**Jurnal Teknik Informatika (JUTIF)**
Vol. 3, No. 4, August 2022, pp. 1119-1124

DOI: https://doi.org/10.20884/1.jutif.2022.3.4.565
p-ISSN: 2723-3863
e-ISSN: 2723-3871

# *APPLICATION OF LEXICON BASED FOR SENTIMENT ANALYSIS OF COVID-19 BOOSTER VACCINATIONS ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES METHOD*

**Muhamad Fahmi[1], Syarif Hidayah*[2], Ahmad Fathan Hidayatullah[3]**

[1,2,3]Informatics Master Study Program, Industrial Technology Faculty, Universitas Islam Indonesia, Indonesia
Email: [1]21917033@students.uii.ac.id, [2]syarif@uii.ac.id, [3]fathan@uii.ac.id

(Article received: August 18, 2022; Revision: August 21, 2022; published: August 22, 2022)

***Abstract***

*To combat the Covid-19 epidemic, the government issues laws governing vaccination implementation. Health Minister Number Ten of 2021 issued the regulation. This program raises advantages and disadvantages, necessitating examination through feedback. The opinions and narratives that individuals share on social media sites like Twitter can be used to get feedback. This work seeks to construct a model to assess public opinion of the Covid-19 Booster Vaccination by using the Lexicon Based technique to identify sentiment on tweet data. Naïve Bayes and logistic regression are the classification techniques employed in this study. The comparison of the two methods' findings reveals that Logistic Regression, with an accuracy of 72%, is superior to Naïve Bayes, which has an accuracy of 70%. There were 607 tweet messages from Twitter that were processed. From January 1 to July 30, 2022, the model was tested for its ability to interpret public opinion on Twitter. The model found that people's attitudes toward the COVID-19 booster shot tended to be favorable. It can be developed by including datasets for additional research. For further research, it can be developed by adding datasets.*

**Keywords**: *covid-19, lexicon based, logistic regression method, naïve bayes method, sentiment analysis, twitter, vaccination booster.*

## 1. INTRODUCTION

The total number of COVID-19 cases worldwide as of August 8, 2022, was 584,590,338 cases. The majority of these instances were initially concentrated in China, but with time, they started to spread to other nations. Up until this point, India and Brazil each had tens of millions of cases, while the United States had 91,939,428 instances. While in our nation, confirmed cases on that date totaled 6,244,978, with a 157,095-person death rate being a somewhat high number. Since the COVID-19 pandemic has not yet been contained, Indonesia's government has mandated vaccinations for all citizens. To combat the COVID-19 pandemic, the Minister of Health enacted Regulation Number 10 of 2021. In addition, the government has started administering booster shots to combat the current pandemic in Indonesia.

Sentiment analysis is a process in which we evaluate people idea, opinion, feeling, attitude, thought, and belief about a particular subject on specified topic or concept. The topic could be a business organization, a news forum, an enterprise or an online product. It is sometimes referred to as opinion mining [1]. In these growing scenarios social media plays a prime role in dealing with such an enormous amount of information. Traditional data mining techniques does not yield good results due to the increasing amount of data in web each second. So to overcome the problem of data mining different machine learning procedure is enforced. Machine learning classifier can easily deal with large amount of data which was not possible by traditional techniques. The information collected from social media can serve as an important parameter for online enterprises if the information is properly dealt with for knowledge discovery purposes [1].

Several studies related to sentiment analysis have been carried out, one of which is the application of Lexicon Based for Sentiment Analysis on Twitter on Covid-19 Issues. Research conducted shows that public opinion on Twitter who believes that Covid-19 is something real is higher than that of community groups who believe the issue of Covid-19 is a conspiracy. The percentage of grouping can be seen from the positive sentiment category at 58.08%, the negative sentiment opinion category at 37.61%, and the neutral opinion sentiment category at 4.31% [2].

In another study, we compared two classification methods, namely the K-Nearest Neighbor (KNN) method with one of the most frequently used classification methods, namely Naïve Bayes. The data used is 1098 data, the opinion is taken from social media twitter. This study focuses on the comparison between the two methods used and also the tendency of public sentiment towards COVID-19. The result of this study is that

there are more people who have a positive sentiment towards COVID-19 with 122 more opinions. In this study, the results showed that the Naïve Bayes method had greater accuracy than the KNN. The level of accuracy obtained in the Naïve Bayes method is 63.21%, and KNN is 58.10%. So in this study the method used to classify is the Naïve Bayes method because it gets higher accuracy [3].

In research related to sentiment analysis on vaccination actions in an effort to overcome the Covid-19 pandemic. In this study, an accuracy of 85.59% was obtained using the Naïve Bayes method and the SVM method obtained an accuracy of 84.41%. In this study, the results showed that the Naïve Bayes method obtained greater accuracy than SVM in the case of vaccination with data that tended to be positive. The data used per keyword, namely the keyword sinovac vaccine 253 data and the red and white vaccine 253 data [4].

In this study, we will create a model that can see the trend of twitter users sentiment towards keywords, namely the covid-19 booster vaccination, on January 1, 2022 to July 30, 2022, with grouping negative, positive and neutral opinions. This work attempts to build a model to assess public opinion about the Covid-19 Booster Vaccination by using the Lexicon Based technique to identify sentiment in tweet data. Nave Bayes and logistic regression are classification techniques used in this study.

## 2. METHODOLOGY

The research methodology is a step-by-step will do in creating and completing study. The following is a methodology diagram on this research.

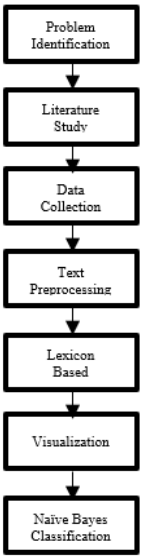


Figure 1. Research Stages

### 2.1 Problem Identification

The initial stage in this research is to define the issue of how to use the Lexicon Based technique to conduct sentiment analysis on Twitter social media. Any message, comment, or opinion can be posted on Twitter by a user. However, because Twitter only allows 280 characters for writing, users often employ acronyms in their posts. Finding sentiment in Twitter data is difficult in and of itself. To do that, we need a method that can choose slang phrases and then turn them into words that are more appropriate for sentiment analysis.

### 2.2 Literature Study

Currently, the process of looking for, analyzing, and applying various types of literature in the form of books, journals, papers, e-books, or other literature linked to the Lexicon Based approach is underway, particularly those utilized for sentiment analysis.

### 2.3 Data Collection

The Indonesian Twitter dataset was currently gathered and looked up. The information gathered is comprised of Twitter community tweets, which include messages, comments, or opinions related to Covid-19 Indonesia. From Januari to August 2022, tweet data was collected.

Using the Python programming language and the snscrape library, the data is obtained using Twitter scraping.

### 2.4 Text Preprocessing

There is a need for a procedure to eliminate noise from the tweet data that has been obtained from Twitter in order for the sentiment analysis process to become more accurate and be used generally. The following figure depicts how text preparation unfolded in this study:

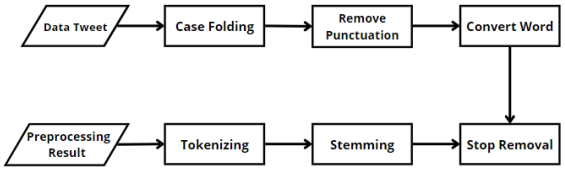


Figure 2. Text Preprocessing

### 2.5 Lexicon Based

The next step in the preprocessing process is classification using the lexical technique after the data has been cleaned. First, the tweets are translated into English, and then the words in the dataset are checked by weighting the words with the Lexicon lexicon with the use of the Vader Sentiment library.

### 2.6 Visualization

The next step is to display the outcomes of the sentiment analysis classification in the form of bar

charts or word clouds after the data classification stage using the Lexicon approach has been finished.

### 2.7 Naïve Bayes Classification

The Nave Bayes algorithm is used as the primary technique for categorization. This technique can forecast the likelihood that a set of data will fall into a particular class. As a method for comparison, logistic regression can explain the connection between explanatory variables and categorical or binary responses. If the data utilized have been weighted, as they were in this study using TF/IDF, the classification process can be completed.

## 3. RESULT AND DISCUSSION

This chapter summarizes the findings of a study that used the Lexicon Based technique in accordance with the steps outlined in the preceding chapter to conduct community sentiment analysis on Twitter regarding the Covid-19 issue.

### 3.1. Stages of Data Collection

Twitter data from January 1, 2021, to August 30, 2022, was the source of the data for this study. 607 tweets were retrieved when the scraping procedure was completed. The code for scraping data can be seen in the following image.

```
import snscrape.modules.twitter as sntwitter
import pandas as pd
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession

# Creating list to append tweet data to
tweets_list = []

# Using TwitterSearchScraper to scrape data and append tweets to list
for i,tweet in enumerate(sntwitter.TwitterSearchScraper
  ("vaksinasi booster covid-19 indonesia since:2022-01-01 until:2022-07-30 ").get_items()):
    if i>2000:
        break
    tweets_list.append([tweet.date, tweet.username, tweet.id, tweet.content])

# Creating a dataframe from the tweets list above
tweets_df = pd.DataFrame(tweets_list, columns = ['Datetime', 'Username', 'Tweet Id', 'Text'])
```

Figure 3. Twitter Data Scraping Code

Figure 3 is a code to get twitter data by scraping where there are username, datetime, tweet id data, and the text content of the tweet.

### 3.2. Stages of Text Preprocessing

The text preprocessing stage is carried out because the crawled data from Twitter contains a lot of noise, such as data duplication because messages are retweeted (RT) or shared by Twitter users, contain information that is not useful for the sentiment classification process such as hashtags (#), mentions (@).), links (http://www), punctuation marks, numbers, use of slang or abbreviations.

### 3.3. Stages of Case Folding

The case folding stage aims to change all the letters in the document into the same form into lower case or lower case letters. Data that has been stored in a csv file will be loaded and then stored in a variable. Then every tweet on the data will be carried out a case folding process.

### 3.4. Stages of Remove Punctuation

The remove punctuation stage aims to remove characters that do not have an influence on the sentiment classification process after the case folding process is carried out, such as removing commas (,), periods (.), hashtags (#), mentions (@), links, characters other than letters. and numbers. In addition, this stage also cleans up duplication of data originating from retweeting (RT) twitter users and removes words that are considered unnecessary for this research such as the words "covid" and "Indonesia" because these words are keywords in data search.

### 3.5. Stages of Convert Word

At this stage, the process of normalizing inappropriate words to the standard form according to the Big Indonesian Dictionary (KBBI). This stage is done because on Twitter, users are limited to only 280 characters. This makes users write tweets with words in the form of abbreviations or non-standard words so that the word needs to be changed into its standard form. Abbreviation words such as "bgt", "yg", "ttg", "sdh" and so on will be changed to their normal form. The source of the slang dictionary is obtained from the GitHub repository (GitHub - ramaprakoso) with the addition of several words that are considered not yet in the library. However, in practice, there are still words that cannot be converted because in some tweet data there are typo sentences or typos that make the words difficult to normalize.

### 3.6. Stages of Stop Removal

The stopword removal stage is carried out to remove words that have no effect in the sentiment classification process such as time, link, and others [5]. These words are deleted based on the words that have been prepared in the stopword dictionary. The source of the stopword dictionary used is taken from the GitHub repository [6].

### 3.7. Stages of Stemming

The stemming stage aims to return the word to its basic form. In this study, the Sastrawi library will be used to carry out the stemming process. Sastrawi is a library in the python programming language built with the NA. algorithm.

### 3.8. Stages of Tokenizing

The last stage of the preprocessing process is tokenizing, where the text data after cleaning is then broken down into tokens based on the delimiter, namely space. The results of this tokenizing process

will then be classified using the Lexicon dictionary with Vader Sentiment.

### 3.9. Stages of Lexicon Based

After all the preprocessing stages have been completed, a total of 588 clean data are obtained which are ready for the sentiment classification stage using the lexicon dictionary. This stage plays an important role in classification. Because this study uses an approach at the word level, where the data that is processed is the word to obtain a sentiment score using vader sentiment.

### 3.10. Stages of Data Visualization

The next stage is visualizing the results of the classification of the Lexicon Based method for sentiment analysis on Twitter data for positive, negative, and neutral classes. The results of the visualization in the bar chart are presented in the following figure.

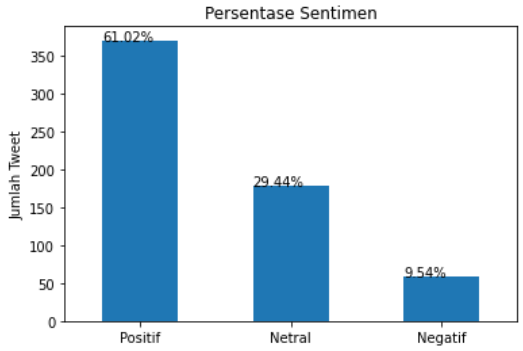


Figure 4. Visualization of Sentiment

From the picture above, it can be seen that public opinion on Twitter has a higher frequency of opinions with positive sentiments.

Sentiment classification results using the Lexicon Based method by calculating the weight of each word using Vader Sentiment on Twitter public opinion about the Covid-19 booster vaccination in Indonesia from January 2022 to August 2022. From a total of 588 twitter data shows that the percentage of public opinion for the sentiment class positive by 61.02% and public opinion for the neutral sentiment class of 29.44%. The remaining 9.54% of public opinion is on negative sentiment.

Then a visualization in the form of a wordcloud that shows words that often appear in the topic of Indonesia's Covid-19 booster vaccination in the positive sentiment class shown in the following image.


Figure 5. WordCloud On Positive Sentiment

Figure 5 shows a visualization of the word output that appears the most on positive sentiment related to the covid-19 booster vaccination. In the picture it can be seen that the words that appear most often are the words "booster", "covid".

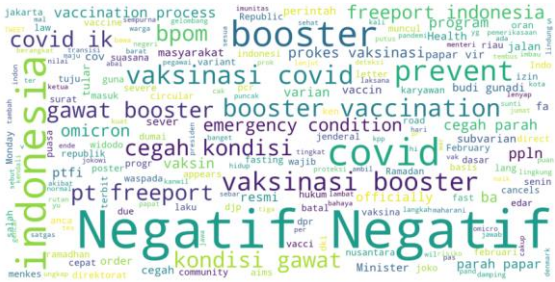


Figure 6. WordCloud On Negative Sentiment

Figure 6 shows a visualization of the word output that appears the most on negative sentiment related to the covid-19 booster vaccination.

Based on the results that have been carried out on research on community sentiment analysis on Twitter related to the Covid-19 booster vaccination in Indonesia using the Lexicon Based method by calculating the weight of each word with vader sentiment, from 588 twitter data shows the results of public opinion with a positive sentiment category of 61.02%, sentiment opinion in the negative category is 9.54%, and sentiment opinion is in the neutral category at 29.44%. So it can be concluded from the results of the sentiment analysis that public opinion on twitter who believes in the Covid-19 booster vaccination is actually still quite high compared to people who have negative opinions regarding the Covid-19 booster vaccination.

## 4. RESULT

The Naïve Bayes Classification method is used to create a model that can be determines whether a tweet falls into one of the three sentiment classes. This model utilizes the value of probability of each tweet to fall into a predetermined class category. This method obtains an accuracy level of 0.70 with details of precision, f1-score and recall in Figure 7.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negatif | 1.00 | 0.08 | 0.15 | 12 |
| Netral | 0.74 | 0.39 | 0.51 | 36 |
| Positif | 0.69 | 0.95 | 0.80 | 74 |
| accuracy | | | 0.70 | 122 |
| macro avg | 0.81 | 0.47 | 0.49 | 122 |
| weighted avg | 0.73 | 0.70 | 0.65 | 122 |

Fig 7. Precision, F1-Score and Recall Details

Figure 7 shows an accuracy level of 0.7 with details on precision, f1-score and recall.

The Logistic Regression Classification method is used to create a model that can determine whether a tweet fall into one of the three classes of sentiment. This model gets an accuracy rate of 0.72.


Fig 8. Logistic Regression Accuracy

Figure 8 is the code used in obtaining the accuracy level with the Logistic Regression comparison method, where the accuracy rate is 72%.

## 5. DISCUSSION

A naive Bayesian approach was employed to do the sentiment analysis, and lexicon-based features were used for the classification of individual words, with the sentiment value of each word being weighted according to its lexicon-based features. Several steps are taken in the process, beginning with the preparation of the text, followed by the weighting of terms, then the training of latent class models with the weighting of lexical characteristics, and finally the testing with latent class models involving the weighting of these features. There is a marked difference between the tells produced by the two methods, and a comparison of the tells obtained by those two methods reveals that the first, logistic regression, produces greater accuracy than the latter, naive bayes, which produces accurate results only 70% of the time. It is, therefore, reasonable to conclude that the implementation of the Naive Bayes approach with Lexicon-Based Features would not lead to a substantial improvement in outcomes.

## 6. CONCLUSION

Based on the test, it shows that the Logistic Regression model is better than Naïve Bayes in the case of sentiment analysis on the Indonesian Covid-19 booster vaccination. This is indicated by the level of accuracy obtained with Logistic Regression of 0.72 or 72% and Naïve Bayes of 0.70 or 70%.

The trend of sentiment from January 1 to August 30, 2022 is dominated by positive sentiment of 61.02%. Then neutral sentiment of 29.44% which is even greater than negative sentiment. Where negative sentiment is only 9.54%.

Suggestions for further research is to add the number of datasets and add other comparison methods.

## REFERENCES

[1] A. Prabhat and V. Khullar, “Sentiment classification on big data using Naïve bayes and logistic regression,” *2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017*, no. March, 2017, doi: 10.1109/ICCCI.2017.8117734.

[2] Y. Nooryuda Prasetya and D. Winarso, “Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twiter Terhadap Isu Covid-19,” *J. Fasilkom*, vol. 11, no. 2, pp. 97–103, 2021.

[3] M. Syarifuddinn, “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, doi: 10.33480/inti.v15i1.1347.

[4] B. Laurensz and Eko Sediyono, “Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 118–123, 2021, doi: 10.22146/jnteti.v10i2.1421.

[5] K. Jahanbin, V. Rahmanian, N. Sharifi, and F. Rahmanian, “Sentiment Analysis and Opinion Mining about COVID-19 vaccines of Twitter Data,” *Pakistan J. Med. \& Heal. Sci.*, vol. 15, no. 1, pp. 694–695, 2021.

[6] A. Syakur, “Implementasi Metode Lexicon Base Untuk Analisis Sentimen Kebijakan Pemerintah Dalam Pencegahan Penyebaran Virus Corona Covid-19 Pada Twitter,” *J. Ilm. Inform. Komput.*, vol. 26, no. 3, pp. 247–260, 2021, doi: 10.35760/ik.2021.v26i3.4720.

[7] G. N. Aulia and E. Patriya, “Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019,” *J. Ilm. Inform. Komput.*, vol. 24, no. 2, pp. 140–153, 2019, doi: 10.35760/ik.2019.v24i2.2369.

[8] D. Muriyatmoko, T. Harmini, and M. K. Ardiansyah, “Sentiment Analysis Covid-19 Vaccination on Twitter Social Media Using Naïve Bayes Method,” *Procedia Eng. Life Sci.*, vol. 2, no. 2, 2021, doi: 10.21070/pels.v2i0.1144.

[9] A. Sasmito Aribowo, “Analisis Sentimen Publik pada Program Kesehatan Masyarakat menggunakan Twitter Opinion Mining,” *Semin. Nas. Inform. Medis*, vol. 0, no. 0, pp.

17–23, 2018, [Online]. Available: https://journal.uii.ac.id/snimed/article/view/11877.

[10] A. Muzaki and A. Witanti, "Sentiment Analysis of the Community in the Twitter To the 2020 Election in Pandemic Covid-19 By Method Naive Bayes Classifier," *J. Tek. Inform.*, vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.

[11] S. Sudianto, P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "Comparison Of Random Forest And Support Vector Machine Methods On Twitter Sentiment Analysis ( Case Study : Internet Selebgram Rachel Vennya Escape From Quarantine ) Perbandingan Metode Random Forest Dan Support Vector Machine Pada Analisis Sentimen Twitt," *Jutif*, vol. 3, no. 1, pp. 141–145, 2022.

[12] R. D. Septiana, A. B. Susanto, and T. Tukiyat, "Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection Chi-Squared Statistic dan Particle Swarm Optimization," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 5, no. 1, pp. 49–56, 2021, doi: 10.47970/siskom-kb.v5i1.228.

[13] M. Kartika, S. Saepudin, and D. Gustian, "Analisis Sentimen Dampak Covid-19 Terhadap Pembatalan Keberangkatan Ibadah Haji Pada Tahun 2020," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 964–972, 2021.

[14] A. F. Hidayatullah, S. Cahyaningtyas, and A. M. Hakim, "Sentiment Analysis on Twitter using Neural Network: Indonesian Presidential Election 2019 Dataset," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012001, 2021, doi: 10.1088/1757-899x/1077/1/012001.

[15] K. Hikmah, A. C. Fauzan, and Harliana, "Sentiment Analysis of Vaccine Booster during Covid-19- Indonesian Netizen Perspective Based on Twitter Dataset," no. 22, pp. 102–106, 2022, [Online]. Available: http://ojs.stmikpringsewu.ac.id/index.php/jtksi.

[16] N. A. Alkhaldi, Y. Asiri, A. M. Mashraqi, H. T. Halawani, S. Abdel-Khalek, and R. F. Mansour, "Leveraging Tweets for Artificial Intelligence Driven Sentiment Analysis on the COVID-19 Pandemic," *Healthcare*, vol. 10, no. 5, p. 910, 2022, doi: 10.3390/healthcare10050910.