ISSN: 2460 – 7223

Twitter Sentiment Analysis on The use of Sinovac Vaccine in Indonesia

1st Sherli Trisnawati
Faculty of Computer Science
Institute of Informatics and Business (IIB) Darmajaya
Bandar Lampung, Indonesia
sherlitrisnawati@darmajaya.ac.id

2nd Akhmad Unggul Priantoro
Faculty of Computer Science
Institute of Informatics and Business (IIB) Darmajaya
Bandar Lampung, Indonesia
akhmadunggul@darmajaya.ac.id

3rd Chairani Fauzi
Faculty of Computer Science
Institute of Informatics and Business (IIB) Darmajaya
Bandar Lampung, Indonesia
chairani@darmajaya.ac.id

4th Riko Hermanto
Faculty of Computer Science
Institute of Informatics and Business (IIB) Darmajaya
Bandar Lampung, Indonesia
rikohermanto@darmajaya.ac.id

5th Hary Sabita
Faculty of Computer Science
Institute of Informatics and Business (IIB) Darmajaya
Bandar Lampung, Indonesia
hary.sabita@darmajaya.ac.id

***Abstract*— Coronavirus Diseases (Covid-19) was reported the first time in Wuhan, Hubeu Province, China 2019. On March 11th 2021 World Health Organization (WHO) declared covid-19 as a world pandemic. To reduce the number of deaths and the number of the transmission of covid-19 is by the vaccination. Several vaccines have been evaluated by WHO to against covid-19, Indonesian Government officially announced the use of Sinovac Vaccine produced by Sinovac Life Science Co, China. This vaccination topic becomes one of the massive topics discussed by the Indonesian people and various responses on social media such as Twitter. The technique used, is crawling tweets from Twitter users worldwide using the English language from May 24th 2021 – August 31st 2021 with the keyword "sinovac vaccine". This study aims to analyze public sentiment regarding the usage of sinovac vaccine in Indonesia. The method used is Naïve Bayes because it has a simple algorithm with high accuracy. The result shows the classification accuracy rate is 80.99% and majorities' responds are neutral and positive. However, the preprocessing data has the rule to get excellent result for the accuracy. Future study can also classify the tweets into different queries and other classification methods can be applied such as Super Vector Machine and KNN.**

Keywords*— *Covid-19, Sinovac Vaccine, Sentiment Analysis, Naïve Bayes

I. INTRODUCTION

Coronavirus Diseases (Covid-19) was reported the first time in Wuhan, Hubeu Province, China 2019. On March 2nd 2021 Indonesian Government announced have found the first two cases of covid-19 in Indonesia and on March 11th 2021 World Health Organization (WHO) declared covid-19 as a world pandemic[1].

To reduce the number of deaths and the number of the transmission of covid-19 is by the vaccination. Researchers around the globe have developed vaccines to overcome the situation. There are several vaccines have been evaluated by WHO to against covid-19, these include AstraZeneca/Oxford vaccine, Johnson and Johnson, Moderna, Pfizer/Biotech, Sinoparm, and Sinovac[2]. From the various types of covid-19 vaccines, Indonesian Government officially announced use Sinovac Vaccine produced by Sinovac Life Science Co, China[3]. Sinovac Vaccine have been passing the evaluation for the clinical trials and the level of effectiveness. In Indonesia reaches 65.3% of effectiveness[4], in Brazil reaches 50% of effectiveness and in Turkey reaches 83.5 of effectiveness. And the target for the vaccination in Indonesia is 208.382.680[5].

However, this vaccination topic becomes one of the massive topic discussed by the Indonesian people and rise the series of the questions on social media such as Twitter and most of the people discussed about the safety of the vaccine process. There are currently 206 million active Twitter users worldwide and Indonesia entered in to the 6th position with 15.7 million users[6].

Sentiment Analysis is one of the part of machine learning domain. This studies that analyze the opinions, sentiments, attitudes, research, and emotions from the public point of view toward the public services, problems, or to any massive issues or topics.

It is interesting to know how the world community responds and given their opinions and perspectives toward the vaccination process for sinovac vaccine on social media. Thus, the response will be divided into two group of sentiment: sentiment positive, sentiment negative, and sentiment netral. The data source in this study were taken from tweets user worldwide written in English with the keyword "Sinovac Vaccine" using the crawling technique using Twitter API in the range of May 24th 2021 – August 31st 2021. The result contributes to the government or any stake holder that related to this topic.

The next section covers the related work done on sentiment analysis, while chapter three explains the research methodology. Next, the segment four will describe the result and discussion.

II. LITERATURE STUDYS

A. Sentiment Analysis

Sentiment analysis is known as opinion mining, which is a process of understanding, extracting, and processing textual data automatically to obtain sentiment information in a sentence. In texts, sentences, and messages, the fundamental

concept of sentiment analysis is to detect text polarity. In this sentiment analysis is use the TextBlob as can see on fig.3.

TextBlob is python library for Natural Language Processing (NLP)[7]. TextBlob has semantics labels that help with fine-grained analysis. Higher subjectivity means that the text contains personal opinion rather than factual information. TextBlob calculates subjectivity by looking at intensity, which the intensity determines whether a word modifies another word.

Preprocessing is to minimize the vocabulary of words contained in the text message. Below is the pre-processing task that used in this research[8].

1) Case folding: changing all the character on dataset to lowercase. The purposes of case folding are to make the same letterform.
2) Document filtering: removes the special characters and symbols, punctuation marks, urls address, and usernames.
3) Tokenizing: separates words in each sentence into the separate words.
4) Stopword removal: remove the unimportant words or have no meaning such is the conjunctions or the prepositions.
5) Stemming: transforms a word into a root of word.

In this study, the sentiment divided into three classes, sentiment positive, sentiment negative, and sentiment neutral. Thus, this study uses Naïve Bayes (NB) algorithm. Naïve Bayes methods are a set of supervised learning algorithm based on applying Bayes' theorem with the Naïve assumption of independence between every pair of features. Below is the Bayes theorem.

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

Where:

TABEL 1 BAYES THEOREM

P(A)	:	The prior probability of belonging to class i
P(B)	:	Normalizing constant, or probability of seeing the given input vector over all possible inputs vectors
P(B|A)	:	The conditional probability of seeing input vector B given we know the class is A

B. *Related Study*

Several previous studies of sentiment analysis in Indonesia regarding the vaccination have done reviewed. These include how the Indonesian people given their reaction on facebook page that own by the Ministry of Health[9], in this research found that the Indonesian citizen given their negative opinion through the process of the vaccination in Indonesia. Analyzing the public opinion regarding the covid-19[10] and the result shown that the community had the positive reaction during the covid-19 pandemic, public opinion on online learning during the pandemic in Indonesia[11] and the result is most the community given their negative opinion on that.

The researchers on these research are using the machine learning algorithm Naïve Bayes to analysis the sentiment.

III. METHODOLOGY

This study uses the python as the programming language with the several libraries such as Tweepy and TextBlob and rapidminer. There are six steps in this study: 1). Data preparation including the crawling and labelling, 2). Pre-processing including case folding, filtering, tokenizing, stopword removal, and stemming, 3). Feature selection, 4). Modeling, 5). Label prediction, and 6). Result including the analyzing and result. As can be seen in fig. 1.

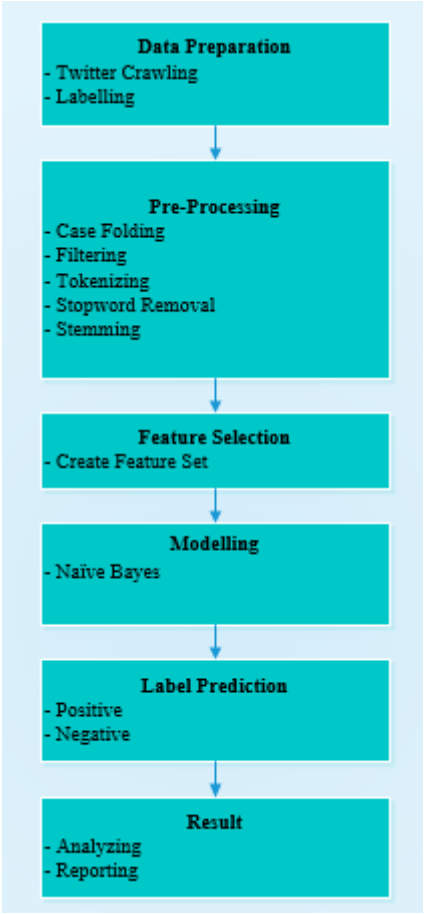

Fig. 1 Research Methodology

A. *Data Preparation*

The key word for crawling data on Twitter is "Sinovac vaccine" with the tweets written in English using English language. In this study there are fourth variables that were retrieved: "date" where the user posts the tweets, "user" means username that posts the tweets, "tweet" mean the tweet or opinion that posted by the Twitter users, and "sentiment". Search tweet API, Tweepy, and TextBlob are libraries that are used in python for crawling data from May 24th 2021 – to August 31st 2021. As can be seen in fig. 2.

```
import tweepy
import re
from textblob import TextBlob
import numpy as np
```

```
[ ]  api_key = "zCbw9hBTSAG3nlW1oNFN0aoS4"
     api_secret_key = "yveQJ2P9Cf90LV1hbvYFPFANFp8KcvJaKgfCRSTIyRnNLTpgWK"
     access_token = "180278318-IOPrftLCRztWf1d2o36m3nZR8noLugVNd4pp97tD"
     access_token_secret = "jDk1xpW9dyKsHWHyqnyO0jDh6QgoA1yd4kw7fxDbdVMmy"
```

```
[ ]  auth = tweepy.OAuthHandler(api_key, api_secret_key)
     auth.set_access_token(access_token, access_token_secret)
     api = tweepy.API(auth)
```

Fig. 2 Crawling with API Key

From crawling data on May to August 2021, there are 78.145 rows raw tweets from users worldwide with three group polarity; positive, negative, and neutral. As mentioned in the previous stage, the polarity for each tweets above had been classifying by TextBlob. On this raw tweets have a lots of noise such as the url link, emoji's, and special character, as can see in table 1.

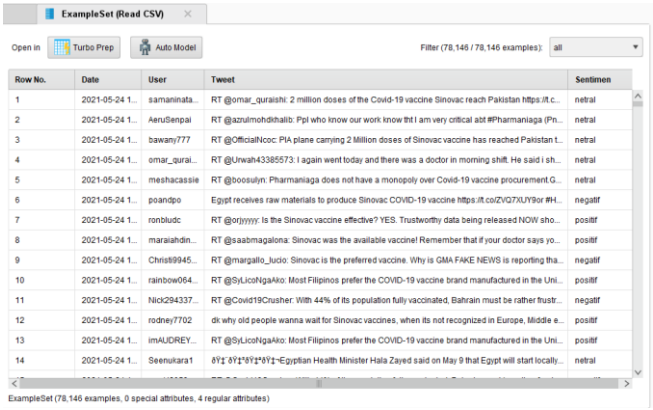

Fig. 3 Raw Tweets

Thus, the next step is pre-processing to get the clean dataset.

B. *Pre-Processing*

- Case folding: in this study all the tweets changing to the lowercase.
- Document filtering: all the special character, has tag, url address, etc are removed.
- Tokenizing: chopping the text or into the words.
- Stopword: the irrelevant words are removed.
- Stemming: converting the words into the root words.

After the pre-processing process using Rapidminer has been done, there are 41.605 rows clean tweets. As can see in fig. 3.

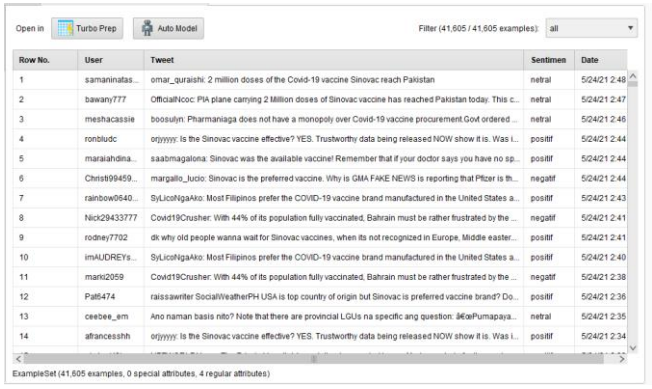

Fig. 4 Clean tweets

Most of the polarity for clean tweets are neutral, positive, and negative as can see in fig. 4.

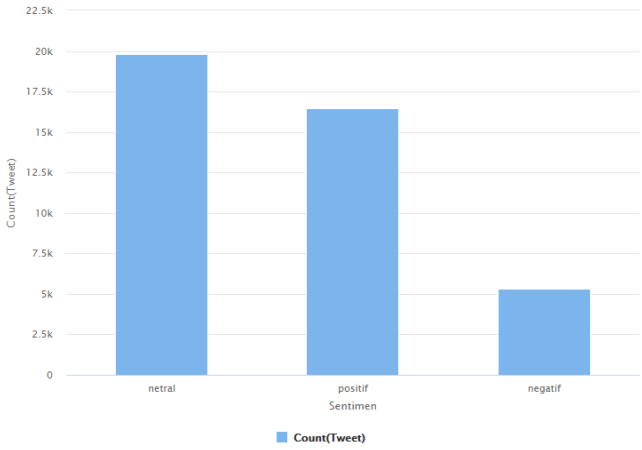

Fig. 5 Visualization Tweets Polarity

C. *Feature Section*

After getting the result in pre-processing data, there are 16.461 tweets for positive tweets, 5.313 tweets for negative tweets, and 19.831 tweets for neutral tweets. As can be found in table 2.

TABEL 2 CLEAN TWEET

Clean tweets	**41.605 tweets**
Neutral	19.831 tweets
Positive	16.461 tweets
Negative	5.313 tweets

D. *Modeling*

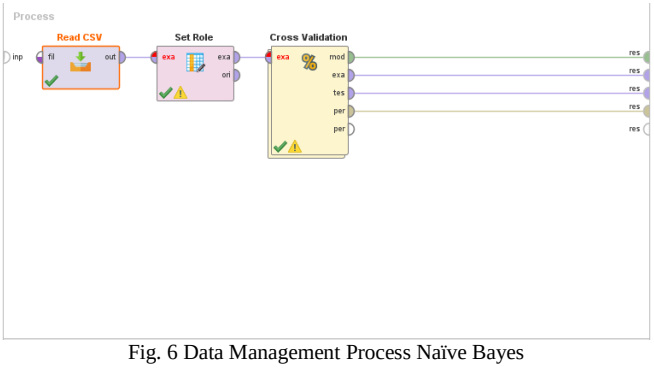

Fig. 6 Data Management Process Naïve Bayes

To build the modelling management using Naïve Bayes algorithm, is using the Rapidminer. As can see in fig. 6 and fig 7. The dataset is divided into two, namely training data and testing data which are automatically done by the Cross Validation on fig. 6. Cross-validation method is used to avoid overlapping in testing data and the performance is for performance evaluation on classification task.

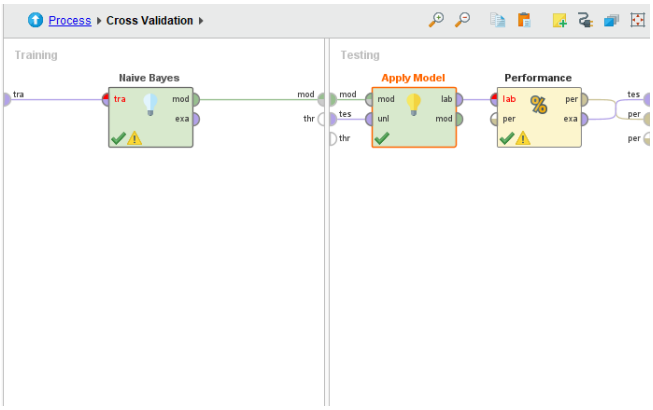

Fig. 7 Data Management Process Using Naïve Bayes

The accuracy result for Naïve Bayes algorithm as can see bellow in fig. 8.

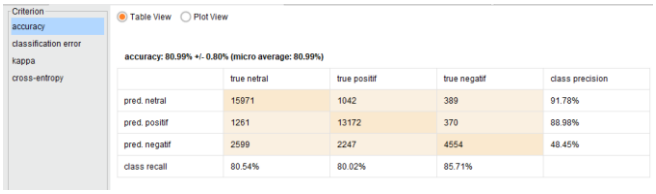

Fig. 8 Naïve Bayes Accuracy

E. Label Prediction

From the previous stage, as explained in fig. 8, the accuracy for the sentiments label is 80,99% as shown in next stage.

IV. RESULT AND DISCUSSSION

The result shows the accuracy using Naïve Bayes algorithm with rapidminer is 80, 99% with three polarities: positive, negative, and neutral. Due to the public awareness about the vaccination against the covid-19, most of the community worldwide has given neutral and positive reaction.

Thus, the writing style and the structure of the language posted on Twitter by user's also huge challenges for sentiment analysis process in this study. Especially on pre-processing process with the huge of dataset text.

V. CONCLUSION AND RECOMMENDATION

This study aimed to analyzed sentiments on the use of Sinovac Vaccine in Indonesia according to positive, negative, and neutral polarities on social media Twitter. Based on the results, the accuracy using Naïve Bayes algorithm is 80,99% and majorities responds are neutral and positive. However, public awareness to get the vaccine against the covid-19 has the main role in this period.

Future study can also classify the tweets into different queries and other classification methods can be applied such as Super Vector Machine and KNN.

REFERENCES

[1] "Listings of WHO's response to COVID-19." https://www.who.int/news/item/29-06-2020-covidtimeline (accessed Nov. 03, 2021).

[2] "COVID-19 Vaccines Advice." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines/advice (accessed Nov. 02, 2021).

[3] "Consulate General of The Republic of Indonesia In Chicago, THE UNITED STATES OF AMERICA," *Kementerian Luar Negeri Repulik Indonesia*. https://kemlu.go.id/chicago/en (accessed Nov. 03, 2021).

[4] "BPOM Issues EUA for Sinovac COVID-19 Vaccine," *Sekretariat Kabinet Republik Indonesia*, Jan. 11, 2021. https://setkab.go.id/en/bpom-issues-eua-for-sinovac-covid-19-vaccine/ (accessed Nov. 03, 2021).

[5] S. Moutinho, "Chinese COVID-19 vaccine maintains protection in variant-plagued Brazil," 2021, doi: 10.1126/SCIENCE.ABI9414.

[6] "Twitter: most users by country," *Statista*. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed Nov. 02, 2021).

[7] "TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation." https://textblob.readthedocs.io/en/dev/ (accessed Nov. 10, 2021).

[8] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From The Perspective of Sinovac and Pfizer," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Apr. 2021, pp. 122–127. doi: 10.1109/EIConCIT50028.2021.9431852.

[9] A. Harun and D. P. Ananda, "Analysis of Public Opinion Sentiment About Covid-19 Vaccination in Indonesia Using Naïve Bayes and Decission Tree," vol. 1, p. 6, 2021.

[10] M. Syarifuddin, "ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN," vol. 15, no. 1, p. 6, 2020.

[11] U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," vol. 5, p. 7, 2021.