Vol. 07, No. 01, January 2023, pp. 1-18

INTERNATIONAL JOURNAL OF INNOVATION IN ENTERPRISE SYSTEM

Journal homepage: https://ijies.sie.telkomuniversity.ac.id
e-ISSN: 2580-3050, DOI: https://doi.org/10.25124/ijies.v7i01.171

Application of the Naïve Bayes Classifier Algorithm to Analyze Sentiment for the Covid-19 Vaccine on Twitter in Jakarta

Ire Puspa Wardhani[1], Yudi Irawan Chandra[2*], Ferri Yusra[3]

[1]Information System
STMIK Jakarta STI&K, Jakarta, Indonesia
irepuspa@gmail.com

[2]Information System
STMIK Jakarta STI&K, Jakarta, Indonesia
yirawanc@gmail.com

[3]Information System
STMIK Jakarta STI&K, Jakarta, Indonesia
ferri.yusra@jak-stik.ac.id

*yirawanc@gmail.com

ARTICLE INFO

Article history:
Received 09 May 2022
Accepted 25 November 2022
Published 31 January 2023

Keywords:
Sentiment Analysis; Text Pre-Processing; Naïve Bayes Classifier; TF-IDF; Twitter.

ABSTRACT

The epidemic of a new disease caused by the coronavirus (2019-nCoV), commonly referred to as COVID-19, has been declared a global virus epidemic by the World Health Organization (WHO). President Joko Widodo has officially ratified Presidential Decree No. 99 of 2020 concerning the provision of vaccines and the implementation of vaccination activities. Twitter is a social media platform that allows users to share information and opinions directly with fellow users. Tweets given can be in any form, either positively or negatively, so one of the methods used is sentiment analysis. Sentiment analysis helps determine an opinion or comment on an issue, whether the response is positive or negative. The Naïve Bayes algorithm is used in sentiment analysis because it is suitable for tweets or text data that is not too long or short text. The initial stage of sentiment analysis is text pre-processing which consists of Cleaning, case folding, tokenizing, and stopword removal. Then the data is labeled manually. The analysis results are visualized as bar charts, pie charts, and word clouds. Then the word weighting is carried out using the term frequency-inverse document (TF-IDF), and classification is carried out using the Naïve Bayes classifier. From the test results, the accuracy value of the confusion matrix is 82% from 2600 tweet data with 80% training data composition and 20% test data.

1. INTRODUCTION

The epidemic of a new variant disease caused by the coronavirus (2019-nCoV) or often referred to as COVID-19, was inaugurated as a global (global) virus epidemic by the world health organization or commonly known as the World Health Organization (WHO) on March 11, 2020 [1]. The spread of this disease is speedy, so in a short time, many victims have fallen, and the virus has been identified. However, this variant of the virus appeared in December, around the end of 2019 in China, precisely in the city of Wuhan.

This disease virus has entered and spread to all corners of the world's citizens. So that the total number of people detected on April 13, 2021, was approximately 136,115,434 cases and 2,936,916 people died [2]. President Joko Widodo announced and confirmed that the COVID-19 virus disease had entered Indonesia in Indonesia. For the first identified case in the Depok area, West Java, 2 Indonesian citizens tested positive for COVID-19 [3]. From the identification of these cases, the increasing number of cases of people in Indonesia exposed to the Covid-19 virus continued to rise, so that on April 11, 2021, the number of cases of the Covid-19 virus was 1,566,995 confirmed positive cases, 1,414,507 people were identified as cured and 42,530 people died [4].

With the impact of the coronavirus or covid-19, all countries are competing in manufacturing vaccines. Almost all countries are interested in research to produce vaccines. As a result of human creation, vaccines take a lot of time and money to manufacture. Seeing that the level of cases produced by the Covid-19 virus is very high, the vaccine's role is needed in protecting the human body from the dangers of this Covid-19 virus. Also, vaccines can reduce the rate of spread of this disease, breaking the chain of spread and reducing the rate of death and positive cases. In responding to this, the Indonesian government has participated and is also active in the plan to realize the procurement of vaccines. Mr. President of Indonesia, Joko Widodo, on October 5, 2020, officially ratified Presidential Decree No. 99 of 2020 concerning the provision of vaccines and the implementation of vaccination activities [5]. Twitter social media allows users to access information, share, participate in giving an opinion or opinion, and create content to interact with other users on their social media platforms. So active Twitter users reached 350 billion active users on January 25, 2020 [6]. *Twitter* is a social media platform that allows users to share information directly with fellow users. The information provided on Twitter is often referred to as an opinion sentence (Tweet), where this tweet has 140 characters [7].

A theme of the discussion that has been the talk of Twitter users or Indonesian citizens is the COVID-19 vaccine, which the Indonesian people are busy discussing. With so many Twitter users expressing opinions about the COVID-19 vaccine, this tweet can be used to find information. Tweets given by Twitter users can be in the form of anything, either positively or negatively, so the suitable method is sentiment analysis [8]. Sentiment analysis helps determine an opinion or comment on a problem, whether it tends to have a positive or negative view or response with various approaches. Sentiment analysis from Twitter should focus on classifying tweet data used [9]. Agus Tiyansyah Syam, in his research, classification is a stage of producing a model or function of a function that can describe and select a plan or part of the data itself [10]. Tweets from users will be classified with machine learning methods to classify a tweet as positive or negative. In a brief period, it produces many tweets that cannot be classified manually from the social media platform, namely Twitter, which takes a very long time. It can be challenging to classify the tweet text's sentiments by humans manually. Researchers will implement machine learning with this algorithm: supervised learning, namely the Naïve Bayes Classifier [11]. Cahyono, in his research, the Naïve Bayes algorithm is used in sentiment analysis because it is suitable for tweets or text data that is not too long or short text. Moreover, fast in producing a model that can predict and produce a new method of studying and understanding data [12]. The formulation of the problem in this research is how to apply the Naïve Bayes Classifier Algorithm to analyze Twitter user sentiment towards tweets on Twitter social media, how to visualize data on positive or negative sentiments, and know the accuracy of the Naïve Bayes Classifier Algorithm.

The limitations of the problem in this research are how to apply the Naïve Bayes Classifier Algorithm to analyze Twitter user sentiment towards tweets on Twitter social media; the dataset used is tweet data with the "vaksin covid-19 " search topic. The dataset used is tweeting with the search topic "vaksin covid-19". The tweet used is the type of tweet in Indonesian with 5053 tweets of data at the time of data scrapping. Data was taken randomly from April 09, 2021 – to April 18, 2021, and there are two classifications of tweet data, namely positive or negative sentiment classes. The objectives of this research are to apply the Naïve Bayes Classifier algorithm to classify and analyze sentiment on the topic "covid-19 vaccine" on Twitter social media, know the sentiment of Twitter users in Indonesia towards the procurement of the covid-19 vaccine and visualize sentiment tweet data and knowing the level of accuracy with the Naïve Bayes Classifier algorithm in classifying sentiment on status or tweets with the topic "vaksin covid-19".

2. METHOD

The flow of research methods is as follows:

1. Problem Identification
   At this stage, problems are identified so that an application is needed to be built and collect data. Data collection methods can be done by literature review. A literature study aims to obtain data and directions from various sources from books, articles, journals, and others related to research as a reference in compiling this research.

2. Problem Analysis
   At this stage, analyze the problems and the system requirements to understand the problems and provide a solution you want to apply to the problem.

3. Data Collection
   In this phase, collecting tweet data is carried out. This stage aims to store tweets. Then it will be likened to training and test data. This phase will manage and collect tweets using the Twitter scrapping tool library. The scrapping tool is the Twint Library: (https://github.com.twintproject/twint). This tweet was taken as many as 5053 tweet data with the search topic "vaksin covid-19" and saved it into a .csv file.

4. Text Pre-Processing
   At this stage, the tweets that have been collected are then pre-processed text to clean the tweet data and discard various unused characters or text [13]. The pre-processing stages used are:
   a. Cleaning
   b. Case Folding
   c. Tokenizing
   d. Stopword Removal

5. Data Labeling
   At this stage, tweets that have been pre-processed will be given manual data labeling, which consists of two classes of data labeling, namely positive or negative.

6. Data Visualization
   At this stage, the tweets labeled in the next stage are visualized to see how many tweets have been labeled into the sentiment class and the most frequently spoken words. The data visualization used includes:
   a. Pie Chart
   b. Bar Chart
   c. WordCloud

7. TF-IDF Word Weighting
   At this stage, the tweets that have been labeled are weighted using TF-IDF to give weight to the number of word frequencies [14].

8. Naïve Bayes Classifier
   At this stage, tweets that have been weighted TF-IDF will be classified into positive or negative sentiments using the Naïve Bayes classifier classification with multinomial equations.

9. Model
   At this stage, tweets that have been word-weighted, namely TF-IDF and Naïve Bayes classifier classification, will produce a classification model and a weighting model. Then the model is stored for weighting and classification of new data that does not yet have a sentiment label so that the data can be classified directly into positive or negative sentiment.

10. Implementation and Testing
    In this phase, the program's implementation using the python programming language is carried out from the data collection stage to creating a confusion matrix. Then the classification model is then tested by calculating the accuracy of the confusion matrix so that the accuracy and performance of a model in classifying data are obtained.

2.1. Data collection

The tweet data or dataset used in this research is sourced from Twitter. Twitter data retrieval using the Twitter scraping tool. The topic search keyword in this research is "vaksin covid-19", with the initial data at the time of scrapping taken from as many as 5053 tweets and in the Indonesian language taken from April 09, 2021 - to April 18, 2021. The data that has been taken is stored in a file in CSV format [15].

Then the text pre-processing process of tweet data was carried out so that 3694 data from 5053 data were taken at the time of scrapping, then deleted data that did not match, such as the presence of data that was not Indonesian, deleted the same data and deleted data that was empty. So that the data obtained as much as 2600 data, and manual labeling was carried out. The data obtained consisted of 2063 data with positive labels and 537 data with negative labels.

2.2. Text Pre-processing

After collecting tweet data, the pre-processing stage is carried out to process data that is still not appropriate before other stages. The data cleaning to uniform the word form of a sentence and limit the amount of new text from a collection of tweets. In the pre-processing stage, several processes will be carried out, namely:

*yirawanc@gmail.com

1. Cleaning
   In this phase, it cleans the attributes of sentences with no connection to words that already exist in a tweet. Such as attributes, URLs, hashtags, mentions of other users, retweets, punctuation characters, and removing excess spaces and emoticons.
2. Case Folding
   The process is carried out at this stage to make all letters in the tweet data lowercase.
3. Tokenizing
   Tokenizing separates a word from a sentence into a word that composes the sentence.
4. Stopword Removal
   Stopword Removal is a step that removes and deletes words, in this case, a tweet that is not important to use. Examples of a conjunction such as "or", "to", "at", "will", "not", "it", "want", and others.

2.3. Manual Labeling

Labeling all tweet data with a predetermined category class. The author himself provides manual labeling. The dataset used is divided into two categories. The two categories used to classify tweets are:
1. Positive
2. Negative

2.4. Data Visualization

After collecting data, text pre-processing, and labeling the data, the data is carried out in the visualization stage to see how many positive and negative sentiments are and what words are most often discussed for each sentiment. The data visualization is a bar chart, pie chart, and word cloud.

2.5. Word Weighting with TF-IDF

After completing the text pre-processing and manual labeling stages, the word weighting stage is carried out. The data that has been cleaned or (pre-processed) subsequently changes words into a vector or numeric form. The following are the stages of word weighting with TF-IDF: Counts the number of times a word appears in a document. Perform calculations from the IDF. Calculates the weight against a (TF-IDF) by multiplying the occurrence of the word (TF) by the word weight (IDF). TF-IDF weights obtained

2.6. Naïve Bayes Classifier

At this stage, the data is classified by text pre-processing, manual labeling, and word weighting. The following is the calculation of the Naïve Bayes Classifier classification. Calculate the value of the prior probability of each class of data. Calculates the conditional probability value for each existing word. Calculates the posterior probability value of each class's probability and each word's probability. Determine the highest value from each posterior probability class so that the data results are included in the positive or negative sentiment class.

3. RESULT AND DISCUSSION

3.1. Text Pre-Processing

The pre-processing stage aims to eliminate unused parts or new tweet words to get quality data for execution. The following pre-processing stages are carried out consisting of 4 stages, namely:
1. Cleaning
   This stage removes symbols or attributes that have nothing to do with the tweets' words, such as uniform resource locator, symbols, hashtags mentioning other users, retweets, punctuation characters, and removing extra spaces and emoticons. Table 1 is the result of the cleaning stages.

Table 1 - Results of Cleaning tweet data.

No	Before *Cleaning*	After *Cleaning*
1.	Alhamdulillah.. sebagian mitra @GrabID kota Blitar sdh vaksin Covid-19 😍 https://t.co/0RyPTqNiGo	Alhamdulillah sebagian mitra kota Blitar sdh vaksin Covid
2.	Alhamdulillah vaksin covid 19 yg ke dua. Sdh selwsai https://t.co/EHCdMSTsCP	Alhamdulillah vaksin covid yg ke dua Sdh selwsai
3.	Muncul kasus pembekuan darah, vaksin COVID-19 Janssen diselidiki UE https://t.co/UQqv2PB86F https://t.co/43IhXh95VS	Muncul kasus pembekuan darah vaksin COVID Janssen diselidiki UE

4.	Vaksin Covid-19 AstraZeneca Bermasalah, Eropa Temukan Kasus Pembekuan Darah https://t.co/tkW6CpgmpO	Vaksin Covid AstraZeneca Bermasalah Eropa Temukan Kasus Pembekuan Darah
5.	Alhamdulillah vaksin covid 19 tahap pertama berjalan lancar.... https://t.co/PWUF8iIdIK	Alhamdulillah vaksin covid tahap pertama berjalan lancar
6.	BPOM Beri Warning Pembekuan Darah pada Vaksin COVID-19 AstraZeneca https://t.co/5sVRJEn78t	BPOM Beri Warning Pembekuan Darah pada Vaksin COVID AstraZeneca

2. Case Folding
   After the cleaning stage, the next step is to carry out the case folding stage. This stage is done by homogenizing a sentence into lowercase sentences. Table 2 is the result of the case folding stages.

Table 2 - Results of cleaning tweet data

No	Before *Case Folding*	After *Case Folding*
1.	Alhamdulillah sebagian mitra kota Blitar sdh vaksin Covid	alhamdulillah sebagian mitra kota blitar sdh vaksin covid
2.	Alhamdulillah vaksin covid yg ke dua Sdh selwsai	alhamdulillah vaksin covid yg ke dua sdh selwsai
3.	Muncul kasus pembekuan darah vaksin COVID Janssen diselidiki UE	muncul kasus pembekuan darah vaksin covid janssen diselidiki ue
4.	Vaksin Covid AstraZeneca Bermasalah Eropa Temukan Kasus Pembekuan Darah	vaksin covid astrazeneca bermasalah eropa temukan kasus pembekuan darah
5.	Alhamdulillah vaksin covid tahap pertama berjalan lancar	alhamdulillah vaksin covid tahap pertama berjalan lancar
6.	BPOM Beri Warning Pembekuan Darah pada Vaksin COVID AstraZeneca	bpom beri warning pembekuan darah pada vaksin covid astrazeneca

3. Tokenizing
   Then do the tokenizing stage; in this process, a sentence, in this case, the tweet, is given a separator and produces a word from the tweet that composes the tweet. Table 3 is the result of the tokenizing stage.

Table 3 - Results of Tokenizing tweet data

No	Before *Tokenizing*	After *Tokenizing*
1.	alhamdulillah sebagian mitra kota blitar sdh vaksin covid	'alhamdulillah', 'sebagian', 'mitra', 'kota', 'blitar', 'sdh', 'vaksin', 'covid'
2.	alhamdulillah vaksin covid yg ke dua sdh selwsai	'alhamdulillah', 'vaksin', 'covid', 'yg', 'ke', 'dua', 'sdh', 'selwsai'
3.	muncul kasus pembekuan darah vaksin covid janssen diselidiki ue	'muncul', 'kasus', 'pembekuan', 'darah', 'vaksin', 'covid', 'janssen', 'diselidiki', 'ue'
4.	vaksin covid astrazeneca bermasalah eropa temukan kasus pembekuan darah	'vaksin', 'covid', 'astrazeneca', 'bermasalah', 'eropa', 'temukan', 'kasus', 'pembekuan', 'darah'
5.	alhamdulillah vaksin covid tahap pertama berjalan lancar	'alhamdulillah', 'vaksin', 'covid', 'tahap', 'pertama', 'berjalan', 'lancar'
6.	bpom beri warning pembekuan darah pada vaksin covid astrazeneca	'bpom', 'beri', 'warning', 'pembekuan', 'darah', 'pada', 'vaksin', 'covid', 'astrazeneca'

4. Stopword Removal
   After performing the tokenizing stage, the next step is to perform a stopword removal stage that removes tweets containing words that have nothing to do with it. We can also create a separate set of stopwords (stoplist). Table 4 is the result of the stopword removal stages.

Table 4 - Results of Stopword Removal of tweet data

No	Before *Stopword Removal*	After *Stopword Removal*
1.	'alhamdulillah', 'sebagian', 'mitra', 'kota', 'blitar', 'sdh', 'vaksin', 'covid'	alhamdulillah mitra kota blitar sdh vaksin covid
2.	'alhamdulillah', 'vaksin', 'covid', 'yg', 'ke', 'dua', 'sdh', 'selwsai'	alhamdulillah vaksin covid yg sdh selwsai
3.	'muncul', 'kasus', 'pembekuan', 'darah', 'vaksin', 'covid', 'janssen', 'diselidiki', 'ue'	muncul pembekuan darah vaksin covid janssen diselidiki ue
4.	'vaksin', 'covid', 'astrazeneca', 'bermasalah', 'eropa', 'temukan', 'kasus', 'pembekuan', 'darah'	vaksin covid astrazeneca bermasalah eropa temukan pembekuan darah
5.	'alhamdulillah', 'vaksin', 'covid', 'tahap', 'pertama', 'berjalan', 'lancar'	alhamdulillah vaksin covid tahap berjalan lancar
6.	'bpom', 'beri', 'warning', 'pembekuan', 'darah', 'pada', 'vaksin', 'covid', 'astrazeneca'	bpom warning pembekuan darah vaksin covid astrazeneca

3.2. Data Manual Labeling

After the text pre-processing stage has been carried out on the data, the data will be labeled manually. Tweet data is labeled for training needs in this study; the data will be used as a reference in the data classification process for data that does not have a label. The author himself does labeling. Table 5 below is the result of manual data labeling.

Table 5 - Results of Data Labeling

No	Data	Class
1.	alhamdulillah mitra kota blitar sdh vaksin covid	Positif
2.	alhamdulillah vaksin covid yg sdh selwsai	Positif
3.	muncul pembekuan darah vaksin covid janssen diselidiki ue	Negatif
4.	vaksin covid astrazeneca bermasalah eropa temukan pembekuan darah	Negatif

3.3. Manual Calculation On Data

At this stage, a training and test data calculations simulation is carried out using the TF-IDF word weighting first and then classifying the data using the Naïve Bayes Classifier algorithm. Then select some training data that has been done in the pre-processing and data labeling stages for TF-IDF calculations and classification. The training data used can be seen in table 6 and test data in table 7.

Table 6 - Training Data

No	Data	Kelas
1.	alhamdulillah mitra kota blitar sdh vaksin covid	Positif
2.	alhamdulillah vaksin covid yg sdh selwsai	Positif
3.	muncul pembekuan darah vaksin covid janssen diselidiki ue	Negatif
4.	vaksin covid astrazeneca bermasalah eropa temukan pembekuan darah	Negatif

Table 7 - Test Data

No	Data	Kelas
1.	alhamdulillah vaksin covid tahap berjalan lancar	Belum diketahui
2.	bpom warning pembekuan darah vaksin covid astrazeneca	Belum diketahui

3.4. TF-IDF Word Weighting

The tweet words generated from the previous stage are given a weight in the word weighting process. The author uses word weighting using TF-IDF. Before getting the TF-IDF value, first get the TF, DF, and IDF values.

1. In Table 8, the TF value with the word "alhamdulillah" in data 1 and the data is one each because "alhamdulillah" is only one word in data 1 and 2. So the DF value for the word "alhamdulillah" is two because from the first to the second data, the word "alhamdulillah" appears only once in data 1 and data 2, data 3 and data 4 do not find the word "Alhamdulillah." After the TF and IDF values have been known, the next step is calculating the IDF value. Table 8 shows the IDF word weighting result based on the training data in table 6.

Table 8 - IDF Word Weighting Results on the overall data

No	Words	TF				D	DF	IDF
		Data1	Data2	Data3	Data4			$log\left(\frac{d}{df_t}\right)$
1	alhamdulillah	1	1			4	2	log(4/2) = 0,301
2	mitra	1				4	1	log(4/1) = 0,602
3	kota	1				4	1	log(4/1) = 0,602
4	blitar	1				4	1	log(4/1) = 0,602
5	sdh	1	1			4	2	log(4/2) = 0,301
6	vaksin	1	1	1	1	4	4	log(4/4) = 0
7	covid	1	1	1	1	4	4	log(4/4) = 0
8	yg		1			4	1	log(4/1) = 0,602
9	selwsai		1			4	1	log(4/1) = 0,602
10	muncul			1		4	1	log(4/1) = 0,602
11	pembekuan			1	1	4	2	log(4/2) = 0,301
12	darah			1	1	4	2	log(4/2) = 0,301
13	janssen			1		4	1	log(4/1) = 0,602
14	diselidiki			1		4	1	log(4/1) = 0,602
15	ue			1		4	1	log(4/1) = 0,602
16	astrazeneca				1	4	1	log(4/1) = 0,602
17	bermasalah				1	4	1	log(4/1) = 0,602
18	eropa				1	4	1	log(4/1) = 0,602
19	temukan				1	4	1	log(4/1) = 0,602
							Total	**9,03**

2. TF-IDF Word Weighting
   The tweet words generated from the previous stage are given a weight in the word weighting process. The author uses word weighting using TF-IDF. Before getting the TF-IDF value, first get the TF, DF, and IDF values.
   In Table 8, the TF value with the word "alhamdulillah" in data 1 and the data is one each because "alhamdulillah" is only one word in data 1 and 2. So the DF value for the word "alhamdulillah" is two because from the first to the second data, the word "alhamdulillah" appears only once in data 1 and data 2, data 3 and data 4 do not find the word "Alhamdulillah." After the TF and IDF values have been known, the next step is calculating the IDF value. Table 9 is the result of the IDF word weighting based on the training data in table 8:

Table 9 - Results of TF-IDF Class Positive Word Weighting

No	Words	DF	IDF	TF-IDF
1	alhamdulillah	2	0,301	2 * 0,301 = 0,602
2	mitra	1	0,602	1 * 0,602 = 0,602
3	kota	1	0,602	1 * 0,602 = 0,602
4	blitar	1	0,602	1 * 0,602 = 0,602
5	sdh	2	0,301	2 * 0,301 = 0,602
6	vaksin	2	0	2 * 0 = 0
7	covid	2	0	2 * 0 = 0
8	yg	1	0,602	1 * 0,602 = 0,602
9	selwsai	1	0,602	1 * 0,602 = 0,602
Total				**4,214**

3. Then calculate the TF-IDF value for harmful class data. Based on table 10, the word "freezing" has a DF value of 2 because the total occurrence of the word "freezing" is two times in data 3 and 4, based on table 8. Then take the IDF value of the word "freezing" based on table 8, worth "0.301". Then calculate the TF-IDF value so that the TF-IDF value with the word "freezing" is "0.602". In table 10 below are the results of the weighting of the hostile class TF-IDF words

Table 10 - Results of Negative Class TF-IDF Word Weighting

No	Words	DF	IDF	TF-IDF
1	muncul	1	0,602	1 * 0,602 = 0,602
2	pembekuan	2	0,301	2 * 0,301 = 0,602
3	darah	2	0,301	2 * 0,301 = 0,602
4	janssen	1	0,602	1 * 0,602 = 0,602
5	diselidiki	1	0,602	1 * 0,602 = 0,602
6	ue	1	0,602	1 * 0,602 = 0,602
7	astrazeneca	1	0,602	1 * 0,602 = 0,602
8	bermasalah	1	0,602	1 * 0,602 = 0,602
9	eropa	1	0,602	1 * 0,602 = 0,602
10	temukan	1	0,602	1 * 0,602 = 0,602
11	vaksin	1	0	1 * 0 = 0
12	covid	1	0	1 * 0 = 0
			Total	**6.02**

3.5. Naïve Bayes Classifier

After going through the TF-IDF stage, the word collection and the weighting of the TF-IDF values are processed by calculations using the classification of the Naïve Bayes classifier on the test data that has gone through the data cleaning or pre-processing stage. The test data used can be seen in Table 11 below.

Table 11. Classification Test Data

No	Data
1.	alhamdulillah vaksin covid tahap berjalan lancar
2.	bpom warning pembekuan darah vaksin covid astrazeneca

After pre-processing the test data, calculations are carried out to find the probability value for positive and negative classes using this study's formula. The following are the stages in the classification process :

1. Calculate the probability value of tweets for positive and negative labeled classes by calculating the prior probability in the following way:

   $P(positive) = \left(\frac{2}{4}\right) = 0{,}5$

   $P(negative) = \left(\frac{2}{4}\right) = 0{,}5$

   Information :
   a. Value 2 comes from the total number of positive and negative class data based on table 6
   b. The value of 4 comes from the total amount of data, in table 6 has 4 data.

2. After calculating the prior probability, then calculate each Conditional Probability Tweet of each positive and negative class on the test data, namely:
   a. Tweet with alhamdulillah, vaccine, covid, stage, running, smoothly. By calculating the conditional probability as shown in table 12 below:

Table 12 - Calculation of Positive Conditional Probability Test Data (a)

Calculation of Positive Conditional Probability Test Data
$P(alhamdulillah \mid positive) = \frac{0{,}602 + 1}{4{,}214 + 9{,}03} = \frac{1{,}602}{13{,}244} = 0{,}121$
$P(vaksin \mid positive) = \frac{0 + 1}{4{,}214 + 9{,}03} = \frac{1}{13{,}244} = 0{,}075$
$P(covid \mid positive) = \frac{0 + 1}{4{,}214 + 9{,}03} = \frac{1}{13{,}244} = 0{,}075$

$$P(tahap \mid positive) = \frac{0+1}{4{,}214 + 9{,}03} = \frac{1}{13{,}244} = 0{,}075$$

$$P(berjalan \mid positive) = \frac{0+1}{4{,}214 + 9{,}03} = \frac{1}{13{,}244} = 0{,}075$$

$$P(lancar \mid positive) = \frac{0+1}{4{,}214 + 9{,}03} = \frac{1}{13{,}244} = 0{,}075$$

Next, calculate the positive posterior probability on the test data (a) with the calculations in table 13 below:

Table 13 - Calculation of Positive Posterior Probability Test Data (a)

$P(positve \mid \begin{matrix} alhamdulillah\ vaksin\ covid\ tahap \\ berjalan\ lancar \end{matrix})$
$= (0{,}5 * 0{,}121 * 0{,}075 * 0{,}075 * 0{,}075 * 0{,}075 * 0{,}075)$
$= 0{,}0000001436$

Next, calculate the negative Conditional Probability of the test data (a) in the Tweet with the words alhamdulillah, vaccine, covid, stage, running, smoothly with the calculations in table 14:

Table 14 - Calculation of Negative Conditional Probability Test Data (a)

Calculation of Negative Conditional Probability Test Data
$P(alhamdulillah \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(vaksin \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(covid \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(tahap \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(berjalan \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(lancar \mid negative) = \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$

Next, calculate the negative posterior probability on the test data (a) with the calculations in table 15 below:

Table 15 - Calculation of Negative Posterior Probability Test Data (a)

$P(negative) \mid alhamdulillah\ vaksin\ covid\ tahap\ berjalan\ lancar)$
$= (0{,}5 * 0{,}066 * 0{,}066 * 0{,}066 * 0{,}066 * 0{,}066 * 0{,}066)$
$= 0{,}00000004133$

b. Tweet The word BPOM, warning, clotting, blood, vaccine, covid, and AstraZeneca. By calculating the conditional probability as shown in table 16 below:

Table 16 - Calculation of Positive Conditional Probability Test Data (b)

Calculation of Positive Conditional Probability Test Data	
$P(bpom \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(warning \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(pembekuan \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(darah \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(vaksin \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(covid \mid positive)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$
$P(astrazeneca \mid positif)$	$= \frac{0+1}{4{,}214 + 9{,}03}$
	$= \frac{1}{13{,}244} = 0{,}075$

Next, calculate the positive posterior probability on the test data (b) with the calculations in table 17 below:

Table 17 - Calculation of Positive Posterior Probability Test Data (b)

$P\left(positive \mid \begin{matrix} bpom\ warning\ pembekuan \\ darah\ vaksin\ covid\ astrazeneca \end{matrix}\right)$
$= (\ 0{,}5 * 0{,}075 * \ 0{,}075 * 0{,}075 * 0{,}075 * 0{,}075 * 0{,}075 * 0{,}075\ )$
$= 0{,}000000006674$

Then calculate the negative Conditional Probability of test data (b) in Tweets with the words BPOM, warning, clotting, blood, vaccine, covid, and AstraZeneca. With the calculations in table 18:

Table 18 - Calculation of Negative Conditional Probability Test Data (b)

Calculation of Negative Conditional Probability Test Data	
$P(bpom \mid negative)$	$= \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(warning \mid negative)$	$= \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(pembekuan \mid negative)$	$= \frac{0{,}602+1}{6{,}02 + 9{,}03} = \frac{1{,}602}{15{,}05} = 0{,}106$
$P(darah \mid negative)$	$= \frac{0{,}602+1}{6{,}02 + 9{,}03} = \frac{1{,}602}{15{,}05} = 0{,}106$
$P(vaksin \mid negative)$	$= \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(covid \mid negative)$	$= \frac{0+1}{6{,}02 + 9{,}03} = \frac{1}{15{,}05} = 0{,}066$
$P(astrazeneca \mid negative)$	$= \frac{0{,}602+1}{6{,}02 + 9{,}03} = \frac{1{,}602}{15{,}05} = 0{,}106$

Next, calculate the negative posterior probability on the test data (b) with the calculations in table 19 below:

Table 19 - Calculation of Negative Posterior Probability Test Data (b)

$P(positif \mid \underset{-}{bpom\ warning\ pembekuan\ darah\ vaksin\ covid\ aztrazeneca})$
$= (0{,}5 * 0{,}066 * 0{,}066 * 0{,}106 * 0{,}106 * 0{,}066 * 0{,}066 * 0{,}106)$ $= 0{,}00000001130$

3. The probability of the test data for each class of sentiment is known; the next step is to find the highest value of the posterior probability value of test data (a) and test data (b):
   a. In test data (a) with the tweet "Alhamdulillah vaksin covid tahap pertama berjalan lancar." For the posterior probability value, the positive class is worth (0.0000001436) and the negative class (0.0000004133), so the test data (a) is included in the positive sentiment class. Because the positive class posterior probability value is greater than the negative class.
   b. In the test data (b) with the tweet "bpom beri warning pembekuan darah pada vaksin covid astrazeneca,." For the posterior probability value, the positive class is worth (0.000000006674) and the negative class (0.00000001130), so the test data (b) is included in the negative sentiment class because the posterior probability value for the negative class is greater than the positive class.

3.6. Program Implementation

The implementation stage is the core stage of making a program, where the data processing starts from data collection to create a confusion matrix. The software used in this program is Jupyter Notebook as a text editor and data processing, Google Chrome as a browser, and the programming language used is Python version 3.7.

Data collection

The data collection process uses a Twitter scrapping tool called twint. In this process, the data is taken using keywords with the topic "vaksin covid-19" with a total of 5000 from April 09, 2021 - to April 18, 2021; then, the data is stored in a CSV file. The following is the process of data collection which can be seen in Figure 1.

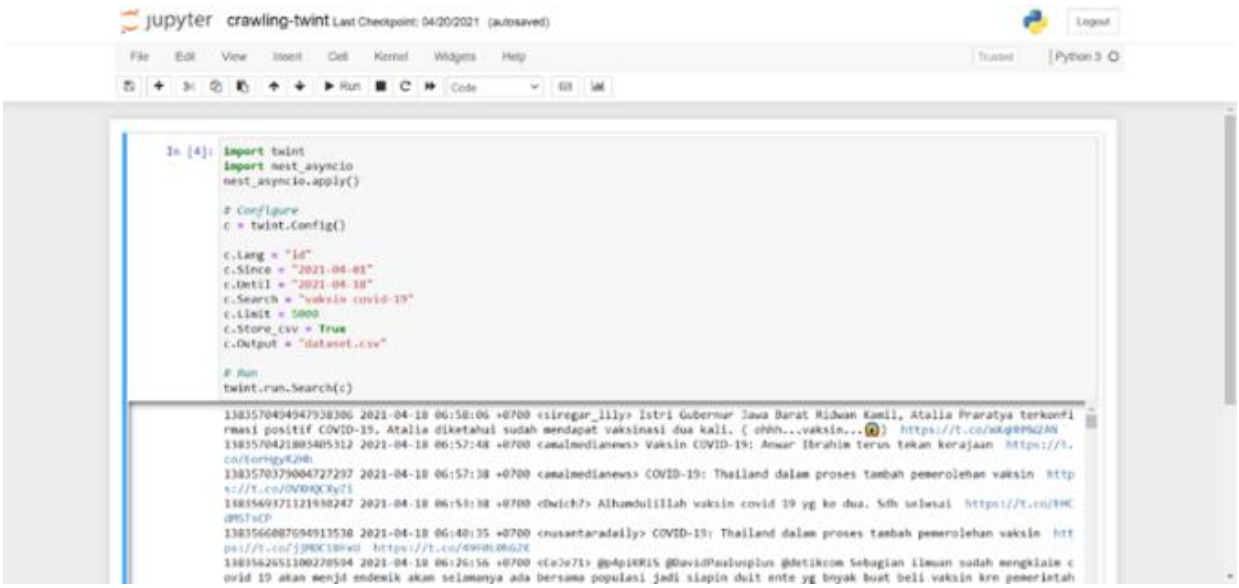

Figure 1 - Scrapping Tweet data

The pre-processing stage will be carried out by a number of processes, namely the cleaning steps can be seen in Figures 2a, Figure 2b, and 2c.

a.

Pre-Processing Cleaning

```
In [10]: # remove user
def remove_user(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)
    return input_txt
tweet_df['tweet_remove_user'] = np.vectorize(remove_user)(tweet_df['tweet'], "@[\w]*")
tweet_df
```

Out[10]:

	tweet	tweet_remove_user
0	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...
1	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...
2	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...
3	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...
4	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...
...	...	...
5048	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...
5049	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...
5050	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was-was tentang vaksin? Jom! daftar untuk Pr...
5051	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin, masyarakat tetap haru...
5052	Harus ada libur nasional untuk memperingati ha...	Harus ada libur nasional untuk memperingati ha...

5053 rows × 2 columns

b.

```
In [11]: # cleaning
def cleaning(strTweet):
    #remove non-ascii
    strTweet = unicodedata.normalize('NFKD', strTweet).encode('ascii', 'ignore').decode('utf-8', 'ignore')

    #remove URLs
    strTweet = re.sub(r'(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\))
    #remove punctuations
    strTweet = re.sub(r'[^\w]|_',' ',strTweet)

    #remove digit from string
    strTweet = re.sub("\S*\d\S*", "", strTweet).strip()

    #remove digit or numbers
    strTweet = re.sub(r"\b\d+\b", " ", strTweet)

    #Remove additional white spaces
    strTweet = re.sub('[\s]+', ' ', strTweet)

    #remove rt
    strTweet = re.sub(r'^RT[\s]+', '', strTweet)

    #remove tab, new line, and backslice
    strTweet = strTweet.replace('\\t', " ").replace('\\n', " ").replace('\\u', " ").replace('\\', "")

    # remove whitespace leading & trailing
    strTweet = strTweet.strip()

    # remove multiple whitespace into single whitespace
    strTweet = re.sub('\s+',' ',strTweet)

    # remove single character
    strTweet = re.sub(r"\b[a-zA-Z]\b", "", strTweet)

    return strTweet
tweet_df['tweet_cleaning'] = tweet_df['tweet_remove_user'].apply(cleaning)
tweet_df
```

c.

Out[11]:

	tweet	tweet_remove_user	tweet_cleaning
0	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil Atalia ...
1	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID Anwar Ibrahim terus tekan kerajaan
2	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...
3	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid yg ke dua Sdh selwsai
4	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...
...	...	...	...
5048	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP Elakkan diri daripada m...
5049	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid Ternyata Aman Bagi Orang Tua Foru...
5050	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was was tentang vaksin Jom daftar untuk Progra...
5051	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin masyarakat tetap harus...
5052	Harus ada libur nasional untuk memperingati ha...	Harus ada libur nasional untuk memperingati ha...	Harus ada libur nasional untuk memperingati ha...

5053 rows × 3 columns

Figure 2 - Tweet Cleaning Process (a) part 1, (b) part 2, and (c) part 3

The stages of case-folding can be seen in Figure 3 below.

Pre-Processing Casefolding

```
In [12]: # casefolding
def caseFolding(s):
    newStrTweetCaseFold = s.lower()

    return newStrTweetCaseFold
tweet_df['tweet_case_folding'] = tweet_df['tweet_cleaning'].apply(caseFolding)
tweet_df
```

Out[12]:

	tweet	tweet_remove_user	tweet_cleaning	tweet_case_folding
0	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil Atalia ...	istri gubernur jawa barat ridwan kamil atalia ...
1	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID Anwar Ibrahim terus tekan kerajaan	vaksin covid anwar ibrahim terus tekan kerajaan
2	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...	covid thailand dalam proses tambah pemerolehan...
3	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid yg ke dua Sdh selwsai	alhamdulillah vaksin covid yg ke dua sdh selwsai
4	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...	covid thailand dalam proses tambah pemerolehan...
...	...	...	...	...
5048	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP Elakkan diri daripada m...	stress pembunuh senyap elakkan diri daripada m...
5049	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid Ternyata Aman Bagi Orang Tua Foru...	vaksin covid ternyata aman bagi orang tua foru...
5050	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was was tentang vaksin Jom daftar untuk Progra...	was was tentang vaksin jom daftar untuk progra...
5051	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin masyarakat tetap harus...	setelah menerima vaksin masyarakat tetap harus...

Figure 3 - The Process of Case Folding Tweet

Tokenizing stages, which can be seen in Figure 4 below.

Out[12]:

	tweet	tweet_remove_user	tweet_cleaning	tweet_case_folding
0	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil Atalia ...	istri gubernur jawa barat ridwan kamil atalia ...
1	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID Anwar Ibrahim terus tekan kerajaan	vaksin covid anwar ibrahim terus tekan kerajaan
2	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...	covid thailand dalam proses tambah pemerolehan...
3	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid 19 yg ke dua. Sdh s...	Alhamdulillah vaksin covid yg ke dua Sdh selwsai	alhamdulillah vaksin covid yg ke dua sdh selwsai
4	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...	covid thailand dalam proses tambah pemerolehan...
...	...	...	...	...
5048	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP. Elakkan diri daripada ...	STRESS PEMBUNUH SENYAP Elakkan diri daripada m...	stress pembunuh senyap elakkan diri daripada m...
5049	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid-19 Ternyata Aman Bagi Orang Tua ...	Vaksin Covid Ternyata Aman Bagi Orang Tua Foru...	vaksin covid ternyata aman bagi orang tua foru...
5050	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was-was tentang vaksin? Jom! daftar untuk Pr...	Was was tentang vaksin Jom daftar untuk Progra...	was was tentang vaksin jom daftar untuk progra...
5051	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin, masyarakat tetap haru...	Setelah menerima vaksin masyarakat tetap harus...	setelah menerima vaksin masyarakat tetap harus...
5052	Harus ada libur nasional untuk memperingati ha...	Harus ada libur nasional untuk memperingati ha...	Harus ada libur nasional untuk memperingati ha...	harus ada libur nasional untuk memperingati ha...

5053 rows × 4 columns

Figure 4 - Tweet Tokenizing Process

The stopword removal steps can be seen in Figure 5 below.

Pre-Processing Stopword

In [16]:
```
# stopwords
list_stopwords = stopwords.words('indonesian')

list_stopwords.extend(["dg", "ny","d", "klo",
                       "amp", "krn", "n", 'u', 'jd', "nyg",
                       "hehe", "nder", "der", "pen", "sis", "jg",
                       "bgt", 'dah', 'ni', 'so', 'x', 'ri', 'dos', 'eee',
                       'skrng', 'skr', 'kpd', 'j', 's', 'b', 'jgn2', 'gara2',
                       'utk', 'y', 'g', 'm', 'pm', 't', 'dm', 'rm', 'p', 'indonesi', 'https',
                       'ampe', 'rt'
                      ])

list_stopwords = set(list_stopwords)

def removeStopwords(words):
    return ' '.join([word for word in words if word not in list_stopwords])
tweet_df['tweet_stopwords'] = tweet_df['tweet_tokenization'].apply(removeStopwords)
tweet_df
```

Out[16]:

	tweet	tweet_remove_user	tweet_cleaning	tweet_case_folding	tweet_tokenization	tweet_stopwords
0	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil, Atalia...	Istri Gubernur Jawa Barat Ridwan Kamil Atalia ...	istri gubernur jawa barat ridwan kamil atalia ...	[istri, gubernur, jawa, barat, ridwan, kamil, ...	istri gubernur jawa barat ridwan kamil atalia ...
1	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID-19: Anwar Ibrahim terus tekan ker...	Vaksin COVID Anwar Ibrahim terus tekan kerajaan	vaksin covid anwar ibrahim terus tekan kerajaan	[vaksin, covid, anwar, ibrahim, terus, tekan, ...	vaksin covid anwar ibrahim tekan kerajaan
2	COVID-19: Thailand dalam proses tambah pemerol...	COVID-19: Thailand dalam proses tambah pemerol...	COVID Thailand dalam proses tambah pemerolehan...	covid thailand dalam proses tambah pemerolehan...	[covid, thailand, dalam, proses, tambah, pemer...	covid thailand proses pemerolehan vaksin
3	Alhamdulillah vaksin	Alhamdulillah vaksin covid 19	Alhamdulillah vaksin covid	alhamdulillah vaksin covid	[alhamdulillah, vaksin,	alhamdulillah vaksin covid

Figure 5 - Stopword Removal Tweet

3.7. Data Visualization

After pre-processing, the data was labeled manually, both positively and negatively. Then at the next stage, data visualization is carried out with 3 data visualization displays consisting of bar charts, pie charts, and word clouds. Here are the results of visualizing tweet data to see the existing sentiments and the words that appear most often for each view. The following are the results of the bar chart visualization, which can be seen in Figure 6a, Figure 6b, and Figure 6c

a. b. c.

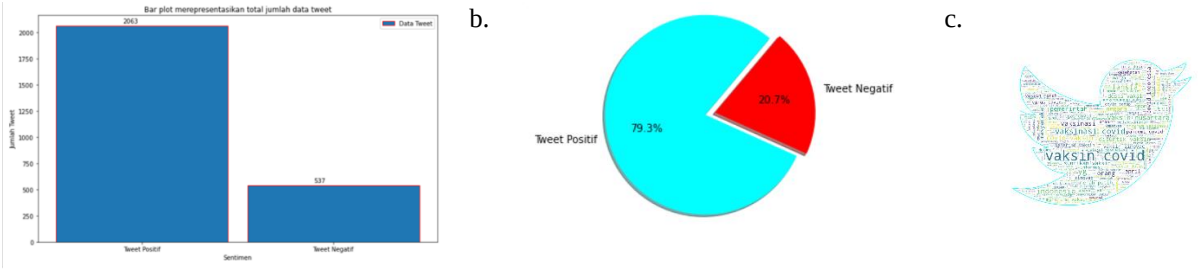

Figure 6 – Data Visualization (a) Bar chart, (b) Pie chart, and (c) Word Clouds

3.8. Sharing Training and Test Data

The data is divided into 80% training data, 2080 tweets, and 20% test data, 520 tweets randomly. The following is the implementation of the test and training data distribution, which can be seen in Figure 7.

```
In [14]: # spliting data
         X_train, X_test, y_train, y_test = train_test_split(df['tweet_cleans'],
                                                             df['label_sentimen'], test_size=0.2, stratify=df['label_sentimen'], random_st
         X_train.shape, X_test.shape, y_train.shape, y_test.shape

Out[14]: ((2080,), (520,), (2080,), (520,))
```

Figure 7 - Sharing Test and Train Data

3.9. TF-IDF Word Weighting

The word weighting in this study uses TF-IDF by utilizing the TfidfVectorizer library from scikit-learn. The following implementation of tf-idf can be seen in Figure 8.

```
In [15]: tfidf_model = TfidfVectorizer(smooth_idf=False).fit(X_train)
         X_train = tfidf_model.transform(X_train)
         X_test = tfidf_model.transform(X_test)
```

Figure 8 - TF-IDF implementation using the tfidfvectorizer library

3.10. Naïve Bayes Classifier

Classified data with a Naïve Bayes classifier with a multinomial model in the scikit-learn library. The following classification stage using multinomialNB can be seen in Figure 9.

Klasifikasi Naive Bayes Classifier

```
In [24]: mnb = MultinomialNB()
         mnb.fit(X_train, y_train)
         predict = mnb.predict(X_test)
         print(predict)

[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1]
```

Figure 9 - Implementation of the Naïve Bayes Classifier using the multinomialNB library

3.11. Model Storage

The model that will be saved is the model of the TF-IDF word weighting process. Furthermore, the model of data classification uses multinomial Naïve Bayes. Figure 10 shows the process of storing the model.

```
In [22]: #simpan tfidf model
         filename_tfidf = 'tfidf.pkl'
         pickle.dump(tfidf_model, open(filename_tfidf, 'wb'))

In [23]: # simpan naive bayes model
         with open('model_analisis.pkl', 'wb') as r:
             pickle.dump(mnb, r)

In [ ]:
```

Figure 10 - Storage Model

3.12. New Data Classification

Figure 11 is the new data classification process stage by calling the model that has been created and calling the file that does not have a label. The model can classify and label tweets into a positive class with a value of 1 and a negative class with a value of 0.

Figure 11 - New Data Classification Results From Model

3.13. Testing

In this test, we will calculate the accuracy of the model that has been built. The results of model testing on 520 tweets of test data can be seen in the confusion matrix in Figure 12 below.

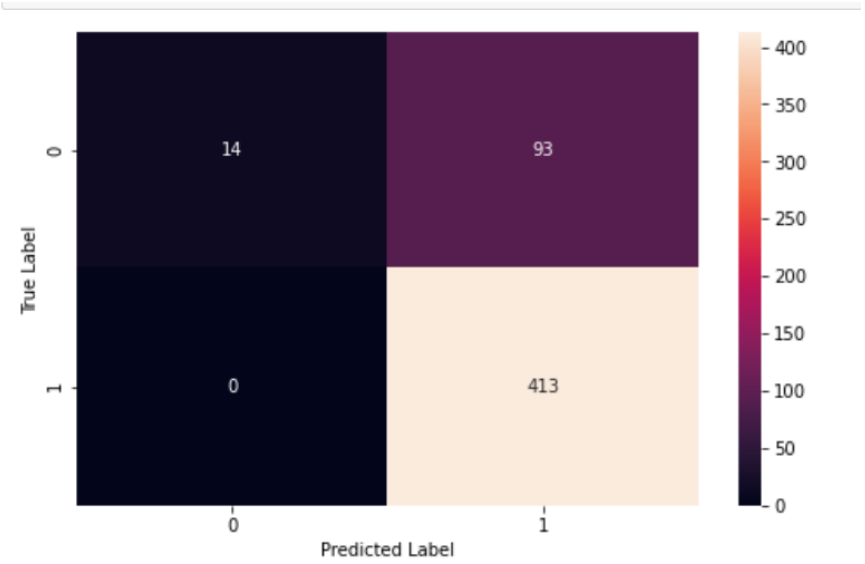

Figure 12 - The results of the confusion matrix from the classification model

Figure 12 can be seen the value:
TP (True Positive) = 413
TN (True Negative) = 14
FP (False Positive) = 93
FN (False Negative) = 0

By looking at the confusion matrix, the accuracy value of the classification model can be calculated. Accuracy shows how accurately the model classifies data correctly. So to find out the accuracy value, you can calculate the information that is predicted to be correct with the overall data as follows:

$$accuracy = \frac{413 + 14}{413 + 14 + 93 + 0} \text{ X } 100\% = 82\%$$

Based on the calculations, it is known that the results of sentiment classification using the Naïve Bayes classifier in this study resulted in an overall accuracy of 82%. In this test, we will calculate the accuracy of the model that has been built. The results of model testing on 520 tweets of test data can be seen in the confusion matrix in Figure 13 below.

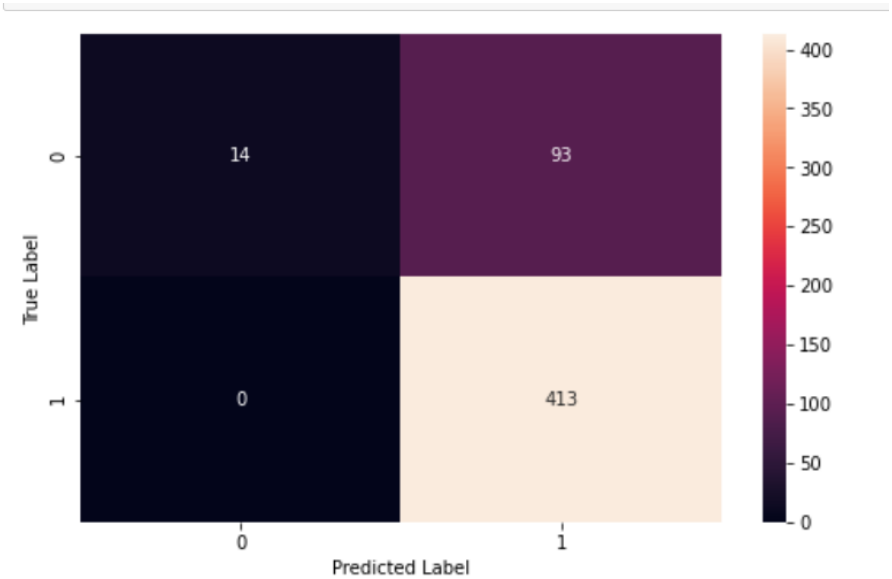

Figure 13 - The results of the confusion matrix from the classification model

Based on the calculations, it is known that the results of sentiment classification using the Naïve Bayes classifier in this study resulted in an overall accuracy of 82%.

4. CONCLUSION

Based on the author's research, conclusions can be drawn, including the Naïve Bayes classifier algorithm can be applied to classify tweets into positive or negative sentiment classes. The data visualization of tweets was successfully displayed via bar charts, pie charts, and word clouds. So it can be seen that many people support the procurement of the COVID-19 vaccine. It can be seen from the percentage of positive sentiment of as much as 79.3% and sentiment analysis on the Covid-19 vaccine topic using the Naïve Bayes classifier algorithm can produce an accuracy value from the model with an accuracy rate of 82%, with a composition of 80% training data and 20% test data. In further research, the labeling process can be carried out with people who are more skilled in the language field to be more precise in labeling tweets. It is possible to apply an algorithm to correct writing errors or non-standard words and add a negation detection feature (convert negation). It is also possible to apply this sentiment analysis to a website-based system to perform real-time data scrapping and other processes automatically. It is expected to be able to apply machine learning algorithm techniques with other types, such as unsupervised learning so that there is no need to label the data first, and the research can improve training data and a better and more stopword word dictionary so that the pre-processing process can be better.

REFERENCES

[1] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Indones. Heal. Inf. Manag. J.*, vol. 8, no. 2, pp. 100–109, Dec. 2020.

[2] C. A. Madeline *et al.*, "Klasifikasi Analisis Sentimen Terhadap Bipolar Disorder Pada Media Sosial Twitter Dengan Menggunakan Metode Support Vector Machine (Svm)," pp. 1–10, 2019.

[3] A. R. T. Lestari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.

[4] M. Syarifuddin, "Analisis Sentimen Opini Publik Terhadap Efek PSBB Pada Twitter Dengan Algoritma Decision Tree-KNN-Naïve Bayes," *Inti Nusa Mandiri*, vol. 15, no. 1, pp. 87–94, 2020.

[5] "KLASIFIKASI KOMENTAR SPAM PADA INSTAGRAM MENGGUNAKAN METODE SUPPORT VECTOR MACHINE - FTI ARS University." [Online]. Available: https://fti.ars.ac.id/publikasi/2107031535541. [Accessed: 09-May-2022].

[6] Y. Cahyono, "Analisis Sentiment pada Sosial Media Twitter Menggunakan Naïve Bayes Classifier dengan Feature Selection Particle Swarm Optimization dan Term Frequency," *J. Inform. Univ. Pamulang*, vol. 2, no. 1, p. 14, 2017.

[7] T. Wahyono, *Fundamental Of Python For Machine Learning*, 1st ed. Jakarta: GAVA Media, 2018.

[8] "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine - PDF Free Download." [Online]. Available: https://docplayer.info/33991922-Analisis-sentimen-pada-twitter-untuk-mengenai-penggunaan-transportasi-umum-darat-dalam-kota-dengan-metode-support-vector-machine.html. [Accessed: 09-May-2022].

[9] T. Mardiana, H. Syahreva, and T. Tuslaela, "Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, 2019.

[10] A. Yuni Muallifah Fakultas Saintek and U. Sunan Kalijaga, "Mengurai Hadis Tahnik dan Gerakan Anti Vaksin," vol. 2, 2017.

[11] "Analisis Sentimen Twitter Terhadap Pembayaran ShopeePayLater Pada Aplikasi Belanja Online (Shopee) Menggunakan Metode Lexicon Based Dan Naïve Bayes Classifier," *J. Ilm. Komputasi*, vol. 19, no. 4, Dec. 2020.

[12] C. M. K. Imam Mulya, "Analisis Sentimen Terhadap Universitas Gunadarma Berdasarkan Opini Pengguna Twitter Menggunakan Metode Naïve Bayes Classifier," *J. Ilm. Komputasi*, vol. 19, no. 4, pp. 507–521, 2020.

[13] Z. Efendi and M. Mustakim, "Text Mining Classification sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi," *Semin. Nas. Teknol. Inf. Komun. dan Ind.*, vol. 0, no. 0, pp. 235–242, 2017.

[14] M. K. Maulidina, "ANALISIS SENTIMEN KOMENTAR WARGANET TERHADAP POSTINGAN INSTAGRAM MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN TF-IDF (Studi Kasus: Instagram Gubernur Jawa Barat Ridwan Kamil)," *Naskah Publ. Univ. Teknol. Yogyakarta*, pp. 1–15, 2020.

[15] D. Musfiroh *et al.*, "Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 24–33, 2021.