JournalofInformation Technology andComputerScience
Volume 6, Number 2, August 2021, pp. 195-208
JournalHomepage: www.jitecs.ub.ac.id

Sentiment Visualization Of Covid-19 Vaccine Based On Naïve Bayes Analysis

Nabilah Putri Aprilia[1], Dian Pratiwi*[2], Anung Barlianto[3]

[1,2,3]Faculty of Industrial Engineering,Trisakti University

[1]nabilah064001700022@trisakti.ac.id, [2]dian.pratiwi@trisakti.ac.id,
[3]anung@trisakti.ac.id

*Corresponding Author

Received 22 July 2021; accepted 07 October 2021

Abstract. COVID-19 is one of the topics that is being discussed intensively. The virus which was declared a global pandemic on March 11 by WHO caused around 2.09 million Indonesians to be infected with the COVID-19 virus. To overcome this, the government carried out a vaccination program. The data taken for this study is public opinion about the COVID-19 vaccine written on Twitter. The idea of this research is to classify the covid vaccination dataset using the naive Bayes classifier method and visualization using word cloud. Crawling to obtain the dataset from Twitter, text pre-processing and labelling to determine the positive and negative classes, TF-IDF feature extraction for word weighting, data splitting with a percentage of 80% for train data and 20% for data testing, and finally classification using nave Bayes are the stages in this research The system's sentiment analysis research yielded significant results, the accuracy value is 73.1%, the precision value is 73% and the recall value is 83%.

Keywords: sentiment analysis, naïve bayes, lexicon-based, twitter

1 Introduction

Social media is an online media whose role is to discuss, communicate and exchange ideas. Quoted from We Are Social reports as many as 150 million of the 268 million population. Based on this information, there is a lot of information or opinions of the Indonesian population that is spilled onto social media. Social media is currently a very popular communication feature, especially Twitter. Daily active users climbed by 17% in the third quarter of 2019, according to Twitter's financial report, and the company now has 145 million users. Every day, Indonesia is said to be one of the most active users [1].

For example, a new disease outbreak caused by the corona virus (2019-nCov). COVID-19 was declared a global pandemic by the World Health Organization (WHO) on March 11, 2020 [2]. With the first epicentre in Wuhan City, China, the virus has now spread to the entire world community with 179 million cases and 3.89 million deaths as of June 28, 2021 [3]. On March 2, 2020, the first case of COVID-19 arrived in Indonesia, which infected 2 Indonesian citizens from Depok, West Java. Starting from this case, the number of corona virus cases is increasing every day. As of June 26, 2021, there were 2.09 million cases with a death rate of 56,729 thousand.

As a result of these conditions, the government now has to impose health regulations such as wearing masks, keeping a safe distance, washing hands, and enforcing PSBB, which has hampered all community activities.Seeing the rapid spread of COVID-19 if it is not handled immediately, one way to prevent the spread of the corona virus is through vaccines. Vaccines not only protect the people being vaccinated but also the wider community by reducing the spread of disease. This chain of human-to-human transmission can be broken, even if there is no 100% immunity, it is called herd immunity which is an important benefit of vaccination [4].The Indonesian government also plans to carry out vaccination activities that will be given to the community later. The government carried out the COVID-19 vaccination program for the first time, on Wednesday (13/1) at the state palace, with President Joko Widodo being injected with the vaccine made by Sinovac[5].

Sentiment analysis is a type of social media analysis in which consumers or professionals express their opinions. Sentiment analysis is important because Sentiment analysis is the process of understanding, processing, and extracting text data automatically, aiming to provide emotional information or the essence of opinion sentences to understand the tendency of a person's problem, whether it is positive or negative.There are three types of sentiment analysis opinions: positive, negative, and neutral [6]. Sentiment analysis uses lexicon-based. Lexicon-based generally uses a dictionary to support sentiment classification [7].The NB approach is one of many algorithms that can be used to analyze sentiment and identify tweets containing hate speech.

The NB approach can be used to classify data on Twitter that contains hate speech. This is because the NB method has been proven to have high accuracy and fast computing power [8]. In addition to classifying hate speech, it can also be visualized using Word cloud. Word cloud is a program designed to visualize words by highlighting the frequency with which the text appears. Word cloud serves as a graphical representation of a document, which is done by plotting words that usually appear in a document in a two-dimensional space. The frequency of words that often appear is usually large, indicating that the word often appears in documents [9]. Word cloud is used as an illustration of the classification results of hate speech, namely Word cloud regarding positive class text and Word cloud regarding negative class text. In this study, the researcher will use the Naive Bayes method by retrieving real-time data on Twitter regarding the COVID-19 vaccine by crawling Twitter with the keywords “Vaksin Covid-19” and “Vaksin Corona” and using the TF-IDF extraction feature for word weighting and the Lexicon campus to perform automatic labeling on the dataset as well as classification using nave Bayes and visualization for seeing sentiment trends with word cloud

2 Related Works

This study using information from previous studies as a guide and also a comparison, both in terms of advantages and disadvantages that already exist. There is also research on sentiment analysis of covid-19, in this study [10] authors analyzes about analysis sentiment COVID-19 vaccine by taking data with the keywords "Jokowi" and "Coronavirus". In addition, this study divides 8 emotions into 8 categories, namely: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. By using the nave Bayes method, the results are 35% positive sentiment, 46% negative sentiment, and 20% neutral. More negative sentiments are because people are dissatisfied with the regulations from President Jokowi at the beginning of Covid-19. In addition, in emotion analysis, there are 3 dominant emotions as anticipation, sadness, and anger. There is study who use polarity sentiment for 2 languagesFilipino and English, authors use rapid miner and naïve bayes classifier algorithms[11].The Rapid Miner search Twitter operator collects every week from March 1 to 31 with the keywords “#covidvaccineph”, “#covid19vaccineph” and results in a total of 11,974 tweets. By using nave Bayes, the results are 89.98% for positive and 9& neutral and 8% for negative sentiment.

And there is a research on sentiment analysis of covid-19[12], authors analyzes the sentiment of covid-19 by taking datasets from Twitter with keywords using hashtags such as "jobs", "coronavirus", "school", and "COVID-19" using the Twint library in python and also uses the Text Blob library to define sentiment. The dataset obtained is 42516 tweets taken during the lockdown using the naive Bayes model and calculating mean, max, min sentiment with 37% positive and 27% negative results and for neutral getting 36%. and there are also researchers who use Text Blob library for sentiment popularity, Text Blob has an API for NLP and can determine the sentiment polarity, which is positive negative, or neutral. author [13] comparing three machine learning, Logistic Regression, Multinomial Naïve Bayes, and Support Vector Machineand the highest results were obtained, LR97.3%, SVM 96.26%, and MNB 88%. However, for LR run-time, the longest is around 3 minutes, while for the fastest, MNB, it is only about 3-5 secs. And it can be concluded that all models provide accuracy above 85%

In previous studies on sentiment analysis, many methods were. the methods that are often used are Naive Bayes, Support Vector Machine, and Logistic Regression and display different results for the level of accuracy but all model is accurate for real data.

3 Materials and method

The purpose of this research is to see how well the Naive Bayes algorithm performs when it comes to sentiment classification in Twitter tweets. The flow of this system is shown in Fig. 1 there are several processes, The first step in this research is the collection of tweet data from Twitter users using the crawling method.

Then enter the next stage, the preprocessing stage, feature selection, and labeling. In the preprocessing process, there are 3 processes, case folding, cleansing, and tokenization.Then enter the feature selection stage using the process ofremoving stopwords and stemming. And enter the labeling stage, at the labeling stage it is divided into 2 classes, negative and positive with automatic labeling using In the lexicon dictionary or dictionary-based, this labeling stage is carried out on the dataset after the cleansing process is carried out because in the process the result of the stop words there are punctuation marks "[ ]" and " ' " which makes it difficult during the labeling process.

In the cleaning stage, tweets will be cleaned such as removing links, mentions, hashtags, then changing all letters to lower case.Next tokenization is done to separate each word from another word. Then there is normalization, which is done such as changing non-standard words into standard words. And in the feature selection stage, there is remove stop words, in the remove stop words stage it will remove meaningless words such as the words: yang, tau, and. Then finally there is stemming, at this stage changing each word to its original form. The next step is feature extraction wherein in this step the word weighting process is carried out using TF-IDF.

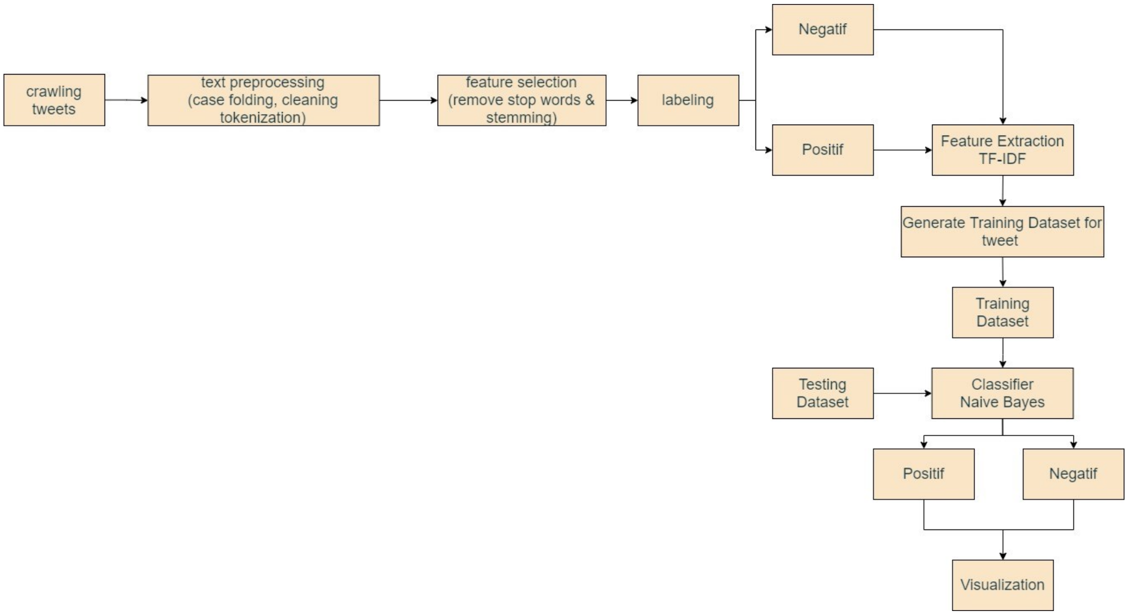

crawling tweets
text preprocessing (case folding, cleaning tokenization)
feature selection (remove stop words & stemming)
labeling
Negatif
Positif
Feature Extraction TF-IDF
Generate Training Dataset for tweet
Training Dataset
Testing Dataset
Classifier Naive Bayes
Positif
Negatif
Visualization

Fig. 1 Flow Diagram

After performing the feature extraction, the next step is to generate a dataset by dividing the 2 datasets into training data and testing data with a ratio of 80:20. After that, enter the classifier and get the results of accuracy and precision, recall, and f-measure or F1-Score for each class. After classifying then visualization using word cloud.

3.1 Dataset

The dataset used tweets from Twitter users taken from November 2020 to March 2021 using the keyword "covid vaccine". The dataset was taken using the Twitter API and was successfully retrieved about 10,000 tweets. After crawling, the dataset will be given a positive or negative label which is done automatically using Python by utilizing the lexicon-based method or positive-negative-based dictionary. The positive and negative dictionaries used in this study are positive and negative dictionaries from several datasets that are made into one data, one dataset is researched by Liu, Hu, & Cheng which has [19]been translated into Indonesian & has been made language adjustments. This dictionary has previously been applied to several previous studies [20], [21].The labeling process is a process for get the expected corpus representation. The labeling process uses a lexicon-based method. In a system that uses a lexicon-based approach, the dictionary is a critical component. Dictionaries are used in the normalization process of sentences and keyword extraction is useful for determining positive and negative sentiments. Examples of dictionary content such as positive keywords "baik", "terimakasih" and negative keywords "sakit", "gamau".The value obtained from the calculation produces a score and will be calculated, from the calculated results, the tweet is entered into a positive or negative class. and get the results of the amount of data from the sentiment class labeling process using Python by using a jupyter notebook

3.2 Preprocessing

The preprocessing stage is where meaningless or unstructured words are removed. There are various phases in preprocessing, including case folding, cleaning, tokenization, normalization, removing stop words, and stemming.Case folding is changing words into lowercase letters, cleaning is the removal of punctuation marks, symbols, numbers that are not needed to eliminate noise that can cause the classification process to be less than optimal. Tokenization is the process of making tokens or solving sentences that are made into several parts or what are called tokens. Normalization is changing non-standard words into standard words. Remove stop words is the removal of unimportant words in sentences such as the words "yang", "at", "and". Stemming is a word converter into a basic word that is useful for increasing the level of accuracy.

Table 1 Preprocessing Data

Phase	Input	Output
Case Folding	b'**Warga** yang pernah mengikuti kegiatan berkerumun akan diminta jalani tes vaksin. https://t.co/Axwc8vKZ7s'	b'**warga** yang pernah mengikuti kegiatan berkerumun akan diminta jalani tes vaksin. https://t.co/Axwc8vKZ7s'
Cleaning	**b'**wargayangpernahmengikutikegiatanberkerumun akandimintajalanitesvaksin. **https://t.co/ axwc8vkz7s'**	wargayangpernahmengikutikegiatan berkerumun akandimintajalanitesvaksin
Tokenization	wargayangpernahmengikutikegiatanberkerumun akandimintajalanitesvaksin	['warga', 'yang', 'pernah', 'mengikuti', 'kegiatan', 'berkerumun', 'akan', 'diminta', 'jalani', 'tes', 'vaksin']
Remove stop words	warga yang pernah mengikuti kegiatan berkerumun akan diminta jalaini tes vaksin	['warga', 'mengikuti', 'kegiatan', 'berkerumun', 'jalani', 'tes', 'vaksin']
normalization	['cuman', 'tolak', 'tes', 'vaksin', 'kabur', 'bawa', 'kabur', 'pasien', 'covid', 'didenda', 'rp', 'juta']	['Cuma', 'tolak', 'tes', 'vaksin', 'kabur', 'bawa', 'kabur', 'pasien', 'covid', 'didenda', 'rp', 'juta']
Stemming	['warga', 'mengikuti', 'kegiatan', 'berkerumun', 'jalani', 'tes', 'vaksin']	['warga', 'ikut', 'giat', 'kerumun', 'jalan', 'tes', 'vaksin']

3.3 Feature Extraction

In the feature extraction method, TF-IDF is applied. The TF-IDF is a method for determiningtheimportanceof a word (term) in a document.The frequencywithwhich a word appears in a document and the inverse frequency with which the document that contains the term appears are combined in this method. The number of times the term documents appears in papers indicates how common it is.

$$W_{dt} = TD(t,d) \tag{1}$$

TD represents the frequency of term t in document d. TFIDF also includes an inverse document frequency. IDF which aims to give high weight to low-value conditions in general conditions. The formula is as follows:

$$IDF_t = \log\left(\frac{N}{Nt}\right) \tag{2}$$

Where $N_t$is the total number of documents that include the term and The document's total number of documents is represented by N. Based on the formula below, TFIDF is a mix of TF and IDF:

$$W_t = TF\ (t,d) \times\ IDF_t \tag{3}$$

TF is considered to have the same weight, but there are some terms whose weights are less important and do not need to be counted, such as affixes, “di-”, or i “and”. So TF-IDF must reduce and add weight to each term and get an equation for the TF-IDF score [15]

3.4 Performance Evaluation Measure

In measuring classification performance there are several ways, but the method is by calculating accuracy, precision, recall, and f-measure. The confusion matrix is used at this evaluation stage.

Table 2Confusion Matrix

Actual	Predict	
	Negative	positive
Negative	TN	FP
Positive	FN	TP

Accuracy is the percentage of the total sentiment that is correctly recognized. The calculation of accuracy is done by dividing the number of sentimental data that takes a moment by the total data as well as the test data. To calculate the value of accuracy is done by using equation (1).

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

To measure the classification performance, the methods used in addition to calculating accuracy are calculating precision, recall, and f-measure. Precision is a comparison of the amount of relevant data found. The number of positive data and the positive value of false data are divided to calculate precision. The value of the false-positive data is taken from the number of true positive values completed in the appropriate column for each class. To calculate the precision value can be done by using the equation:

$$precision = \frac{TP}{TP+FP} \quad (2)$$

The recall is a comparison of the amount of relevant material discovered based on the amount of relevant material. Recall calculation is done by dividing the correct data with a positive value with the sum of the correct data with a negative value. The value of false data that is negative is taken from the number of values other than true positive rows that correspond to each class. Recall calculation can use the equation:

$$recall = \frac{TP}{TP+FN} \quad (3)$$

The F-measure or F1-score is a single parameter measuring retrieval success that combines recall and precision. The F-measure value is obtained from the calculation of the multiplication of precision and recall divided by the result of the addition of precision and recall then multiplied by two and the f-measure calculation using the equation

$$f - measure = 2 \times \frac{precision \times recall}{preicison+recall} \quad (4)$$

3.5 Naïve Bayes

Naïve Bayes Classifier is a text classification algorithm that uses probability and statistical calculations proposed by Thomas Bayes. This algorithm is used to predict future possibilities based on past experiences [21]. The Naïve Bayes Classifier has the advantage of a fast calculation process, easy to use with a simple and efficient structure implementation [22]. The Naive Bayes Classification technique can obtain very good accuracy values and good processing time complexity in very large data during text classification. The Naïve Bayes classification equation is included in equation[8]

$$P(C|X) = \frac{P(x_1|C)...P(x_n|C)P(C)}{P(x)} \qquad (1)$$

Information:

C :data whose class is unknown
X : a data hypothesis specification class C
P(C|X) :probability of C in class X
$P(X_1|C)$:probability$X_{1,}$… X_n is a tweet on C
P(X) :probability of occurrence X
P(C) : probability of event C

4Result and Decision

After all the designs have been compiled and the system has been built, the next step is to use the system to generate a probability model from the training data, then use data testing to evaluate the model's accuracy.

4.1 AnalisisSentimen

After doing text preprocessing, the sentiment class labeling will then be carried out. Sentiment analysis is divided into two types, document-level analysis, and sentence analysis. In this study, sentiment analysis is carried out on each sentence because every tweet has a sentiment.

In sentiment labeling, the analysis is carried out based on the lexicon dictionary. Lexicon dictionary labeling is done by labeling the data by selecting each word into a positive word or a negative word. After that, the words that have entered the positive and negative categories will be calculated and it can be seen whether the tweet is a positive tweet or a negative tweet.

In the sentiment class thickening process, it is done automatically using Jupyter notebook. The assessment score system will be generated in the form of positive or negative, if the sentiment score is below 0 (sentiment < 0) then the system will judge it as negative sentiment, while if the sentiment score is above 0(sentiment >= 0) then the system will judge it as a positive sentiment. In determining whether a tweet is included in the positive or negative class, each word contained in the covid vaccine dataset must be counted by equating the covid vaccine dataset with the lexicon dataset. From the results of the analysis, 11,750 words were found in the covid vaccine dataset and 8,575 words that were not in the lexicon dataset, and only 1,945 words in the lexicon with a total of 10,248 words in the lexicon dataset or about 84% not in the lexicon dataset and 16 % of words contained in the lexicon dataset. Words contained in the lexicon have a weight which will be calculated in the system to determine whether the sentence is included in the positive or negative class and words that are not in the lexicon will not be weighed, there is an example of calculating the weight in Table3

Docs 1 warga yang pernah mengikuti kegiatan berkerumun akan
diminta jalani tes vaksin

Docs 2 kemenag usul beli vaksin covid buatan saudi demi kelancaran haji

Docs 3 perbandingan antara vaksin covid

Table 3 Lexicon Weight Calculation Example

TweetDoc 1	Weight	TweetDoc 2	Weight	TweetDoc 3	Weight
Warga	None	Kemenag	None	Perbandingan	3
Yang	-5	Usul	2	Antara	None
Pernah	None	Beli	2	Vaksin	None
Mengikuti	3	Vaksin	None		
Kegiatan	3	Covid	None		
Berkerumun	-4	Buatan	3		
Akan	None	Saudi	None	covid	None
dimintai	2	Demi	None		
Jalani	None	Kelancaran	4		
Tes	None	Haji	None		
vaksin	None				
Result	**-1**		**11**		**3**

In Table 3, it can be seen from the example of 3 docs that there are only a few words contained in the lexicon, as in tweet docs 1 of 11 words there are only 5 words, namely which, following, activities, crowding, and asking with words that each have a weight. contained in the lexicon dataset. The 5 words will be calculated in number and get a weight of -1. For tweet docs 2 out of 10 words there are only 4 words in the lexicon dataset and the result is a weight of 11. And in the tweet doc 3 of 4 words, there is only 1 word and the result is a weight of 3.

Table 4 Labeling Result

Tweet	Sentiment	Label
warga yang pernah mengikuti kegiatan berkerumun akan diminta jalani tes vaksin	-1	Negative
kemenag usul beli vaksin covid buatan saudi demi kelancaran haji	11	Positive
perbandingan antara vaksin covid	3	Positive

After the process of labeling the sentiment analysis class, it can be seen that the number of tweets for negative and positive opinions on the covid vaccine is more than the tweets for negative opinions on the covid vaccine. For tweets that fall into the positive sentiment class, 5,885 tweets are obtained and 2,954 tweets enter the negative sentiment class

4.2 Naïve Bayes Classifier

In the classification process, the research was carried out by building a system using training data and test data from all random COVID vaccine data. The data classification method is carried out using probability calculations per class. There is an example of the calculation of nave Bayes to determine the probability class of tweets can be seen below:

In this example calculation, we use 2 training data whose class is known and 1 testing data whose class is not known.

Example of training data:

Positive class

document: ['kemenag', 'usul', 'beli', 'vaksin', 'covid', 'buat', 'saudi','lancar', 'haji']

Negative class

document: ['cuma', 'tolak', 'tes', 'vaksin', 'kabur', 'bawa', 'kabur', 'pasien', 'covid', 'denda', 'rp', 'juta']

Examples of Testing Data whose class is not yet known

document: ['riset', 'vaksin', 'covid', 'lomba', 'taruh', 'gengsi', 'antarnegara', 'presiden', 'rusia']

Table 5 Prior Calculation Example

Calculate the probability of each class	**Prior P(c)**
P (positif)= 1/2	0.50
P (negatif)= 1/2	0.50

In Table 4.14 the first thing to do is calculate the prior value of each class contained in the data train, in the example Table 5 there is 1 tweet in the positive class and 1 tweet in the negative class and the probability value of each class is 0.50. then look for the likelihood value can be seen in Table 4 and 7

Table 6 Positive Class Likelihood Calculation Example

No	**Term**	**Frequency**	**Likelihood P(X|C)**	**Result**
1	kemenag	1	P(Kemenag| pos)= 1/9	0.11
2	usul	1	P(usul| pos)= 1/9	0.11
3	beli	1	P(beli| pos)= 1/9	0.11
4	vaksin	1	P(vaksin| pos)= 1/9	0.11
5	covid	1	P(covid| pos)= 1/9	0.11
6	buat	1	P(buat| pos)= 1/9	0.11
7	saudi	1	P(saudi| pos)= 1/9	0.11
8	lancar	1	P(lancar| pos)= 1/9	0.11
9	haji	1	P(haji| pos)= 1/9	0.11
Total Words		**9**		

Table 7 Negative Class Likelihood Calculation Example

No	Term	Frequency	Likelihood P(X|C)	Result
1	Cuma	1	P(Cuma|negatif) = 1/12	0.0833
2	tolak	1	P(tolak|negatif) = 1/12	0.0833
3	tes	1	P(tes|negatif) = 1/12	0.0833
4	vaksin	1	P(vaksin|negatif) = 1/12	0.0833
5	kabur	2	P(kabur|negatif) = 2/12	0.1667
6	bawa	1	P(bawa|negatif) = 1/12	0.0833
7	pasien	1	P(Pasien|negatif) = 1/12	0.0833
8	covid	1	P(covid|negatif) = 1/12	0.0833
9	denda	1	P(denda|negatif) = 1/12	0.0833
10	rp	1	P(rp|negatif) = 1/12	0.0833
11	juta	1	P(juta|negatif) = 1/12	0.0833
Total Words		**9**		

In Table 6 and Table 7 there are calculations to find the likelihood of each word in the train data in positive and negative classes. To find the likelihood, each TF (term frequency) of each word must be calculated and get the results. After that, to do the calculation of evidence first, it is calculated for a total of 21 terms.

Table 8 Evidance Calculating Example

No	Term	Train Term	Frequency	P(x) Evidance	
1	riset	tidak ada	0	0/21=	0
2	vaksin	ada	2	2/21 =	0.10
3	covid	ada	2	2/21 =	0.10
4	lomba	tidak ada	0	0/21=	0
5	taruh	tidak ada	0	0/21=	0
6	gengsi	tidak ada	0	0/21=	0
7	antarnegara	tidak ada	0	0/21=	0
8	presiden	tidak ada	0	0/21=	0
9	rusia	tidak ada	0	0/21=	0

In Table 8 there are calculations to find evidence in the testing data. The data testing term will be equated with the training data term. If so, the number of terms will be calculated. After getting the prior, likelihood and evidence values, they are calculated using the Nave Bayes equation for each negative or positive class

$$P(C|X) = \frac{P(x_1|c)...P(x_n|c)P(C)}{P(x)}$$

$$P(pos|datatest) = \frac{P(Vaksin|Positif).P(Covid|Positif).P(pos)}{P(vaksin).P(covid)} = \frac{(0.11)(0.11)(0,50)}{(0.10)(0.10)} =$$

0.00605

$$P(neg|datatest) = \frac{P(Vaksin|Positif).P(Covid|Positif).P(neg)}{P(vaksin).P(covid)} = \frac{(0.0833)(0.0833)(0,50)}{(0.10)(0.10)} =$$

0.00347

The probability value in the testing data is greater in the positive class, which is 0.00605 and the negative class is 0.00347 and the testing data is assumed to be in a positive class because the positive class is larger.

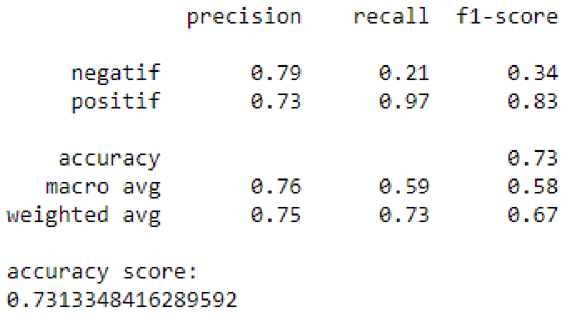
To get an evaluation of the test model that has been done by data testing, a confusion matrix method is needed to calculate the accurate value of the data classification results. The confusion matrix is used to test the prediction results of the classification method,

The Table displayed in this confusion matrix consists of the predicted class and the actual class.

Actual	**Predict**	
	Negative	**Positive**
Negative	120	444
Positive	31	1173

Fig 3. Confusion Matrix

It can be seen in Fig. 3, the confusion matrix 2x2 for each column represents the value of each positive class and negative class.



```
              precision    recall  f1-score

     negatif       0.79      0.21      0.34
     positif       0.73      0.97      0.83

    accuracy                           0.73
   macro avg       0.76      0.59      0.58
weighted avg       0.75      0.73      0.67

accuracy score:
0.7313348416289592
```

Fig 4. Accuracy Results

The results of testing the model that is trained using the naive Bayes classifier utilizing the confusion matrix approach in the evaluation process, which includes a comparison of training data by 80% and test data by 20% from 8837 are shown in Fig. 4. The amount for training data is 7071 and for testing data of 1768 tweet data with an accuracy value of 73.1%. The accuracy value is obtained from the performance evaluation measure formula, 73.1 % is summing the true positive and true negative data, then dividing the true positive, true negative, false positive, and false negative data.for the precision value based on the performance evaluation measure formula, true positive divided by true positive and false negative yields a result of 73%. Meanwhile, The precision value is calculated using the formula 2 times precision multiplied by recall, then divided by precision plus recall, giving an 83% precision value.As a result, the naive Bayes classifier method can be recommended for use in the classification of tweets with Indonesian text on the sentiment of the covid vaccine with a high accuracy value.

4.3 Visualization

Wordcloud is a word visualization method that highlights the frequency of words in a given text. The more frequently the term appears in the data and as a tool for doing analysis, the greater the word appears in the word cloud.The purpose of the visualization is to extract information in the form of topics or opinions that are often discussed by the public about the covid vaccine so that from the many available tweets, information that is considered important can be taken. A word cloud can be used to visualize these words; for example, see Fig. 4 for a word cloud on the covid vaccine.

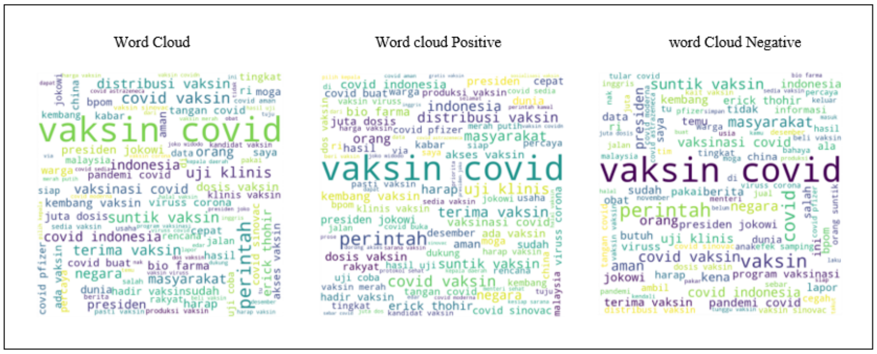


Fig. 4 Wordcloud

In Fig. 4 it can be seen that word cloud is more likely to have positive sentiments because it can be seen in words such as “uji klinis”, “distribusivaksin”, and “pastivaksin”. To see the visualization of the positive and negative sentiment word cloud, see word cloud positive and word cloud negative. In word cloud positive the tweet data is classified as positive using lexicon based sentiment analysis. It can be seen that the covid vaccine is often discussed and there are the words "Uji klinis", "sediavaksin", and "vaksin gratis". In word cloud negative, tweet data is classified as negative using lexicon-based sentiment analysis. The difference between positive and negative word clouds may be recognized in the negative word cloud. "vaksincovid" is still a topic that is often discussed, but in the negative word cloud there are also words such as "tularcovid", "takut", and "efeksamping".

5 Conclusion

Following the purpose of this study, namely implementing the nave Bayes classifier algorithm for sentiment analysis and visualizing the opinion of the covid tweet vaccine in Indonesian, it was successfully carried out. So some conclusions are drawn as follows:

1. From the results of the tests and analyzes that have been carried out, it can be concluded several things, namely the classification of opinion data on covid vaccines in Indonesian can be done with the naive Bayes classifier algorithm by previously using lexicon-based as an automatic sentiment With an accuracy of 73.1% with a precision value of 73%, the recall value is 83%.

2. Visualization of data from the classification results thereare 3-word clouds. Wordcloud whole dataset, In the overall word cloud to see the tendency of sentiment and the results obtained the overall word cloud tends to be a positive sentiment. On word cloud, the entire "covid vaccine" dataset is a topic that is often discussed. In addition, there also the words “uji klinis”, “dosisvaksin”, “terimavaksin”. Positive word cloud, In the positive word cloud, just like the word cloud, the entire "vaksincovid" dataset is a topic that is often discussed and then there are also the words kata “uji klinis”, “sediavaksin” and “gratis vaksin”. In the positive word cloud, people have more opinions such as the distribution of vaccines, what is the price of the vaccine, is the vaccine free, and also the candidate who gets the vaccine. Word cloud negative, In the negative word cloud, the same as the word cloud, the entire dataset and positive word cloud "vaksincovid" is still a topic that is often discussed and there are also words such as seperti “tularcovid”, “takut”, and “efeksamping”.. In the negative word cloud, people have more opinions such asinfectious covid, side effects of vaccines, and people are afraid of getting vaccinated.

References

[1] G. A. Buntoro, “Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naïve Bayes Classifier Dan Support Vector Machine,” *J. Din. Inform.*, vol. 5, no. 2, pp. 1–18, 2016, doi: 10.1016/j.cya.2015.11.011.

[2] WHO, “Virtual press conference on COVID-19 – 11 March 2020,” *J. Phys. A Math. Theor.*, vol. 44, no. 8, pp. 1–9, 2020, doi: 10.1088/1751-8113/44/8/085201.

[3] WHO, “WHO Coronavirus Weekly Update,” 2021, [Online]. Available: https://covid19.who.int/.

[4] I. P. Sari and S. Sriwidodo, “Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin COVID-19,” *Maj. Farmasetika*, vol. 5, no. 5, p. 204, 2020, doi: 10.24198/mfarmasetika.v5i5.28082.

[5] Kementerian Kesehatan Republik Indonesia, “Program Vaksinasi COVID-19 Mulai Dilakukan, Presiden Orang Pertama Penerima Suntikan Vaksin COVID-19 | Direktorat Jendral P2P,” *Kementerian Kesehatan RI*, 2021. http://p2p.kemkes.go.id/program-vaksinasi-covid-19-mulai-dilakukan-presiden-orang-pertama-penerima-suntikan-vaksin-covid-19/.

[6] E. M. Sipayung, H. Maharani, and I. Zefanya, “Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier,” *J. Sist. Inf.*, vol. 8, no. 1, pp. 958–965, 2016, [Online]. Available: https://ejournal.unsri.ac.id/index.php/jsi/article/view/3250/1907.

[7] I. 2017 Kusumawati, “Analisa Sentimen Menggunakan Lexicon Based Kenaikan Harga Rokok Pada Media Sosial Twitter,” *Anal. Sentimen Menggunakan Lex. Based Untuk Melihat Persepsi Masy. Terhadap Kenaikan Harga Rokok Pada Media Sos. Twitter*, 2017.

[8] I. Liu and Y. A. Sari, “Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata,” vol. 3, no. 5, pp. 4914–4922, 2019.

[9] F. Kusuma Wardani, R. Valentinus Hananto, and V. Nurcahyawati, “Analisis Sentimen Untuk Pemeringkatan Popularitas Situs Belanja Online Di Indonesia Menggunakan Metode Naive Bayes,” *Jsika*, vol. 08, no. 01, pp. 1–9, 2019.

[10] Samsir *et al.*, “Naives Bayes Algorithm for Twitter Sentiment Analysis,” *J. Phys. Conf. Ser.*, vol. 1933, no. 1, p. 012019, 2021, doi: 10.1088/1742-6596/1933/1/012019.

[11] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, “Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes,” *Inf.*, vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[12] S. Kumar Singh, P. Verma, P. Kumar, and A. Abdul, “Journal of Critical Reviews Sentiment Analysis of Covid-19 Epidemic Using Machine Learning Algorithms on Twitter,” vol. 7, no. 18, p. 2020, 2020.

[13] C. Author and P. Gupta, “Pr ep rin t n ot pe er r ie we d Pr ep rin t n ot pe er we.”

[14] L. Setyo, “Implementasi Text Mining Untuk Mendeteksi Hate Speech Pada Twitter,” 2019.

[15] M. S. Hadna, P. I. Santosa, and W. W. Winarno, “Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter,” *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 57–64, 2016, [Online]. Available: https://fti.uajy.ac.id/sentika/publikasi/makalah/2016/95.pdf.

[16] M. Syarifuddin, “Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn,” *Inti Nusa Mandiri*, vol. 15, no. 1, pp. 23–28, 2020.

[17] P. Antinasari, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1733–1741, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.

[18] E. B. Santoso and A. Nugroho, “Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook,” *Eksplora Inform.*, vol. 9, no. 1, pp. 60–69, 2019, doi: 10.30864/eksplora.v9i1.254.

[19] B. Liu, M. Hu, and J. Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web,” *Proc. 14th Int. Conf. World Wide Web*, pp. 342–351, 2005, [Online]. Available: http://dl.acm.org/citation.cfm?id=1060797.

[20] F. F. Rachman and S. Pramana, “Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter,” *Heal. Inf. Manag. J.*, vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175.

[21] W. Setyobudi, A. Alwi, and I. P. Astuti, “Sentimen Analisis Twitter Terhadap Penyelenggaraan Gojek Traveloka Liga 1 Indonesia,” *Komputek*, vol. 2, no. 1, p. 56, 2018, doi: 10.24269/jkt.v2i1.68.