



**Sentiment Analysis from Twitter about Covid-19 Vaccination in Indonesia  
using Naive Bayes and XGBoost Classifier Algorithm**

*THESIS REPORT*

Alvin Irwanto  
41518010055

UNIVERSITAS  
DEPARTMENT OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021



**Sentiment Analysis from Twitter about Covid-19 Vaccination in Indonesia  
using Naive Bayes and XGBoost Classifier Algorithm**

*THESIS REPORT*

Submitted to Complete Terms  
Completed a Computer Bachelor Degree

Created By:

Alvin Irwanto  
41518010055

UNIVERSITAS  
MERCU BUANA

**DEPARTMENT OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021**

## ORIGINALITY STATEMENT SHEET

### ORIGINALITY STATEMENT SHEET

The undersigned below:

Student Number : 41518010055

Name : Alvin Irwanto

Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and XGBoost  
Classifier Algorithm

Stating that my Thesis Report is my own and not plagiarism. If it is found in my Thesis Report that there is an element of plagiarism, then I am ready to get academic sanctions related to it.



Jakarta, 17 January 2022



Alvin Irwanto

UNIVERSITAS  
MERCU BUANA

## THESIS PUBLICATION STATEMENT

### THESIS PUBLICATION STATEMENT

As a Universitas Mercu Buana student, I, the undersigned below:

Student Name : Alvin Irwanto  
Student Number : 41518010055  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and  
XGBoost Classifier Algorithm

By giving permission and approval of **None-exclusive Royalty Free Right** to Universitas Mercu Buana for my scientific work entitled above along with the available devices (if necessary).

With this **None-exclusive Royalty Free Right**, Universitas Mercu Buana has the right to store, transfer / format, manage in form of database, administer and publish my thesis.

Furthermore, in sake of science development in Universitas Mercu Buana environment, I give the permission to Researcher in Research Lab of Computer Science Faculty, Universitas Mercu Buana to use and develop existing result of the research of my thesis for the research and publication purpose as long as my name is stated as author / creator and Copyright owner.

Hereby I made this statement in truthfulness.

UNIVERSITAS  
MERCU BUANA

Takarta, 17 January 2022



Alvin Irwanto

## THESIS OUTPUT STATEMENT LETTER

### THESIS OUTPUT STATEMENT LETTER

As a Universitas Mercu Buana student, I, the undersigned below:

Student Name : Alvin Irwanto  
Student Number : 41518010055  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and  
XGBoost Classifier Algorithm

Declared that:

1. My Thesis Output as follows:

No	Output	Type	Status
1	Scientific Publication	Not Accredited National Journal	Submitted ✓
		Accredited National Journal	
		Not Reputable International Journal	Approved
	Reputable International Journal ✓		
	Submitted/Published at:	Journal Name : SINERGI	
	ISSN : 2460-1217		
	Journal Link : <a href="https://publikasi.mercubuana.ac.id/index.php/sinergi">https://publikasi.mercubuana.ac.id/index.php/sinergi</a>		
	Published Journal Link :-		
	File		

2. Willing to complete the entire article publication process starting from submitting, revising the article until it is declared that it can be published in the intended journal.
3. Asked to attach a scanned ID card and a statement letter (see the HKI document attachment), for the purpose of registering HKI if needed.

This statement I made in truth.

Jakarta, 17 January 2022

  
Alvin Irwanto

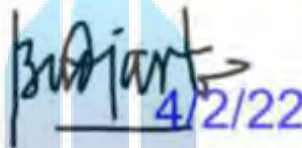
## COMMITTEE APPROVAL SHEET

Student Number : 41518010055  
Student Name : Alvin Irwanto  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and  
XGBoost Classifier Algorithm

This Thesis has been checked and examined as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 17 January 2022

Approved



(Prof. Dr. Rahmat Budiarto. M. Eng)

UNIVERSITAS  
MERCU BUANA

## COMMITTEE APPROVAL SHEET

Student Number : 41518010055  
Student Name : Alvin Irwanto  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and  
XGBoost Classifier Algorithm

This Thesis has been checked and examined as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 17 January 2022

Approved



(Emil Robert Kaburuan, S.T., M.A., Ph.D)

UNIVERSITAS  
MERCU BUANA

## COMMITTEE APPROVAL SHEET

Student Number : 41518010055  
Student Name : Alvin Irwanto  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and  
XGBoost Classifier Algorithm

This Thesis has been checked and examined as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 17 January 2022

Approved



Anis Cherid, SE, MTI

UNIVERSITAS  
MERCU BUANA




## VALIDITY SHEET

Student Number : 41518010055  
Student Name : Alvin Irwanto  
Thesis Title : Sentiment Analysis from Twitter about Covid-19  
Vaccination in Indonesia using Naive Bayes and XGBoost  
Classifier Algorithm

This Thesis has been checked and examined as one of the requirement to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 17 January 2022

Approved,



(Dr. Leonard Goeirmanto)

Thesis Supervisor

UNIVERSITAS  
MERCU BUANA

Acknowledge,



(Wawan Gunawan, S.Kom, MT)  
Informatics Thesis Coordinator



(Emil Robert Kaburuan, S.T., M.A., Ph.D)  
Head of Informatics Department

## PREFACE

Praise and gratitude gently given to Allah SWT who has given His grace and guidance so that the author can compile this Thesis report with title "Sentiment Analysis from Twitter about Covid-19 Vaccination in Indonesia using Naive Bayes and XGBoost Classifier Algorithm" on schedule. This final project was prepared to fulfill one of the requirements for obtaining a bachelor's degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

In writing this Thesis report, the author realizes that without the help and guidance of many parties, this research would not be carried out properly. Therefore, the authors would like to thank:

1. Prof. Dr. Ir. Ngadino Surip, MS as the Chancellor of Universitas Mercu Buana who has provided many positive changes and progress for our university.
2. Yaya Sudarya Triana, M.Kom., as Dean of the Faculty of Computer Science, Universitas Mercu Buana
3. Emil Robert Kaburuan, S.T., M.A., Ph.D., as Head of the Department of Informatics Engineering, Universitas Mercu Buana
4. Anis Cherid, SE, MTI, as Head of the International Department of Informatics, Universitas Mercu Buana
5. Dr. Leonard Goeirmanto, ST, M. Sc as a Thesis Supervisor who has helped a lot and is willing to take the time to provide guidance.
6. Lecturers of Informatics Engineering at Universitas Mercu Buana, for their guidance and teaching in lectures
7. My parents and family who have given a lot of enthusiasm and motivation so that I can complete this thesis report.
8. Friends who have always been an encouragement and motivation to the author during the implementation of this final project.

The author realizes that this report is still far from perfect, it is not free from mistakes and shortcomings. As a form of improvement, the author is open to suggestions and constructive criticism from readers for the perfection of this report, also for the writer's improvement. Finally, the author hopes that this final project can be useful for readers to increase knowledge and insight and can be useful for future research.



Jakarta, 17 January 2022



Alvin Irwanto

## TABLE OF CONTENTS

<b>COVER .....</b>	<b>i</b>
<b>TITLE PAGE .....</b>	<b>i</b>
<b>ORIGINALITY STATEMENT SHEET .....</b>	<b>ii</b>
<b>THESIS PUBLICATION STATEMENT .....</b>	<b>iii</b>
<b>THESIS OUTPUT STATEMENT LETTER .....</b>	<b>iv</b>
<b>COMMITTEE APPROVAL SHEET .....</b>	<b>v</b>
<b>VALIDITY SHEET .....</b>	<b>viii</b>
<b>ABSTRAK .....</b>	<b>ix</b>
<b>ABSTRACT .....</b>	<b>x</b>
<b>PREFACE.....</b>	<b>xi</b>
<b>TABLE OF CONTENTS.....</b>	<b>xii</b>
<b>JOURNAL TEXT .....</b>	<b>1</b>
<b>CHAPTER 1. LITERATURE REVIEW .....</b>	<b>12</b>
<b>CHAPTER 2. ANALYSIS AND DESIGN.....</b>	<b>20</b>
<b>CHAPTER 3. SOURCE CODE.....</b>	<b>23</b>
<b>CHAPTER 4. DATASET .....</b>	<b>35</b>
<b>CHAPTER 5. EXPERIMENT STAGE .....</b>	<b>36</b>
<b>CHAPTER 6. RESULTS ALL EXPERIMENTS .....</b>	<b>41</b>
<b>BIBLIOGRAPHY .....</b>	<b>44</b>
<b>ATTACHMENT OF HAKI DOCUMENTS .....</b>	<b>46</b>
<b>ATTACHMENT OF CORRESPONDENCE .....</b>	<b>48</b>

	<p><b>SINERGI</b> Vol. xx, No. x, February 20xx: xxx-xxx <a href="http://publikasi.mercubuana.ac.id/index.php/sinergi">http://publikasi.mercubuana.ac.id/index.php/sinergi</a> <a href="http://doi.org/10.22441/sinergi.xxxx.x.xxx">http://doi.org/10.22441/sinergi.xxxx.x.xxx</a></p>	
---	--	---

## SENTIMENT ANALYSIS FROM TWITTER ABOUT COVID-19 VACCINATION IN INDONESIA USING NAIVE BAYES AND XGBOOST CLASSIFIER ALGORITHM

Alvin Irwanto<sup>1</sup>, Leonard Goeirmanto<sup>2</sup>

<sup>1</sup>Department of Informatics, Faculty of Computer Science, Universitas Mercu Buana

### Abstract

The pandemic that hit the world has big impact in our life. But after some time, it seems that it will be going to end because the vaccine has already been made. In response to this, some people expressed their opinions about this vaccination on social media, for example in the form of tweets on Twitter. Those opinion or tweet that the authors use as a sentiment analysis material to find out the assessment of this vaccination. The tweet data in this study was obtained through data crawling using the Twitter API with the Python programming language. The variables used in this case are public tweets and their sentiments. This sentiment analysis process uses the Classification method with the Naive Bayes Classifier and will be compared with the XGBoost Classifier algorithm. The results of this study indicate that people are more likely to be positive in responding to this vaccination. It was also found that in this case, the Naive Bayes Classifier got a better performance with 0.95 from ROC - AUC Score and 134 ms in runtime compared to the XGBoost Classifier algorithm that have 0.882 in ROC - AUC Score and 1 minutes and 59 seconds in runtime.

This is an open access article under the [CC BY-NC](https://creativecommons.org/licenses/by-nc/4.0/) license



### Keywords:

sentiment analysis, twitter, covid-19 vaccination, naive bayes classifier, xgboost classifier

### Article History:

Received: May 2, 2019  
Revised: May 29, 2019  
Accepted: June 2, 2019  
Published: June 2, 2019

### Corresponding Author:

Leonard Goeirmanto  
Informatics Department,  
Universitas Mercu Buana,  
Indonesia  
Email:  
[leonard@mercubuana.ac.id](mailto:leonard@mercubuana.ac.id)

### INTRODUCTION

The Covid-19 pandemic that happens has destroyed almost all sectors of human life. For example, is in work and college, which previously had to come to a place, are now required to stay at home which is done online. Various efforts have been made to prevent the spread, ranging from wearing masks, physical distancing, to frequent hand washing. However, the most effective solution at this time is by vaccination. The primary objective of

vaccination is to achieve herd immunity, which implies that people are pushed to get vaccinations for at least 70% of the population to be immune. Although it cannot prevent 100% of the spread of Covid-19, it can at least reduce the effects of its spread. In Indonesia itself, there are various types of vaccines used, ranging from Sinovac, Astrazeneca, Pfizer, Moderna, to the most recent, Zifivax.

However, this vaccination program itself has not been able to run optimally, because

there are still some people who have doubts about this vaccine. On social media, there are various kinds of opinions regarding this Covid-19 vaccination, for example, on Twitter. There are still some people on Twitter in particular, who are hesitant about this vaccination. Moreover, coupled with cases of death after vaccination, some people are increasingly afraid and doubt the use of vaccines.

This study aims to analyze the public's response to Covid-19 vaccination by classifying it into positive and negative responses. It is hoped that the results of this sentiment can be used as information and evaluation material for related parties, in seeing public opinion about the Covid-19 vaccination, whether the socialization has been conveyed well and can be understood by the people in Indonesia, or rather there are still miscommunication going on in the community regarding this vaccination.

In this paper, the authors use two algorithms as a model for extract and mine the value from the data. Those algorithms are Naïve Bayes Classifier and XGBoost Classifier. The reason why the authors choose those algorithms because both algorithms already known to have high performance in the application of classification case such as in research and machine learning competition. This paper also discusses which of the algorithms has the best performance for handling this case.

## RELATED WORKS

Sentiment analysis has attracted the attentions of many data mining researches. Sentiment analysis is primarily used to express a particular individual's opinion. In conclusion, the current cutting edge divided classes into two categories: positive and negative. Many research has used various methodologies to accomplish sentiment analysis. According to Abdullah and Hadzikadic [1], sentiment analysis for text uses two primary methodologies: symbolic or lexicon-based and machine learning approaches. In this study, the authors focus on machine learning approaches as a methodology.

Social media platforms such as Twitter is a place where many people express their thoughts and opinions freely. Many researchers and scientists study the topic about sentiment analysis that came from Twitter data. From this research, it can help and figure it out many things, depends on what research is for. For example, help e-commerce businesses in focusing on improving service and company

quality, which lead to improv traffic, sales, and profitability.

It is also used in sentiment analysis of public towards the Republic of Indonesia's presidential candidates for the 2019 - 2024 period using tweet data [2]. In that research, they are using Naive Bayes method to predict the class. After that, they compare with other methods such as Support Vector Machine and K-Nearest Neighbors algorithm. They classify the sentiment by two classes namely positive and negative. The results of their experiments, it is shows that the Naïve Bayes method has a better accuracy (80.90%) compared other methods, such as KNN (75.58%) and SVM (63.99%). For the other results of, they found that the value of the positive sentiment polarity of the Jokowi-Ma'ruf Amin was 45.45% positive and 54.55% negative, while the Prabowo-Sandiaga received a 44.32% positive and 55.68% negative sentiment.

Another example is to investigate customer feedback about US airline services who has been done by Saad [3]. The author tests the models using a sample of tweets from six different airlines in the United States. He applied the Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), Nave Bayes (NB), and Decision Tree (DT) machine learning techniques to categorize tweets in the classification phase, and the K-Fold Cross Validation methodology to test and validate the model. When the results of each classifier were compared, it was discovered that SVM had the greatest accuracy of 83.31%.

## THEORETICAL BACKGROUND

### Twitter

Twitter is one type of social media that is often used by most people to communicate via the internet. Twitter is operated by Twitter, Inc., which offers a social network in the form of a microblog. It is called a microblog because this site allows users to send and read blog messages as usual, but is limited to only 280 characters that are displayed on the user's profile page. Every message posted to Twitter is known as a tweet. In that tweet, the users can upload photos, videos, links and text [4].

### Sentiment Analysis

Sentiment analysis is a branch of data mining that aims to analyze, understand, process, and extract textual data in the form of opinions on entities such as products, services, organizations, individuals, and certain topics [5]. This analysis can be used to obtain information



which can be in the form of positive and negative percentages from a data set, one of which can be obtained through public tweets on Twitter [6]. This sentiment analysis aims to make an assessment that occurs in society on the topic being discussed. The results can be used for evaluation of related parties.

### Vaccination

Vaccination is the provision of certain viral antigenic components, which of course are safe. Individuals who are vaccinated can trigger immunity or immunity against viral infections in accordance with that given [7]. This vaccination can be done using a syringe or even a drop into the mouth, for now, vaccination of Covid-19 itself is done by using a syringe. There are several vaccines that which has been circulating, in Indonesia itself, the circulating vaccines are Sinovac, Astrazeneca, Moderna, Pfizer, and Zifivax.

### SMOTE

Synthetic Minority OverSampling Technique (SMOTE) is an over-sampling algorithm that is commonly used to make synthetic data by making replication from minority data. It works by choosing a random example from the minority class is first. Then k of the nearest neighbors for that example is found, after that, the randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space [8].

### Naïve Bayes Classifier

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, that usually use in classification especially for text classification. Naïve Bayes uses simple statistics based on the Bayes theorem which assumes the presence or absence of a certain feature of a class that is not related to other features. Naïve Bayes method is based on conditional probability and maximum likelihood of occurrence [9]. The Bayes Theorem calculates the probability of an event based on the probability of a previous event. The Bayes Theorem can be defined as:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (1)$$

Explanation:

P(A B)	: Posterior probability
P(B A)	: Likelihood
P(A)	: Class prior probability
P(B)	: Predictor prior probability

### XGBoost Classifier

XGBoost stands for eXtreme Gradient Boosting is an implementation of gradient boosted decision trees designed for execution speed and model performance [10]. It is a machine learning method that's recently dominated Kaggle competitions for structured or tabular data in applied machine learning. It's a part of a supervised learning algorithm that combines the estimates of a set of simpler and weaker models, to try to properly predict a target variable. The beauty of this powerful algorithm is its scalability, which allows for rapid learning via parallel and distributed computing while still utilizing memory efficiently. XGBoost falls under the category of Boosting techniques in Ensemble Learning. Ensemble learning combines different models into a collection of predictors to improve prediction accuracy [11]. The mistakes caused by earlier models are attempted to be rectified by subsequent models by adding weights to the models in the boosting approach (give more weight if it performs better).

## RESEARCH METHOD

### Data Collection

The data collection method used in this research is by collecting tweet data or so-called Crawling data from Twitter. The data is taken from May until October 2021. The reason why the data is taken at that time is because the vaccination mass vaccination was already running in Indonesia and on May, there was a case where it was reported that someone had died after being vaccinated which certainly made the public's trust in this vaccination decrease and doubts about this vaccine. The crawling data end on October because the authors think that at that time, Covid-19 vaccination is almost evenly distributed and has become a must for all citizens, so that opinion tweets about this vaccination are starting to decrease.

The dataset in this research contains 2 columns, which are Comment and Value, with 10208 rows. To get the data itself, the authors using the Python library, namely Tweepy, which can collect tweets from users based on certain keywords. To be able to collect this tweet data, a certain token and key is required, including access token, access token secret, consumer key, and consumer key secret, which is obtained via the Twitter API. In this research, the authors use hashtags as keywords to search for tweets about this vaccination. The hashtags that the authors use are "#vaksinasi", "#vaksin",

"#vaksinasicovid19", "#vaksinCovid19", "#sinovac", "#astrazeneca", "#pfizer", and "#zifivax". The authors here are only crawling the tweet or comment without determining or checking who wrote it. The sample data shown in Table 1.

Table 1. Sample Dataset

No.	Comment
1	Serem nih efek #AstraZeneca 😞 <a href="https://t.co/8xyNHRNB0H">https://t.co/8xyNHRNB0H</a>
2	Efek pasca vaksin #AstraZeneca ini malah lebih berat dibanding pas positif Covid dulu, hehe 😊
3	Jadi seperti itulah cerita mengenai #vaksin #COVID19. Jangan mudah terpengaruh oleh isu - isu negatif, yang bisa ja... <a href="https://t.co/9BetZGU6o3">https://t.co/9BetZGU6o3</a>
4	#AstraZeneca bener banget ..... anw, manfaat nya masih lebih tinggi dari mudarat nya kan <a href="https://t.co/y5zamaRj0F">https://t.co/y5zamaRj0F</a>
5	Vaksin Sinovac Kemampuannya Melampaui Uji Klinis <a href="https://t.co/E8mf9mvzSY">https://t.co/E8mf9mvzSY</a> via @holopiscom #Vaksin #Covid19... <a href="https://t.co/hAZDTqGU3Z">https://t.co/hAZDTqGU3Z</a>

### Labeling Data

Labeling the data is important for the machine learning process. The data labeling process is done manually and the sentence will be determined to have a positive or negative meaning. In this study, the authors only used 2 sentiments, namely positive and negative which are denoted by the numbers 0 (negative) and 1 (positive).

The positive label seen from the contents of the tweet contains sentences that are positive, supportive and statements of agreement. Negative label is a class with data containing negative meaning, ridicule, and contradictory sentences. The consistency is important here, so the machine is not confused and make a bad accuracy. If it already considers profanity to be negative, don't consider the other half of the dataset to be positive if it contains profanity.

The authors label the easiest examples first. The obvious positive/negative examples should be labeled as soon as possible, and the hardest ones should be left to the end when already have a better comprehension of the problem, for example, if the tweet is sarcasm or ironic, the authors delve deeper into this case and look back at the tweet, to avoid misunderstanding the meaning of the tweet. The data after labeling shown in Table 2.

Table 2. Data after labeling

No.	Comment	Value
1	Serem nih efek #AstraZeneca 😞 <a href="https://t.co/8xyNHRNB0H">https://t.co/8xyNHRNB0H</a>	0
2	Efek pasca vaksin #AstraZeneca ini malah lebih berat dibanding pas positif Covid dulu, hehe 😊	0

3	Jadi seperti itulah cerita mengenai #vaksin #COVID19. Jangan mudah terpengaruh oleh isu - isu negatif, yang bisa ja... <a href="https://t.co/9BetZGU6o3">https://t.co/9BetZGU6o3</a>	1
4	#AstraZeneca bener banget ..... anw, manfaat nya masih lebih tinggi dari mudarat nya kan <a href="https://t.co/y5zamaRj0F">https://t.co/y5zamaRj0F</a>	1
5	Vaksin Sinovac Kemampuannya Melampaui Uji Klinis <a href="https://t.co/E8mf9mvzSY">https://t.co/E8mf9mvzSY</a> via @holopiscom #Vaksin #Covid19... <a href="https://t.co/hAZDTqGU3Z">https://t.co/hAZDTqGU3Z</a>	1

### Pre-Processing Data

#### Case Folding

Case folding itself is divided into several steps, which are removing the link, hashtag, and username that usually come after the tweet, changing the sentences into lowercase, and remove punctuation, number, and special characters that will not be used in later stages. Characters other than letters are removed and are considered delimiters. Table 3 show the data after this process.

Table 3. Data after case folding

No.	Comment	Value
1	serem nih efek	0
2	efek pasca vaksin ini malah lebih berat dibanding pas positif covid dulu hehe	0
3	jadi seperti itulah cerita mengenai jangan mudah terpengaruh oleh isu isu negatif yang bisa ja	1
4	bener banget anw manfaat nya masih lebih tinggi dari mudarat nya kan	1
5	vaksin sinovac kemampuannya melampaui uji klinis via	1

#### Remove Redundant and Unrelated Tweet

This process is done after the case folding because there are many tweets that have same sentence but at the end of it have a different link. If run this process first, it will end with many duplicate data. The data also contains many tweet that only contains unrelated tweet, such as tweet that does not use Bahasa Indonesia, talk about another topic, and only contains 1 character that does not even have any meaning. So, in this process the authors also removed it from the dataset. After those processes, the dataset is reduced from 10208 rows into 4176 rows.

#### Tokenizing

This is the process of breaking sentences into separate words called tokens [12]. This process separates words in the dataset based on readable spaces [13]. The sentence from the tweet will be broken down into word for

word in an array. These tokens help in better identification of the context or the model's development. By evaluating the sequence of words, tokenization helps in interpreting the meaning of the text. This is how we teaching machine about the words. Table 4 show the data after this process.

Table 4. Data after tokenizing

No.	Comment	Value
1	['serem', 'nih', 'efek']	0
2	['efek', 'pasca', 'vaksin', 'ini', 'malah', 'lebih', 'berat', 'dibanding', 'pas', 'positif', 'covid', 'dulu', 'hehe']	0
3	['jadi', 'seperti', 'itulah', 'cerita', 'mengenai', 'jangan', 'mudah', 'terpengaruh', 'oleh', 'isu', 'isu', 'negatif', 'yang', 'bisa', 'ja']	1
4	['bener', 'banget', 'anw', 'manfaat', 'nya', 'masih', 'lebih', 'tinggi', 'dari', 'mudarat', 'nya', 'kan']	1
5	['vaksin', 'sinovac', 'kemampuannya', 'melampau', 'uji', 'klinis', 'via']	1

### Filtering or Stopwords Removal

This stage aims to eliminate the stop words in the data. Stopword itself is a common word that usually appears in large numbers and is considered meaningless. Examples of stopwords in Indonesian are “yang”, “di”, “dari”, etc. It must be removed because it will take up space in database, or taking up valuable processing time. So, for this, the authors remove them by storing a list of words that consider as stop words. Table 5 show the data after this process.

Table 5. Data After Filtering

No.	Comment	Value
1	['serem', 'nih', 'efek']	0
2	['efek', 'pasca', 'vaksin', 'berat', 'dibanding', 'positif', 'covid']	0
3	['cerita', 'mengenai', 'jangan', 'mudah', 'terpengaruh', 'oleh', 'isu', 'isu', 'negatif']	1
4	['bener', 'banget', 'manfaat', 'tinggi', 'mudarat']	1
5	['vaksin', 'sinovac', 'kemampuannya', 'melampau', 'uji', 'klinis', 'via']	1

### Stemming

Stemming itself is the process of returning a data to its root word. This stage will eliminate the suffix and prefix in the token / word (reduce inflected), so that a word that has a suffix or prefix will return to its basic form. The example in Bahasa Indonesia is:

Penerima → Terima  
Ditetapkan → Tetap

Melawan → Lawan

Table 6. Data after stemming

No.	Comment	Value
1	serem nih efek	0
2	efek pasca vaksin berat banding positif covid	0
3	cerita mengenai jangan mudah pengaruh isu isu negatif	1
4	bener banget manfaat tinggi mudarat	1
5	vaksin sinovac mampu lampau uji klinis via	1

### TF-IDF

TF-IDF is statistical measure of a word's relevance to a document in a collection of documents. This is accomplished by multiplying two metrics, first is the number of times a word appears in a document and the second is the word's inverse document frequency over a collection of documents. TF itself stands for Term's Frequency and IDF is Inverse Document Frequency. In this process, it also converts the data into a matrix of TF-IDF features. This is because the Machine Learning algorithms cannot work with raw data directly.

### Synthetic Minority Oversampling Technique (SMOTE)

The dataset that used in this research is imbalanced, with 3810 data have positive label and 366 negative labels. This could be a problem because most machine learning algorithms are designated to work best with balanced data that the target classes have similar prior probabilities [14].

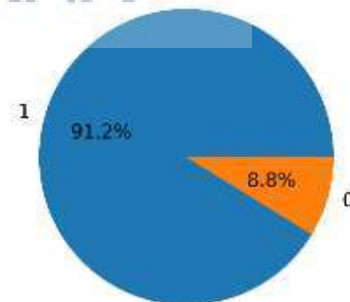


Figure 1. Comparison between the labels before applying SMOTE

If the data imbalanced, the machine would make predictions that are more inclined towards the majority class. To handle that problem, the authors used SMOTE for this research. SMOTE is used to oversample the minority class through producing “synthetic” examples rather than over-sampling with replacements [15]. SMOTE works by choosing



examples in the feature space that are close together, drawing a line in the feature space between the examples, and generating a new sample at a position along that line. To be more specific, a random case from the minority class is selected initially. Then, initialize the  $k$  of the closest neighbors are found (usually  $k=5$ ). After that, A randomly chosen neighbor is picked, and a synthetic example is generated at a randomly chosen point in feature space between the two examples. After applying that technique, the data class are equals.

SMOTE itself is well known in handling imbalanced data cases. It can be seen from several studies that have used this method, for example in research conducted by Jonathan, Putra, and Ruldeviyani [8] and also research by Lu, Cheung, and Tang [16], where the application of SMOTE is very optimal for the results. However, it should also be noted that in applying this method, the given dataset should be clean and clear from noise, because if noise samples exist in the minority class, it will blind oversampling algorithms like SMOTE and will produce additional noises, hence applying oversampling to the noisy minority class may actually degrade performance more [16]. In this research itself, the dataset is already through several preprocessing and data cleaning processes, so that the resulting data is ready to be applied with the SMOTE method. Figure 2 show how the label comparison after applying this method.

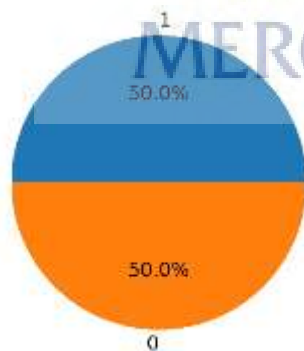


Figure 2. Comparison between the labels after applying SMOTE

### Model Implementation

The implementation of the model is done using Python programming language and Jupyter notebook as a software to use it. The reason why those algorithms was chooses because both algorithms already known to have high performance in classification, for example, the Naïve Bayes Classifier already being used in many researches and has been proven to run

text classification cases well. On the other side, the XGBoost algorithm has recently dominated applied machine learning and Kaggle competitions for structured or tabular data. The beauty of the XGBoost algorithm is its scalability, which allows for rapid learning via parallel and distributed computing while still utilizing memory efficiently. At the end, the authors want to know which is the best for handling case like this one, Hopefully, this research can be used as a reference for the future research.

### Evaluation and Validation

Evaluating a model is a crucial step in creating a successful machine learning model. The counts of test records correctly and incorrectly predicted by a classification model are used to evaluate the model's performance. For the case of imbalanced data, the best classification matrix that we can use is the F1-Score, this is because when the number of samples in one class outnumbers those in another, the classification accuracy deteriorates [16]. F1-Score itself is come from the weighted harmonic mean of precision and recall [17]. Because the minority class only counts for a small percentage of the data, excluding all of the minority class samples has minimal impact on overall accuracy. It is also important to note that the F1-score is a well-balanced combination of precision and recall. As a result, it is a more acceptable statistic for evaluating minority class categorization [18]. In addition, the authors also used the confusion matrix and ROC AUC score and plot to make sure how the model works.

The model validation for this research used 2 kinds of validation. First is K-fold Cross Validation technique. This method is mostly used to create model predictions and measure the performance of a predictive model when it is applied in practice. Furthermore, K-fold cross validation is utilized to reduce bias from the data [19]. The technique includes only one parameter called  $k$ , which specifies the number of groups into which a given data sample should be divided [20]. When a precise value for  $k$  is specified, it can be substituted for  $k$  in the model reference. For this research, the authors choosed value of  $k=10$ , which means, the data will be divided into 10 groups. The first group will become the validation data and the rest will be train data. After the process is done, the second group will become the validation data and the rest will be the train data and this happens over and over again until all fold get part as validation data.

The second technique that used for validation is splitting with ratio 80% data train and 20% data test [21]. This is actually the commonly

used for validation model. The reason of using this technique is because the authors want to test the model also with the data that is not in the train or validation data. This is like a simulation of application this model to real-world data. With this technique, the author can also provide more evaluation metrics such as Confusion Matrix and ROC-AUC Curve to get better understanding about the model.

### RESULTS AND DISCUSSION

The main purpose of this research is to know what is the public opinion in Indonesia about the Covid-19 vaccination also create the machine learning model that can classify the tweets into 2 classes, which are positive and negative. In addition, this research also figures out which algorithm that have best performance between the Naïve Bayes and XGBoost Classifier to handle this kind of case.

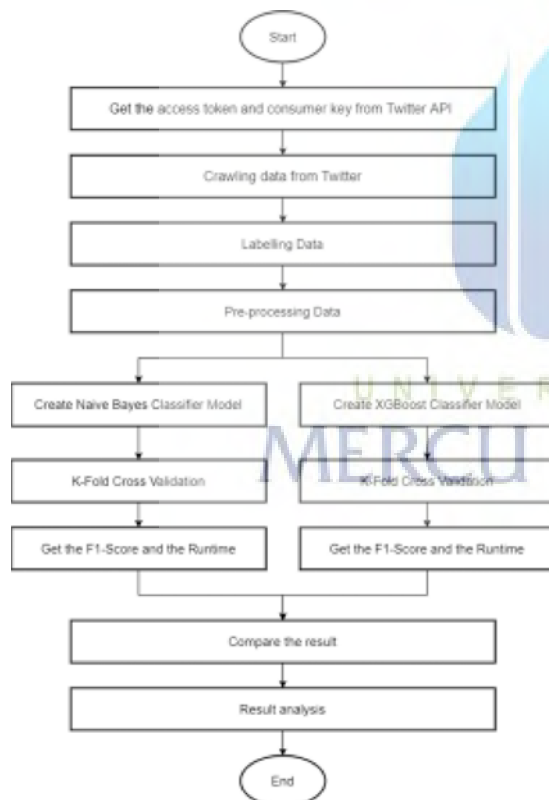


Figure 3. Research process using K-Fold Cross validation as validation

From Figure 1, we can see that the most of the tweets are positive sentiments. It probably means that people in Indonesia already well socialized about this vaccination. The tweets show that many people in Indonesia already known about the important of this vaccination to end this pandemic. The government and the

related parties are also can socialize the procurement of this vaccination well by continuing to invite people to vaccinate.

For the machine learning model, after applying the whole pre-processing data, the next step is to create the model. The model that used in this research are XGBoost classifier and Naïve Bayes Classifier where the process is done using Python as a tool. For increasing the performance of the model, the authors use the hyper parameter for each of the model. For Naïve Bayes Classifier, the hyper parameter that used is alpha, which are a laplace smoothing parameter. The default value for alpha is 1, but in this case the authors found that the optimum value of alpha is 0,6. For XGBoost Classifier, the hyper parameter that the authors used are n\_estimators. The function of n\_estimators is indicating the number of times the modeling cycle should be repeated. The value selection must be careful, because if we put too low value can cause underfitting, and if too high can cause overfitting. The typical values if between 100 to 1000. For this case, the optimum value for n\_estimator is 500. After creating the model is done, the next step is for evaluation process. It is done by using K-fold cross validation with the value of k=10. Table 7 show how the performance of both algorithms.

Table 7. Model Performance

Algorithm	F1-Score	Runtime
Naïve Bayes Classifier	0.945	134 ms
XGBoost Classifier	0.968	1 min 59 s

From Table 7, we can see that F1-Score for XGBoost classifier is higher than Naïve Bayes Classifier with a difference of only about 2%, where the XGBoost Classifier obtain 0.968 while the Naïve Bayes Classifier obtain slightly less which are 0.945. However, when look at the runtime, we can see that the Naïve Bayes Classifier can have a short training time than the XGBoost Classifier. The same thing happens in the testing process, the Naïve Bayes Classifier can predict the data faster than the XGBoost Classifier. Because the efficiency of the model is evaluated by the running time [22], it means that the Naïve Bayes Classifier run more efficient than the XGBoost Classifier.

The second technique, is the splitting 80:20 ratio. The preparation as same like the Cross Validation. The difference is the process after the pre-processing data. The flowchart process is shown in figure 4.

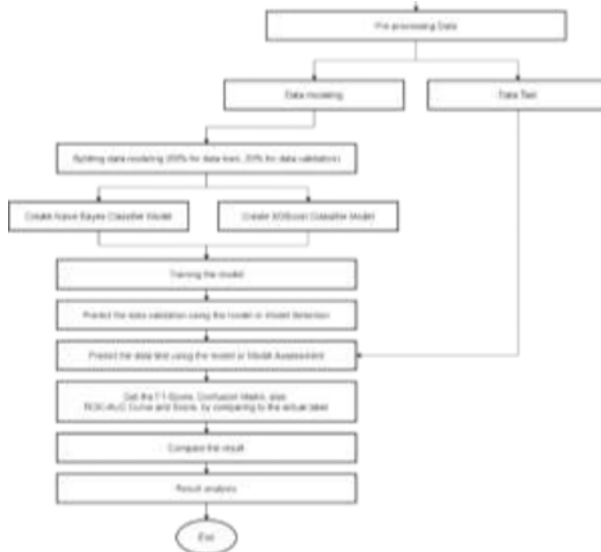


Figure 4. Research process using splitting 80:20 ratio as validation

In this technique, the authors separate data into 2 data, first the data that used for data modeling and the second is for data test. The data test here is used as a simulation when the model meets the data that never met before. It is come from approximately 20% from the whole dataset. The test data is used as a data that will be predicted by the model.

In Figure 4, we can see that the test data is not used in the training model process. This is for make sure that the model that train here has not met the data for validation before. To get the performance, the model will predict the test data without the label. At the end, the predicted label will be compared with the actual label. After running the model, the result is displayed in Table 8.

Table 8. Model Performance 2

Algorithm	F1-Score	ROC - AUC Score
Naive Bayes Classifier	0.947	0.950
XGBoost Classifier	0.976	0.882

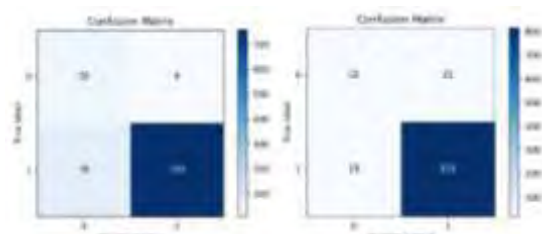


Figure 5. Confusion Matrix of Naive Bayes Classifier model (On the left), and Confusion Matrix of XGBoost Classifier model (On the right)

The result shows that the XGBoost Classifier have a bigger F1-score than the Naive Bayes. However, if we look closely to the confusion matrix in figure 5, it is shown that the Naive Bayes Classifier have a better performance to predict the negative label, while XGBoost have more problem to predict it. The reason why the XGBoost have a better F1-Score is because it predicts more accurate in the positive label, while in the negative label, it performs the opposite. These results are also in line with the ROC - AUC score in table 8.

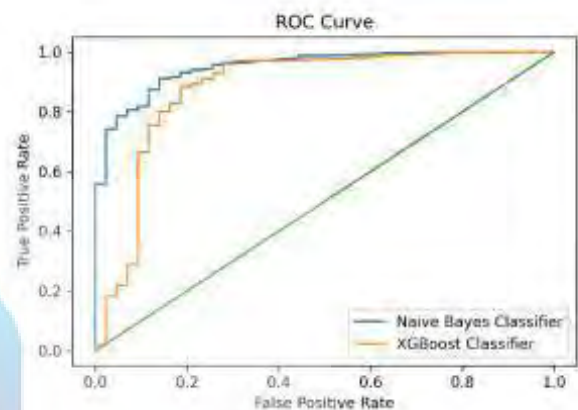


Figure 6. ROC – AUC curve of Naive Bayes and XGBoost Classifier model

In figure 8, it shows how both of the model working on predicting the data. Classifiers that create curves closer to the top-left corner of the ROC space perform better, whereas those that produce curves closer to the 45-degree diagonal of the ROC space perform less in performance. From the curve, we can see that the Naive Bayes Classifier model have a closer to the top-left corner, it tells that the Naive Bayes Classifier model is more capable of distinguishing between classes for this case.

Table 9. Comparison with previous research

Study	Additional technique	Model	F1-Score
Abdelrahman I. Saad [3]	Bag of Words (BoW)	XGB	0,623
		Naive Bayes	0,71
This Study	TF-IDF + SMOTE	XGBoost	0.968
		Naive Bayes	0.945

Compare to the research before [3], this study using TF-IDF instead using BoW. This research also using the SMOTE technique to



handle the imbalanced data. The result shows that the F1-Score for the both Naïve Bayes Classifier and XGBoost Classifier algorithm in this research have a better performance than the previous research. It means that the SMOTE technique is well executed and in the end also have an impact on model that successfully handle the data that have good performance.

## CONCLUSION

Sentiment analysis is a branch of research that investigates at how people express opinions in text that usually do on social media platforms such as Twitter. This study collected a dataset of tweet about vaccination in Indonesia that collected from May until October 2021. Twitter Data or we called as Tweet cannot be used directly, so we already perform the pre-processing, and because the data is imbalanced, we also already use SMOTE method to make it balanced. Our proposed model used 2 algorithms to figure it out who have the better performance of classifying tweets as positive and negative. The model evaluation here using 2 methods, those are 10-fold Cross Validation and Splitting 80:20 ratio to make sure the model works well.

The result shows that more Tweets responded positively to this vaccination program, as seen in figure 1, where the positive response was 91.2% compared to the negative response which was only 8.8%. It means that the socialization has been conveyed well and can be understood by the people in Indonesia, and the people has also understood the importance of this vaccination for the common good and for himself.

In addition, this study also reveals that XGBoost Classifier algorithm have a better F1-Score for predicting the tweets compare to Naïve Bayes Classifier. However, from the confusion matrix and ROC-AUC curve, we can see that the Naïve Bayes Classifier is do better in classifying the tweets both for the positive and negative label. The Naïve Bayes Classifier is also more efficient if we look at the runtime process. So, the authors have a conclusion that in this research, Naïve Bayes Classifier have a better performance than the XGBoost Classifier. As we can see, a higher score does not always indicate a better final result. This research also finds out that the application of the SMOTE method also had a good impact on imbalanced data such as this study. It can be seen from the high F1-Score which indicates that the model is good at learning between positive and negative labels.

## ACKNOWLEDGMENT

This research was supported by Universitas Mercu Buana lecturer. We thank our colleagues from Universitas Mercu Buana also who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

## REFERENCES

- [1] M. Abdullah and M. Hadzikadic, "Sentiment Analysis of Twitter Data: Emotions Revealed Regarding Donald Trump during the 2015-16 Primary Debates," 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), 2017.
- [2] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019.
- [3] A. I. Saad, "Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques," 2020 16th International Computer Engineering Conference (ICENCO), 2020.
- [4] "New user FAQ," Twitter. [Online]. Available: <https://help.twitter.com/en/resources/new-user-faq>. [Accessed: 19-Nov-2021].
- [5] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2019.
- [6] B. M. Pintoko and K. M. L., "Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier," e-Proceeding of Engineering, vol. 5, no. 3, pp. 1–10.
- [7] "Vaccines," World Health Organization. [Online]. Available: <https://www.who.int/travel-advice/vaccines>. [Accessed: 19-Nov-2021].
- [8] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek," 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 2020.
- [9] R. Y. Hayuningtyas and R. Sari, "Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata Tmii Menggunakan Naïve Bayes Dan Pso," Jurnal Techno Nusa Mandiri, vol. 16, no. 1, pp. 37–42, 2019.

- [10] J. Brownlee, "A Gentle Introduction to XGBoost for Applied Machine Learning," *Machine Learning Mastery*, 17-Feb-2021. [Online]. Available: <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>. [Accessed: 19-Nov-2021].
- [11] J. Bhattacharyya, "Understanding XGBoost Algorithm In Detail," *Analytics India Magazine*, 02-Nov-2020. [Online]. Available: <https://analyticsindiamag.com/xgboost-internal-working-to-make-decision-trees-and-dedupe-predictions/>. [Accessed: 19-Nov-2021].
- [12] T. Perry, "What is Tokenization in Natural Language Processing (NLP)?," *Machine Learning Plus*, 01-Feb-2021. [Online]. Available: <https://www.machinelearningplus.com/nlp/what-is-tokenization-in-natural-language-processing/>. [Accessed: 19-Nov-2021].
- [13] E. Dwianto and M. Sadikin, "Analisis Sentimen Transportasi Online pada Twitter Menggunakan Metode Klasifikasi Naïve Bayes dan Support Vector Machine," *Jurnal Ilmiah Teknik Informatika*, vol. 10, no. 1, p. 94, 2021.
- [14] A. Indrawati, H. Subagyo, A. Sihombing, W. Wagiyah, and S. Afandi, "Analyzing The Impact Of Resampling Method For Imbalanced Data Text In Indonesian Scientific Articles Categorization," *Baca: Jurnal Dokumentasi Dan Informasi*, vol. 41, no. 2, p. 133, 2020.
- [15] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, "An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set," *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, 2018.
- [16] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3525–3539, 2020.
- [17] M. Khalafat, J. S. Alqatawna, R. M. H. Al-Sayyed, M. Eshtay, and T. Kobbaey, "Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 15, no. 14, p. 90, 2021.
- [18] Y. Yang, H. Yeh, W. Zhang, C. Lee, E. Meese and C. Lowe, "Feature Extraction, Selection, and K-Nearest Neighbors Algorithm for Shark Behavior Classification Based on Imbalanced Dataset", *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6429-6439, 2020.
- [19] D. Fitriana, S. Dwiasnati, H. H. H, and K. A. Baihaqi, "Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes," *Faktor Exacta*, pp. 1–9, 2020.
- [20] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," *Machine Learning Mastery*, 03-Aug-2020. [Online]. Available: <https://machinelearningmastery.com/k-fold-cross-validation/>. [Accessed: 19-Nov-2021].
- [21] S. Qaiser, N. Yusoff, F. K. Ahmad, and R. Ali, "Sentiment Analysis of Impact of Technology on Employment from Text on Twitter," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 14, no. 07, p. 88, 2020.
- [22] D. Zhang and Y. Gong, "The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure," *IEEE Access*, vol. 8, pp. 220990–221003, 2020.

## WORKING PAPER

This paper is a complete material for a journal article with the title "Sentiment Analysis from Twitter about Covid-19 Vaccination in Indonesia using Naive Bayes and XGBoost Classifier Algorithm". This working paper contains all the material from the Final Project research results. This paper presents several sections which consists of literature review, analysis and design, source code, dataset, experiment stage, and results of all the experiments.

Chapter 1 discusses the literature review which contains journal articles that form the basis or foundation for this research. Chapter 2 describes the analysis and design of the research conducted, the design in this chapter is explained using a Flowchart. Chapter 3 describes the source code used in this research, there is an explanation of how each code is used to process the existing data. Chapter 4 describes the dataset used in this study, including explanations, how to obtain data, data attributes, and adjustments to the final data that are ready to be processed. Chapter 5 describes the stages that must be passed in processing the dataset into the expected research results. Chapter 6 is the last part of this paper which explains the overall results of the experiments that have been carried out, including the calculation of performance results using the Confusion Matrix and F1-Score, as well as the results of the comparison of the two algorithms that have been used.

