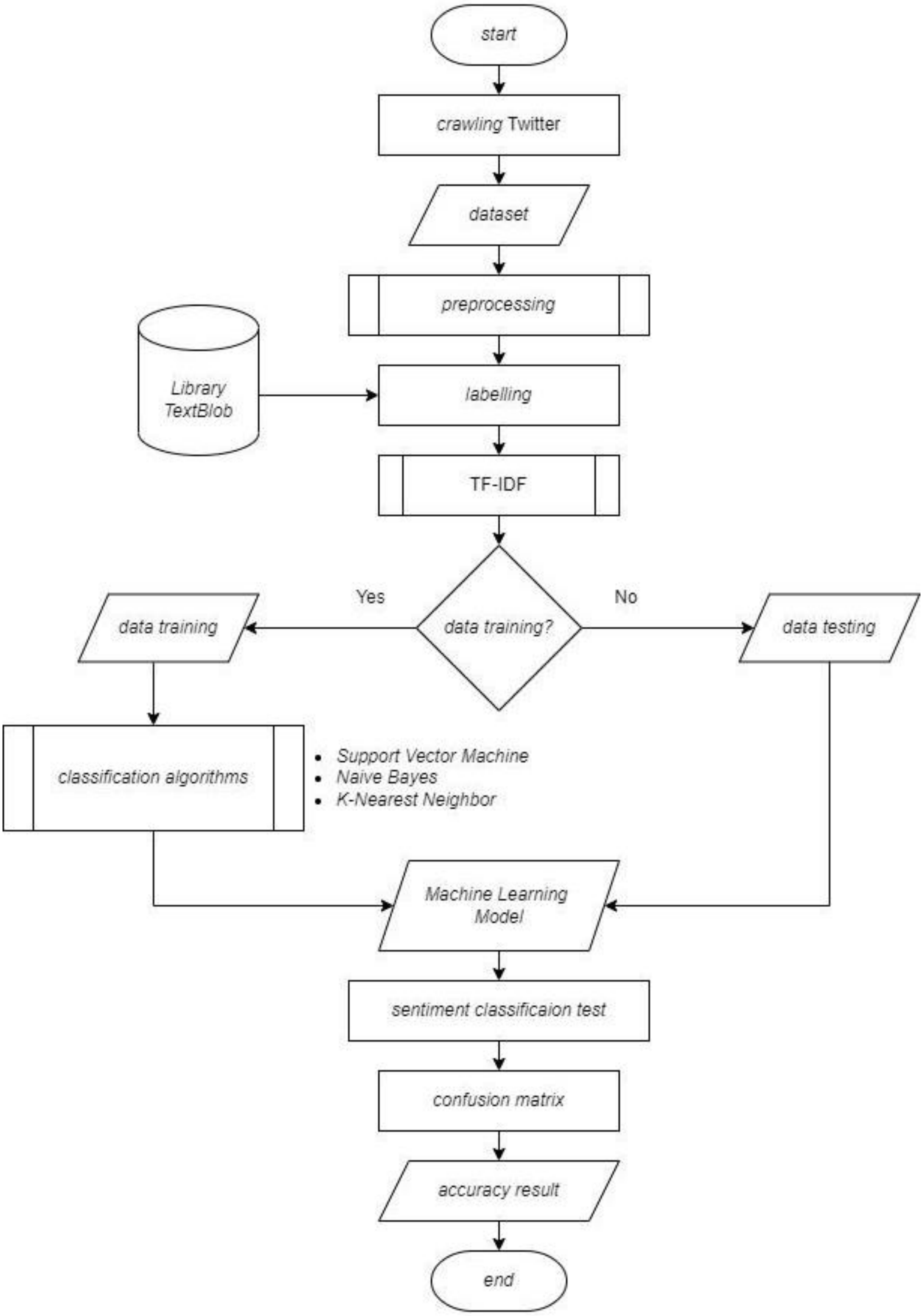


Figure 1. Flowchart Research Design

There are three algorithms that will be applied to sentiment analysis of tweets COVID-19 vaccine Feature Extraction TF-IDF and labeling using TextBlob Library algorithm SVM, Naïve Bayes and KNN. The research on the comparison of three classification algorithms in the sentiment analysis process for the COVID-19 vaccine on Twitter was carried out in several stages. There are six stages in this research, it is the preprocessing, labeling using TextBlob Library, the application of Feature Extraction, the TF-IDF training and testing, the creation of machine learning models, as well as the classification and accuracy testing stages.

4 Results and Discussion

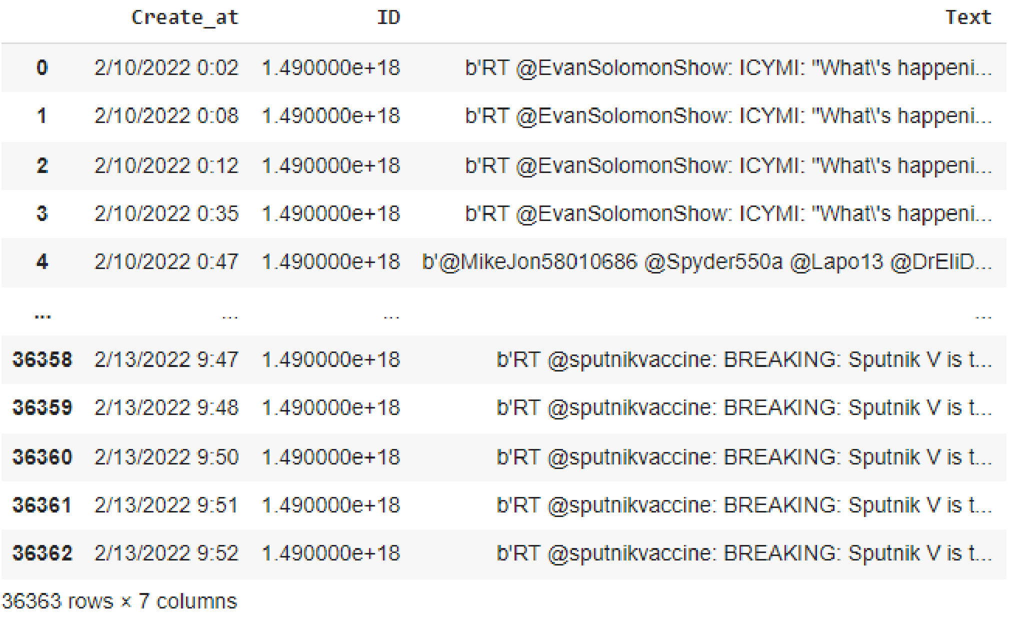
This section is divided into two parts, results and discussion. The results are a description of the data and findings obtained using the methods and procedures described in the data collection method. The discussion is an explanation of the results that answer research questions more comprehensively.

4.1 The results

The results of this study compare the accuracy results of three classification algorithms, namely SVM, Naïve Bayes, and KNN in the process of analyzing COVID-19 vaccine sentiment on Twitter. The results of this study are as follows.

4.1.1 *Results of the Crawling Process*

Data used in this study was taken from the crawling using the Twitter API with Python programming with the keyword #vaccineCOVID-19 in the period 10-13 February 2022. In the crawling , 35,644 data were generated, and the text was retrieved on the process is a tweet that uses English. Sample data from the crawling can be seen in Figure 2.



```
            Create_at            ID                                             Text
0      2/10/2022 0:02  1.490000e+18     b'RT @EvanSolomonShow: ICYMI: "What\'s happeni...
1      2/10/2022 0:08  1.490000e+18     b'RT @EvanSolomonShow: ICYMI: "What\'s happeni...
2      2/10/2022 0:12  1.490000e+18     b'RT @EvanSolomonShow: ICYMI: "What\'s happeni...
3      2/10/2022 0:35  1.490000e+18     b'RT @EvanSolomonShow: ICYMI: "What\'s happeni...
4      2/10/2022 0:47  1.490000e+18  b'@MikeJon58010686 @Spyder550a @Lapo13 @DrEliD...
...               ...           ...                                              ...
36358  2/13/2022 9:47  1.490000e+18       b'RT @sputnikvaccine: BREAKING: Sputnik V is t...
36359  2/13/2022 9:48  1.490000e+18       b'RT @sputnikvaccine: BREAKING: Sputnik V is t...
36360  2/13/2022 9:50  1.490000e+18       b'RT @sputnikvaccine: BREAKING: Sputnik V is t...
36361  2/13/2022 9:51  1.490000e+18       b'RT @sputnikvaccine: BREAKING: Sputnik V is t...
36362  2/13/2022 9:52  1.490000e+18       b'RT @sputnikvaccine: BREAKING: Sputnik V is t...

36363 rows × 7 columns
```

Figure 2. Crawling

4.1.2 *Results Preprocessing*

This stage aims to align the words, remove characters such as numbers, symbols, punctuation marks, etc., and remove unnecessary words so that the data becomes more structured. Stages preprocessing that will be carried out in this research are case folding, cleansing, tokenization, and stopword removal.

4.1.2.1 Case Folding

In case folding, the process of converting uppercase letters to lowercase letters is carried out. This is done so that uppercase and lowercase letters are not detected choosing different meanings. The results of tweets before and after going through the case folding can be seen in Table 1.

Table 1. Case Folding

Input *tweet*	Results *case folding*
b"@zerocovidzoe Yepp it's great to to see a government fighting its own citizens who just want their freedom back and\xe2\x80\xa6 https://t.co/ZlWbg2J2me"	b"@zerocovidzoe yepp it's great to to see a government fighting its own citizens who just want their freedom back and\xe2\x80\xa6 https://t.co/zlwbg2j2me"

b'@JanBenninkCom Nice vaccin!'	b'@janbenninkcom nice vaccin!'
b'This is stunning. Sad but stunning. Holding Canadians hostage at 90% vaccination rate so you can COERCE the few rem\xe2\x80\xa6 https://t.co/btZnJ0S85R'	b'this is stunning. sad but stunning. holding canadians hostage at 90% vaccination rate so you can coerce the few rem\xe2\x80\xa6 https://t.co/btznj0s85r'

4.1.2.2 *Cleansing*

The cleansing stage is carried out to remove punctuation marks, numbers, symbols and other characters so that the process later analysis is easier and does not mix with other characters that are not text. The results of tweets before and after going through the cleansing can be seen in Table 2.

Table 2. Cleansing

Input *tweet*	Results *cleansing*
b"@zerocovidzoe Yepp it's great to to see a government fighting its own citizens who just want their freedom back and\xe2\x80\xa6 https://t.co/ZlWbg2J2me"	yepp its great to to see a government fighting its own citizens who just want their freedom ack and
b"@1Think4yourself @TropicalVertic1 That's not true at all. Furthermore, myocarditis is a much more common complicati\xe2\x80\xa6 https://t.co/Xaf8tlhs2v"	thats not true at all furthermore myocarditis is a much more common complicati
b'@BelovedAmanda0 @jm131995 @iruntoyouj right? take a flight, vaccination card, a mask, 10 days of quarantine, go to\xe2\x80\xa6 https://t.co/MppGL9BHd0'	right take a flight vaccination card a mask days of quarantine go to

4.1.2.3 *Tokenization and Stopword Removal*

The tokenization is used to get word pieces that have value in the preparation of the document matrix in the next process. Meanwhile, in stopword removal, words that have no effect but often appear in tweets. Package used for the tokenization stage and the stopword removal is NLTK. The results of tweets before and after going through the tokenization stage and the stopword removal can be seen in Table 3.

Table 3. Tokenization and Stopword

Input *tweet*	Results Tokenization and Stopword
b'Or if you get your first vaccine dose between July 28, 2021 and February 28, 2022 at a New York City-run vaccine lo\xe2\x80\xa6 https://t.co/MKJ9SSazSp'	['get', 'first', 'vaccine', 'dose', 'between', 'july', 'february', 'new', 'york', 'cityrun', 'vaccine']
b'@blueswampwolf @kcrehabguy @APTAtweets If vaccination prevented transmission, you\xe2\x80\x99d have a point, but it\xe2\x80\x99s clear	['vaccination', 'prevented', 'transmission', 'point', 'clear']

th\xe2\x80\xa6 https://t.co/0AC1jdUZ2v'	
b'@Surelyyouareki2 @Kitcatclaws Not yet and I tried 3 times with the latest vaccination.'	['yet', 'tried', 'times', 'latest', 'vaccination']

1.

4.1.3 *TextBlob Library Labeling Results Sentiment*

Class labeling is usually divided into three classes, called positive, negative, and neutral classes. However, in this study only two classes of sentiment were used, positive and negative. In TextBlob library, the labeling process is conducted by determining the subjectivity and polarity for each tweet. Labeling indicator uses TextBlob library, which is based on the polarity a tweet, where < 0 is a negative class, = 0 is a neutral class, and > 0 is a positive class. Because this study did not use a neutral class, the normalization by taking 2500 data samples in each positive class. and negative. After the labeling process, it is continued with the normalization to get a sample of positive sentiment class data and negative with a target of 5000 data. The results of labeling sentiment classes using the TextBlob library can be seen in Table 4.

Table 4. Results of Labeling TextBlob Library

Tweet	Subjectivity	Polarity	Sentiment
new many think next logical step protect airline passengers covid compulsory testing vaccination	0,301136	0,221591	Positive
he makes money off if forced vaccination	0,2	-0,3	Negative
evidence whether vaccination status matters given governments pretty much normalxe	0.6	0.225	Positive
unvaccinated also doesnxe mean covidxe complicated nat stxe	0.84375	-0.40625	Negative

2.

4.1.4 *Results Feature Extraction TF-IDF*

Data Tweet that has gone through the preprocessing which is still text form will then be converted into vector form using the TF-IDF technique. Numerical data obtained from the word weighting process can be used for classification analysis. The pseudocode for the feature extraction TF-IDF.

```
TF-IDF
-----------------------------------------------------------
Deklarasi:
Doc[tweet]: datasets tweet after normalisasi
Row[tweet]: total documents (row of datasets)
-----------------------------------------------------------
Algoritma:
For i=1 to Doc[tweet] do
     Count_words = total words in Doc[tweet]
     For j=1 to Doc[tweet](i) do
          Count_words(j) = total words in Doc[tweet](i)
          Weight_words =
          count_words(j) x log(row[tweet]/count_words)
     End for
End for
```

Pseudocode feature extraction principle is to give weight to each word in the dataset according to the feature extraction TF-IDF. The results of word weighting using the feature extraction TF-IDF can be seen in Table 5.

Table 5. Results of Feature Extraction TF-IDF

TF-IDF
{'covid': 0.09919466866827557, 'antivaccine': 0.5054833109508403, 'protesters': 0.3573840138852575, 'force': 0.3774057906747702, 'cancellation': 0.4173962106198401, 'act': 0.379513404073856, 'lifeline': 0.41070302777448403, 'charity': 0.41397684426823755, 'ook': 0.3733562727143532, 'fair': 0.39884528317238727, 'police': 0.40164279426245447, 'investigate': 0.4640220416637106, 'alleged': 0.48615143185230913}
{'get': 0.3222691215277407, 'real': 0.5887019296164484, 'trustful': 0.8693384456506991, 'information': 0.5801507028235453, 'science': 0.6134956575402495, 'well': 0.4990935000756765, 'practicing': 0.8693384456506991, 'medical': 0.5162213533898162, 'doctors': 0.6702540601809153}
{'nice': 2.9390679309004892, 'vaccin': 1.319728704818299}

4.1.5 *The classification*

Data Tweet that has passed the Feature Extraction TF-IDF data split using a ratio of 80:20. Furthermore, the classification stage is carried out using three algorithms, it is SVM, Naïve Bayes, and KNN. The following is an example of a classification calculation using the SVM algorithm, based on the results of the confusion matrix data testing can be seen in Table 6.

Table 6. Confusion Matrix Algorithm SVM

Two Class Classification		*Predicted Class*		Total
		1	0	
Actual Class	1	481(*TP*)	19(*FN*)	500
	0	25(*FP*)	475(*TN*)	500
Total		506	494	1000

Table confusion matrix above, the SVM algorithm classifies 481 positive data predicted correct, 19 negative data predicted wrong, 25 positive data which predicted wrong, and 475 negative data predicted correctly. This shows that the SVM algorithm can classify tweets into positive sentiments and negative correctly as many as 956 tweets out of 1000 tweets. Based on Table 6 the confusion matrix of the SVM algorithm above produces the following level of determination.

- $accuracy = \frac{TP+TN}{P+N} x\ 100 = \frac{481+475}{506+494}\ x\ 100$
  $= \frac{956}{1000}\ x\ 100 = 95{,}6\ \%$
- $precision = \frac{TP}{TP+FP} x\ 100 = \frac{481}{481+25}\ x\ 100$
  $= \frac{481}{506}\ x\ 100 = 95{,}1\ \%$
- $recall = \frac{TP}{TP+FN} x\ 100 = \frac{481}{481+19}\ x\ 100$
  $= \frac{481}{500}\ x\ 100 = 96{,}2\ \%$

Based on the results of the evaluation of the calculations above, the accuracy of the SVM algorithm is 95.6%. The precision to determine the level of model accuracy in the classification process is 95.1%. While the recall obtained is to measure the completeness of the overall data which is positive by 96.2%.

In this classification process, 10 tests were carried out. From 10 times of testing, the highest accuracy value will be taken to be used as the final result of this research. The results of 10 tests of each classification algorithm can be seen in Table 7 and Figure 3.

Table 7. Results of 10 Tests

Experimental	Algorithm Classification Algorithm		
	SVM	*Naïve Bayes*	KNN
1	96	94	90,1
2	96	91,2	90,7
3	94,5	92,2	88,2
4	95,1	92,6	91
5	95,1	91,7	89,7
6	94,6	91,2	89,9
7	96	92,4	89,1
8	96,3	93,3	88,8
9	95,2	93,6	89,6
10	94,9	92,3	88,1
Average	95,37%	92,45%	89,52%

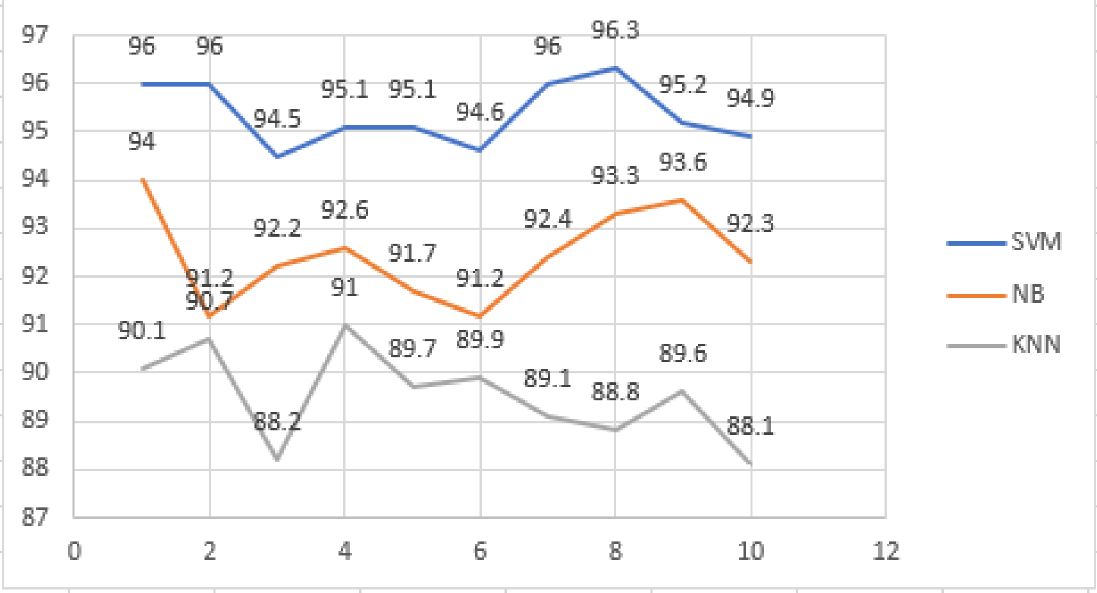


Figure 3. Graph of 10 Times Classification Algorithm Testing

4.2 Discussion

In this study, a comparison of the results of accuracy using the SVM, Naïve Bayes, and KNN algorithm preprocessing, labeling using the TextBlob library, normalization , and TF-IDF. The data used is dataset the on keyword #vaccineCOVID-19 Twitter as many as 35,644 tweet form of *.csv with one main column, namely text, and two supporting columns, namely created_at and id.

Based on the research conducted, the best accuracy and performance on the SVM algorithm were obtained after testing 10 times. In this test, the highest accuracy value is taken for each algorithm to be used as the final result. The results of 10 tests of each classification algorithm can be seen in Table 8 and Figure 4.

Table 8. Final Results Comparison of Accuracy

	Classification Algorithm		
	SVM	*Naïve Bayes*	KNN
Accuracy	96,3%	94%	91%

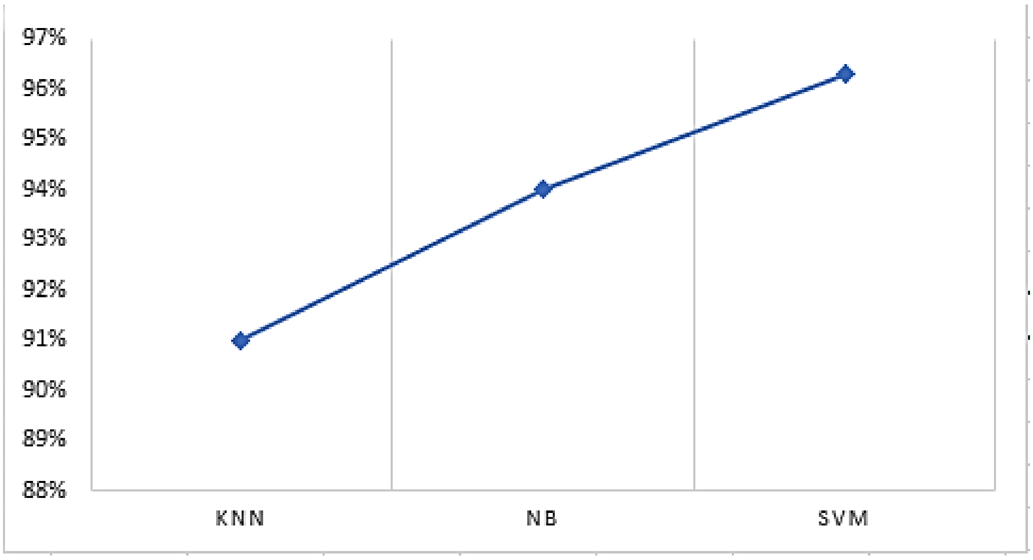


Figure 4. Graph of Accuracy Comparison Results

Based on the research that has been done, the comparison of the three algorithms shows that the final result of the accuracy of the SVM algorithm is higher than the Naïve Bayes and KNN algorithms, which is 96.3%. Meanwhile, in the Naïve Bayes and KNN algorithms, the final results obtained accuracy values of 94% and 91%, respectively. To find out that the method used in this study is better than the previous method, a comparison of the accuracy results with previous studies using datasets with the same keywords was carried out. The comparison results are shown in Table 9 and Figure 5.

Table 9. Comparison of Accuracy with Previous Research

Authors	*Dataset*	*Feature Extraction*	Algorithm	Result
Aygun, I., Kaya, B., & Kaya, M (2022)	928.402 *tweet* vaccine COVID-19	-	*Deep Learning*	87%
Wibowo, D. A., & Musdholifah, A. (2021)	832 *tweet* vaccine COVID-19	*FastText Embedding*	SVM	88,1%
Proposed Method	35.644 *tweet* vaccive COVID-19	Pelabelan *library TextBlob* dan TF-IDF	SVM *Naïve Bayes* KNN	96,3% 94% 91%

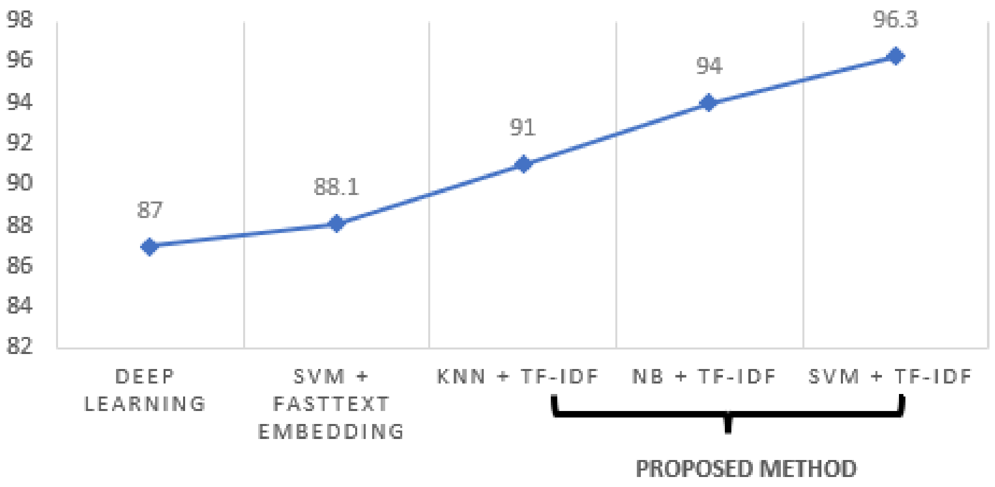


Figure 5. Comparison Graph of Accuracy with Previous Research

The difference between this study and previous research is in the application of the algorithm and feature selection used. The similarity of this study with previous research is the use of the vaccineCOVID19 keyword dataset being crawled. In a study conducted by Aygun, I., Kaya, B., & Kaya, M (2022) applied Deep Learning which resulted in an accuracy of 87%. The total accuracy value was obtained using a dataset of 928,402 tweets. This study also achieved F1 between 84% - 88% (Aygun et al., 2022). Another study, conducted by Wibowo, DA, & Musdholifah, A. (2021) by applying the FastText Embedding and the SVM algorithm resulted in an accuracy of 88.1%. Accuracy results were obtained using a dataset of 832 tweets. This study also compares the Naïve Bayes and FastText-SVM, with the final result of the FastText-SVM model being superior to other models (Wibowo & Musdholifah, 2021).

The accuracy results are quite high in this study, which is above 90% supported by the TF-IDF feature extraction and labeling using the TextBlob library. While the drawback in this study is that after the word weighting process using the TF-IDF feature extraction, the words with the highest weight to the lowest weight have not been able to sort words with the provisions of certain weights being selected for later selection. For further researchers, it is expected to use feature selection in order to be able to select words with certain weight provisions so that they can produce more optimal accuracy values. As well as conducting research using different classification algorithms so that comparisons can be made to the level of accuracy generated in the sentiment analysis process.

5 Conclusion

Based on the results of research that has been conducted on the COVID-19 vaccine on Twitter, it is certain that the way to analyze public sentiment about the COVID-19 vaccine using three classification algorithms, such as SVM, Naïve Bayes, and KNN is carried out in several stages. The first stage is the crawling, followed by preprocessing, labeling using the TextBlob library, normalization, word weighting using TF-IDF, and finally the classification process using the SVM, Naïve Bayes, and KNN algorithms.

This study was conducted to compare the accuracy results of three classification algorithms, namely SVM, Naïve Bayes, and KNN in the process of sentiment analysis of the COVID-19 vaccine on Twitter. After testing three classification algorithms, namely SVM, Naïve Bayes, and KNN 10 times, the highest accuracy values were 96.3% for SVM; 94% for Naïve Bayes and 91% on KNN algorithm. The highest accuracy is obtained using the SVM algorithm, therefore the use of this algorithm is suitable for determining sentiment analysis about the COVID-19 vaccine on Twitter.

6 References

Alamsyah, A., Prasetiyo, B., Hakim, M. F. al, & Pradana, F. D. (2021). Prediction of COVID-19 Using Recurrent Neural Network Model. *Scientific Journal of Informatics*, *8*(1), 98–103.

Ashari, I. A., Muslim, M. A., & Alamsyah, A. (2016). Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing. *Scientific Journal of Informatics*, *3*(2), 149–158.

Athoillah, M. (2018). Klasifikasi Kendaraan Bermotor Dengan Multi Kernel Support Vector Machine. *Buana Matematika : Jurnal Ilmiah Matematika Dan Pendidikan Matematika*, *8*(1:), 1–8.

Aygun, I., Kaya, B., & Kaya, M. (2022). Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic With Deep Learning. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 2360–2369.

Bose, R., Aithal, P. S., & Roy, S. (2020). Sentiment analysis on the basis of tweeter comments of application of drugs by customary language toolkit and textblob opinions of distinct countries. *International Journal of Emerging Trends in Engineering Research*, *8*(7), 3684–3696.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, 112–116.

Fitriana, F., Utami, E., & al Fatta, H. (2021). Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes. *Jurnal Komtika (Komputasi Dan Informatika)*, *5*(1), 19–25.

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 1–4.

Kao, A., & Poteet, S. R. (2007). Overview. In *Natural Language Processing and Text Mining* (pp. 1–7). Springer London.

Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). *Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques*. *4*(1), 39–45.

Larasati, U. I., Muslim, M. A., Arifudin, R., & Alamsyah, A. (2019). Improve the Accuracy of Support Vector Machine Using Chi Square Statistic and Term Frequency Inverse Document Frequency on Movie Review Sentiment Analysis. *Scientific Journal of Informatics*, *6*(1), 138–149.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Mubarok, M. S., Adiwijaya, & Aldhi, M. D. (2017). *Aspect-based sentiment analysis to review products using Naïve Bayes*. 020060.

Mustakim, & Oktaviani F, G. (2016). *Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa*. *13*(2), 195–202.

Pedregosa, Fabian, et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Rasenda, R., Lubis, H., & Ridwan, R. (2020). Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *4*(2), 369.

Sohrabi, C., Alsafi, Z., O’Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., & Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *International Journal of Surgery*, *76*, 71–76.

Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*, *57*, 821–829.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126.

Wibowo, D. A., & Musdholifah, A. (2021). Sentiments Analysis of Indonesian Tweet About Covid-19 Vaccine Using Support Vector Machine and Fasttext Embedding. *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 184–188.

Windasari, I. P., Uzzi, F. N., & Satoto, K. I. (2017). Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek. *Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017*, *2018-Janua*, 266–269.