ANALYSIS OF PUBLIC OPINION SENTIMENT ON COVID-19 VACCINE BASED ON SOCIAL MEDIA

[1]Ida Ayu Gde Suwiprabayanti Putra, [2]Dian Rahmani Putri

[1,2]Information System, Faculty of Informatics and Computers, ITB STIKOM Bali, Denpasar

Author's email: suwiprabayanti@gmail.com[1], rahmani@stikom-bali.ac.id[2]
**Corresponding author: suwiprabayanti@gmail.com*

Abstract

In addition to words or terms in keywords, it serves to make it easier to classify the type of paper, so do not use phrases. To reduce deaths from Covid-19, the Indonesian government in early 2021 started a vaccination program. Using vaccines significantly reduces the overall disease burden. The vaccination program in Indonesia is also trending on Twitter Indonesia. So far, Indonesia has always bought vaccines from producing countries, this makes Indonesia dependent on imported vaccines. There are many countries whose supplies of Covid-19 vaccines have not been met, so they have to scramble to get them. Thus, domestically produced vaccines are also needed. The Red and White Vaccine is the government's effort to meet domestic vaccine needs. Sentiment analysis is the process of determining sentiment and classifying the polarity of text in a document or sentence so that categories can be determined as positive, negative, or neutral sentiment. Currently, researchers widely use sentiment analysis as a branch of research in computer science. Social networks, such as Twitter, are commonly used in sentiment analysis to determine public perception. In sentiment analysis, data mining is performed to analyze, process and extract textual data within an entity, such as a particular service, product, individual, phenomenon or topic. Based on this, in this study an analysis of sentiment towards Covid-19 vaccination will be carried out using social media Twitter. The algorithm used to analyze is the Naive Bayes Classifier Method. This analysis will be carried out using the Weka application. The results of this study are from 14,710 tweet data obtained 9724 word data after preprocessing. Of the 9724 words, there are 5910 words with a tendency of positive sentiment, 897 words of negative sentiment and 2917 words of neutral sentiment. So it can be concluded that for tweets with the topic of "red and white vaccines", the public's tendency to discuss this issue is a positive sentiment. The value of Correctly Classified Instances or data that is classified correctly is 60.77% and Incorrectly Classified Instances or data that is not classified correctly is 39.22%.

Keywords: *twitter, public opinion sentiment, covid-19, vaccine*

1. INTRODUCTION

Covid-19 first appeared in the city of Wuhan, China. The spread is very fast and deadly. It is spread by direct physical contact with humans and is transmitted through the mouth, nose and eyes. The impact of the Covid-19 outbreak is seen in almost all sectors of people's lives. Social activities are prohibited and temporarily suspended, the economy is weakening, transportation services are reduced and strictly regulated, tourism is closed, shopping centers are empty of visitors and the informal sector is closed, such as; Online motorcycle taxis, angkot drivers, street vendors, mobile traders, MSMEs and rough coolies. Trade centers, such as malls, the Tanah

Abang market, which are usually crowded with people, are suddenly deserted and are currently closed temporarily. The tourism sector has decreased, the government has closed tourist attractions and entertainment venues. Work and study is also done at home online. [1]

To reduce deaths from Covid-19, the Indonesian government in early 2021 started a vaccination program. Using vaccines significantly reduces the overall disease burden. [2] So far, Indonesia has always purchased vaccines from producing countries, this has made Indonesia dependent on imported vaccines. There are many countries whose supplies of Covid-19 vaccines have not been met, so they have to scramble to get them. Thus, domestically produced vaccines are also needed. The Red and White Vaccine is the government's effort to meet domestic vaccine needs. Moreover, the current pandemic situation is increasingly concerning with the emergence of various variants of the corona virus that causes Covid-19.

As for the development of the Merah Putih vaccine, the government is collaborating with five universities and two institutions, namely Airlangga University (UNAIR), Gadjah Mada University (UGM), University of Indonesia (UI), Bandung Institute of Technology (ITB), Eijkman Institute for Molecular Biology (LBM). , and the Indonesian Institute of Sciences (LIPI). The vaccine is made and developed with a domestically circulating virus. Currently, the development process is still accelerating and is estimated to be ready for use in 2022.

The vaccination program in Indonesia is also trending on Twitter Indonesia. It can be seen that netizens on Twitter social media care about the vaccination program in Indonesia. Social media with a large number of users is the right place to see the public's response to public policies. Social media is a medium for voicing people's voices. In this case, social media is a platform to respond to the vaccination program in Indonesia. Social media has changed conventional protests to become more digital using social media. [2]

Sentiment analysis is the process of determining sentiment and classifying the polarity of text in a document or sentence so that categories can be determined as positive, negative, or neutral sentiment. Currently, researchers widely use sentiment analysis as a branch of research in computer science. Social networks, such as Twitter, are commonly used in sentiment analysis to determine public perception. Sentiment analysis can also be likened to opinion mining, because it focuses on opinions that state positive or negative. In sentiment analysis, data mining is performed to analyze, process and extract textual data within an entity, such as a particular service, product, individual, phenomenon or topic. The analysis process can include review text, forums, tweets, or blogs, with preprocessing data including tokenization, stopword, deletion, stemming, sentiment identification, and sentiment classification processes. [3]

One of the algorithms that can be used for text mining is the Naïve Bayes Classiffier. Based on this, in this study an analysis of sentiment towards Covid-19 vaccination will be carried out using social media Twitter. The algorithm used to analyze is Naive Bayes Classifier (NBC). The advantage of NBC is that in its simple algorithm, NBC only requires small training data and can be relied on against irrelevant attributes. This analysis will be carried out using the Weka application. Weka has advantages such as having many data mining and machine learning algorithms, ease of use, always up-to-date with new algorithms. [5] The results of this study are the classification of positive, neutral and negative opinions.

2. LITERATURE REVIEW

2.1. Sentiment Analysis

Sentiment analysis or opinion mining is the process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in an opinion sentence. The magnitude of the influence and benefits of sentiment analysis causes research

and applications based on sentiment analysis to develop rapidly. Even in America there are about 20-30 companies that focus on sentiment analysis services. [6]

2.2. Naive Bayes Classifier (NBC)

Bayesian decision theory is a statistical approach that is fundamental to pattern recognition, this approach is based on the quantification of trade-offs between various classification decisions using the probabilities and costs incurred in these decisions. Naive Bayes Classifier or often called Bayesian Classification is a statistical classification method that can be used to predict the probability of membership of a class. Naïve Bayes Classifier (NBC) is proven to have high accuracy and speed when applied to databases with large data. According to Kusirin, the Naïve Bayes Classifier is based on the Bayes theorem which has the following general form [5]:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Description:

X = data with unknown class

H = data hypothesis X is a specific class

P(H|X) = probability of hypothesis H based on condition X (posteriori probability) P(X|H) = probability X based on condition on hypothesis H

P(X) = probability of X

2.3. Weka

Weka was developed in 1994 and has become the most popular open source data mining software. Weka has advantages such as having many data mining and machine learning algorithms, ease of use, always up-to-date with new algorithms. Weka software is not only used for academic purposes but is quite widely used by companies to forecast the business of a company. All Weka techniques are based on the assumption that data is available as a single flat file or relationship, where each data point is described by a number of attributes (usually, attributes). numeric or nominal, but some other attribute types are also supported). [5]

3. RESEARCH METHODS

The conceptual model of this research is shown in Figure 1. Research starts from preparing data. The data is retrieved using the API for social media Twitter. Then the data is processed by doing Data Preprocessing. After the data is ready, it is analyzed using the Naive Bayes Classifier method and using the Weka application. After getting the results from the method, an accuracy test is carried out to see the accuracy of the classification results.

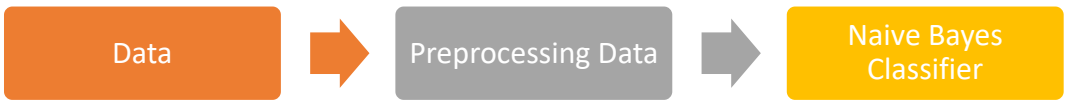

Figure 1. Research Methods

4. RESULTS AND DISCUSSION

The data used in this study were obtained from Twitter. The data is taken using the Tweepy library. The data was taken using the Python programming language and the Jupyter application. From the results of data collection obtained 14,710 tweets. The code and method of data collection are shown in Figure 2.

```
[*]: import tweepy
import csv
import pandas as pd

access_token = "57037292-14jKp9ENhnYNKZJADOwnrxi2a1zeYC8Ck76e3j4Ln"
access_token_secret = "bYwUFtD7sXbDJKnda0gIfBWqHxGlih7zhd9XWicFRT2OP"
api_key = "kgvYELGkIKCTxnjdt0iLGy6Hd"
api_secret_key = "CxNPgKJL3oa1NGdqdsxYDTzYKpua7xSMMr2Z7ei3cjgGKJP1h9"

auth = tweepy.OAuthHandler(api_key, api_secret_key)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)
search_key = "vaksin merah putih"

csvFile = open(search_key+".csv","a+",newline="",encoding="utf-8")
csvWriter = csv.writer(csvFile)
t = []

for tweet in tweepy.Cursor(api.search_tweets,q=search_key,count=99999999999,lang="id",since="2021-01-01").items():
    print(tweet.created_at,tweet.id,tweet.user.name,tweet.text)
    t.append(tweet.text.encode("utf-8"))
    tweets = [tweet.text.encode("utf-8")]
    csvWriter.writerow(tweets)

dictTweets = {"twit":t}
df = pd.DataFrame(dictTweets, columns=["twit"])
df

[ ]:
```

Figure 2. Twitter Data Retrieval Code

In Figure 2 seen the programming language code used to retrieve twitter data. The programming language used is Python. The application used is Jupyter Lab. The API used to connect twitter data is the API available to Twitter Developers. The API used consists of api_key, api_secret_key, access_token and access_token_secret. Then the API library is imported from the Tweepy library. Tweepy is provided by Twitter to access and retrieve data contained within Twitter. The keywords used to collect the data are "red and white vaccines" and the data taken are 14,710 tweets.

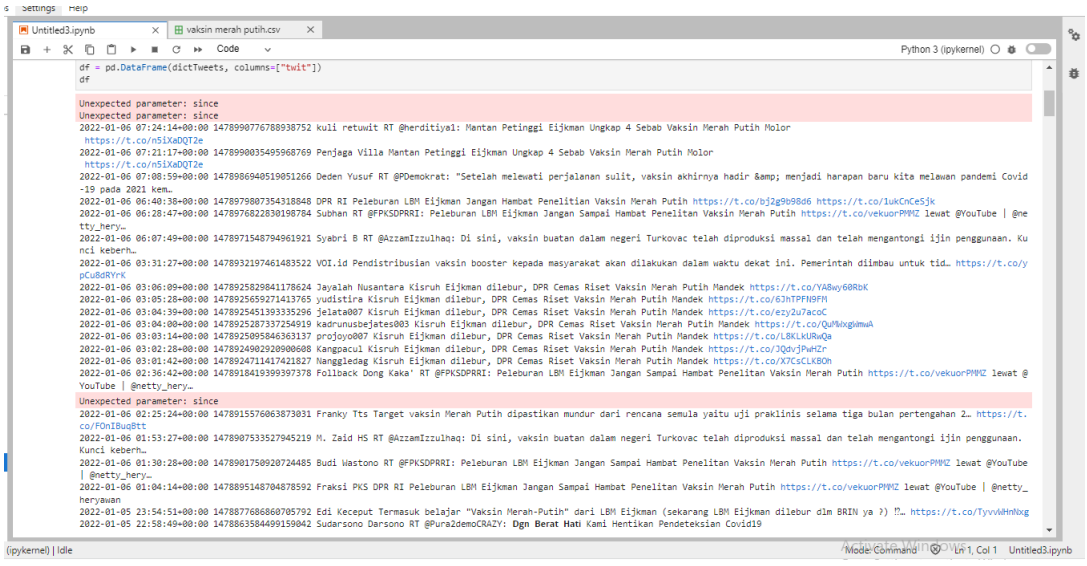

Figure 3. Twitter Data Collection Results

In Figure 3 see the code that has been run on the Jupyter application. At the bottom of the code there are tweet data results obtained using the keyword "red and white vaccine". The data is then transferred to excel to be processed word by word.

The next step is to process the tweet sentences that are obtained word for word. The Tweets obtained are entered into excel form as shown in Figure 4.

Figure 4. Tweet

Furthermore, word cleaning is carried out by tokenization, this process will remove punctuation marks, spaces, and separate word by word. This process uses the NLTK library in Python. The code in this process is shown in Figure 5.

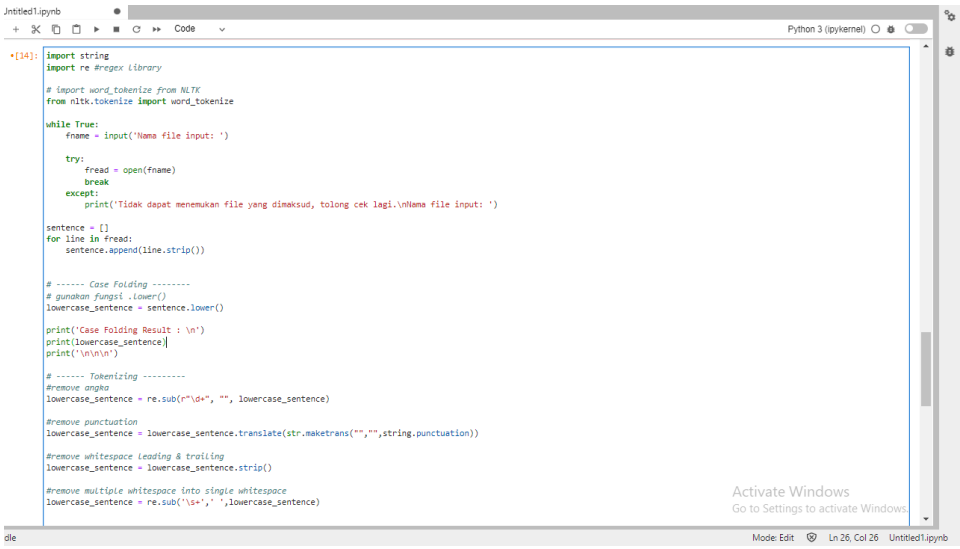

Figure 5. Tokenization

The next process is to remove the stopword. Stopwords are basic words that do not provide important information. The stopword used is imported from the NLTK library. From the tweets that were generated earlier, if any of the words in the tweet are entered into the stopword, those words will be removed. The stopword code used is shown in Figure 6.

Figure 6. Stopwords Code

The last process is to change words into basic words or often called stemming. Stemming will be done by referring to the basic data dictionary that has been provided. The stemming process in Indonesian is done by removing affixes and prefixes in words. The library used is Sastrawi. The code for stemming is shown in Figure 7.

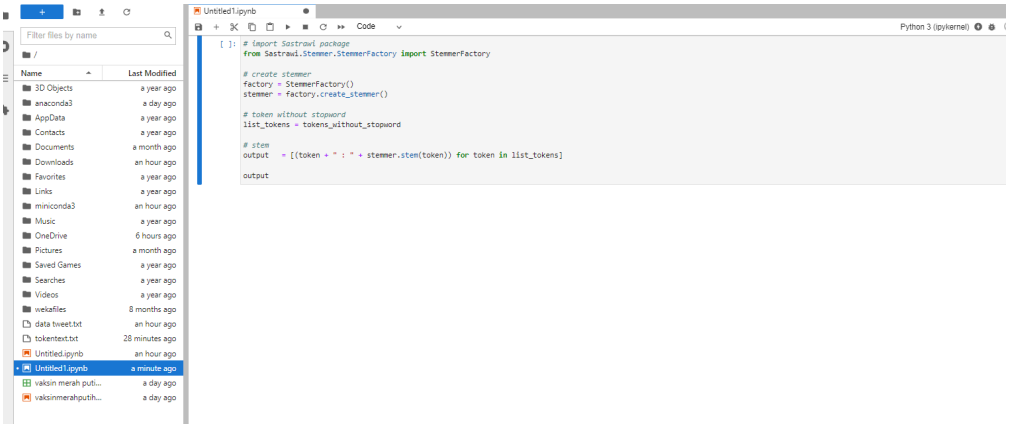

Figure 7. Stemming Code

After all these processes are carried out, there are 9724 words that can be processed for analysis using the Weka application. Examples of words obtained can be seen in Figure 8.

```
BRIN
pengembang
lanjut
proses
integrasi
Eijkman
lembaga
Kepala
pengembangan
Biologi
Molekuler
berhenti
lembaga
lebur
perjalanan
sulit
akhirnya
hadir
harapan
baru
kita
melawan
pandemi
```

Figure 8. Words

The classification process is carried out using the Naive Bayes algorithm and with the help of the Weka application. Before classification, training data is prepared in the form of words that are included in positive, negative and neutral sentiments. Positive and negative words are obtained from the research results in the form of positive sentiment words and negative sentiment words. If it is not included in both, it is categorized as neutral. Examples of words with positive and negative sentiments are shown in Figure 9.

Table 3. Top words that are associated with the small list of positive words.

#	Hebrew	English	Score	#	Hebrew	English	Score
1	לי	for me	47249	26	אם	if	7126
2	אני	me	45924	27	כמה	a few	6862
3	(את) אתה	you	37334	28	שיהיה	that will be	6549
4	זה	this	35197	29	כבר	already	6396
5	היה	was	27990	30	עוד	more	6330
6	יותר	more	19957	31	הזה	this	6135
7	אבל	but	19054	32	כמו	like	5860
8	כל	all	18378	33	מזל	luck	5786
9	עם	with	18332	34	איזה	which	5751
10	על	on	18030	35	קצת	a little	5714
11	שלי	mine	17289	36	לו	to him	5598
12	אז	then	14596	37	שזה	that this	5593
13	רק	only	14159	38	אחד	one	5437
14	מה	what	12978	39	לך	for you	5378
15	הוא	he	11435	40	הכל	everything	5315
16	כי	because	11392	41	משהו	something	5178
17	הכי	the most	10470	42	עכשיו	now	5240
18	כך	like this	10119	43	מאוד	very	5199
19	יש	there is	8747	44	אנשים	people	5153
20	יום	day	8550	45	רוצה	want	5160
21	אותי	me	8039	46	שהוא	that he	5117
22	הרבה	a lot	7955	47	באמת	indeed	4978
23	להיות	to exist	7927	48	אותו	him	4937
24	יהיה	will be	7772	49	שם	there	4904
25	היום	today	7363	50	וזה	and this	4830

Figure 9. Positive and Negative Words

Then from the 9724 words that have been obtained from pre-processing, they are compared with the training data so that the results will be how many words are included in positive, negative or neutral sentiments. Of the 9724 words, after being compared with the training data, they are then classified into three sentiment classes and converted into an arff file to be entered into the Weka application as shown in Figure 10.

```
@relation datatwitter

@attribute data{positif,negatif,netral}

@data
positif
positif
positif
negatif
positif
netral
netral
positif
netral
positif
positif
positif
positif
negatif
positif
netral
netral
positif
netral
positif
positif
positif
positif
negatif
positif
netral
netral
positif
netral
positif
positif
positif
```

Figure 10. Weka Data

After the data is entered into the Weka application, the results from the application are in the form of data visualization as shown in Figure 11. There are 5910 positive words, 897 negative words and 2917 neutral words.

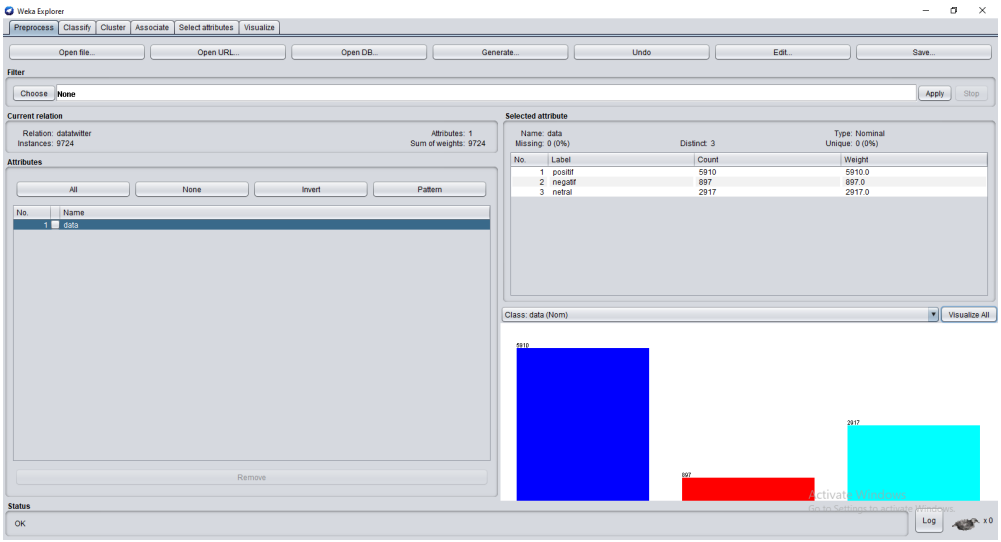

Figure 11. Data Visualization

For data analysis, it can be seen in Figure 12. From the results, it can be seen that the value of Correctly Classified Instances or data that is classified correctly is 60.77% and Incorrectly Classified Instances or data that is not classified correctly is 39.22%.

```
=== Summary ===

Correctly Classified Instances        5910               60.7775 %
Incorrectly Classified Instances      3814               39.2225 %
Kappa statistic                          0
Mean absolute error                      0.3548
Root mean squared error                  0.4212
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             9724
```

Figure 12. Data Analysis

CONCLUSION

The conclusions obtained from this research are:

1. Classification has been successfully carried out using the Naive Bayes Classifier algorithm and the help of the Weka application.
2. From 14,710 tweet data obtained 9724 word data after preprocessing. Of the 9724 words, there are 5910 words with a tendency of positive sentiment, 897 words of negative sentiment and 2917 words of neutral sentiment. So it can be concluded that for tweets with the topic of “red and white vaccines”, the public's tendency to discuss this issue is a positive sentiment.
3. The value of Correctly Classified Instances or data that is classified correctly is 60.77% and Incorrectly Classified Instances or data that is not classified correctly is 39.22%.

REFERENCES

S. Syafrida and R. Hartati, “Bersama Melawan Virus Covid 19 di Indonesia,” SALAM J. Sos. dan Budaya Syar-i, vol. 7, no. 6, pp. 495–508, 2020, doi: 10.15408/sjsbs.v7i6.15325.

D. Kurniawandanarissy and J. Sutan, “PENGGUNAAN SOSIAL MEDIA DALAM MENYEBARKAN PROGRAM VAKSINASI COVID-19 DI INDONESIA,” pp. 27–34, 2015.

Samsir, Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, “Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes,” J. Media Inform. Budidarma, vol. 5, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.

I.- Indriati, B. Rahayudi, and C. Dewi, "Analisis Sentimen Mengenai Moda Raya Terpadu (MRT) Jakarta dengan Metode BM25 dan K-Nearest Neighbor," J. Teknol. Inf. dan Ilmu Komput., vol. 8, no. 2, p. 389, 2021, doi: 10.25126/jtiik.2021824508.

F. L. D. Cahyanti, W. Gata, and F. Sarasati, "Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Dalam Menentukan Tingkat Keberhasilan Immunotherapy Untuk Pengobatan Penyakit Kanker Kulit," J. Ilm. Univ. Batanghari Jambi, vol. 21, no. 1, p. 259, 2021, doi: 10.33087/jiubj.v21i1.1189.

G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," INTEGER J. Inf. Technol., vol. 1, no. 1, pp. 32–41, 2017, [Online].

Yaakov HaCohen-Kernera, Haim Badasha, "Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew" Procedia Computer Science 96 (2016) 733 – 743.