Jurnal Borneo Administrator
Volume 18 (3) 2022: 213-230
P-ISSN: 1858-0300, E-ISSN: 2407-6767
DOI: 10.24258/jba.v18i3.992
Accredited by Kemenristek/BRIN No.148/M/KPT/2020

ARTICLE

Sentiment Analysis for Opinions on the COVID-19 Vaccination Program Using a Naive Bayes Classifier

Hasna Melani Puspasari and Pujiatmo Subarkah

Centre of Data and System Information, The National Institute of Public Administration, Jakarta, Indonesia

How to cite: Puspasari, Hasna Melani & Subarkah, Pujiatmo. (2022). Sentiment Analysis for Opinions on the Covid-19 Vaccination Program Using a Naive Bayes Classifier. *Jurnal Borneo Administrator*, 18(3), 213-230. https://doi.org/10.2428/jba.v18i3.992

Article History
Received: 5 November 2021
Accepted: 12 September 2022

Keywords:
COVID-19 Vaccination,
Naive Bayes Classifier,
YouTube,
Web Crawling,
Sentiment Analysis.

ABSTRACT
Implementing the COVID-19 vaccination program was not easy because it received various responses from the public. This study will explore public sentiment on the program which was taken based on public comments on several videos on YouTube using data crawling techniques using Coberry tools. The data was public opinions in *Bahasa Indonesia* that will be collected based on videos from official news channels with high engagement in the last eight months. Sentiment analysis was carried out using the Naive Bayes Classifier method that can be used to make deep analysis through filtering and classifying the opinions of the Indonesian people, as stated on YouTube. This research showed that public sentiment was dominated by negative sentiment based on some public doubts regarding the side effects of vaccines and the government follow-up regarding the country's economic recovery. Compared with the previous studies, the conclusions of sentiment obtained by this study were not the same due to differences in data sources and the selection of timeframes for community responses. An analysis of the public responses that have been carried out using a data collection method like this one will be very effective in providing an overview of the public desire to facilitate policymakers in formulating policy designs for the public interest.

A. INTRODUCTION

The world was experiencing a massive wave of infection with the novel coronavirus disease (COVID -19) after the WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China. The rapid increase in cases outside China led WHO to announce that the outbreak was a pandemic. In Indonesia, the pandemic started in March 2020, after two people were confirmed positive for COVID-19. Since then, COVID-19 has spread massively to more than 700,000 infected people by the end of 2020. Because of that, the Indonesia Government was trying to accelerate the COVID-19 vaccination program in early 2021 to achieve herd immunity in 2022. Instead of allowing more and more people to be infected to create immunity against COVID-19, the government has chosen to achieve immunity by vaccination (Hardy, 2020). This program was expected to be one of the ways to end this pandemic.

* Corresponding Author
Email : puspasari1995@gmail.com, pujiatmo.subarkah@gmail.com

The Vaccination Program started on January 13th, 2021. It was split into four phases that allow the healthcare workers to receive the first batch of vaccines, followed by public servants and then other members of the public services. As of September 28th, 2021, as many as 88,521,137 Indonesian people have received their first dose, and 49,655,718 have received their second dose vaccinations. Meanwhile, as of September 28th, 2021, 10,412,252 people have received their first dose, and 7,687,969 people have received their second vaccination dose in Jakarta.

Many creative ways are implemented to encourage vaccines among people who are still hesitant to get vaccinated. Local authorities offer many incentives in rural areas as a vaccination campaign for older residents to get vaccinated. Many older adults thought vaccines did not prevent Covid-19; instead, they would give some serious diseases and even death. The other things related to their belief were that many were concerned about whether the vaccines were permissible or prohibited by Islam because there was a lot of terrible news that the vaccine contained pork. It was worsened by the death of a 22-year-old-man who died one day after getting AstraZeneca vaccine injections. This incident created negative sentiment that challenged the trust of the community.

Hesitancy and misinformation have hampered vaccination programs (Hernikawati, 2021). It has turned out to be a severe challenge to the vaccination campaigns. The Ministry of Communication and Information Technology has detected at least 70 fake news items from October 2020 to January 18th, 2021, regarding the COVID-19 vaccine. A civil society group Anti-Slander Society (MAFINDO), identified three main groups that spread fake news regarding vaccines, religious background, anti-Chinese, and anti-west bias. Vaccination would not be smoothly implemented without good communication because media literacy is relatively low. People tend to believe clickbait and fake news widely spread on social media.

The digital transformation not only influenced business and made the world more accessible, but it also changed how we communicate through social media, which has changed social interaction. Social media refers to the means of interactions among people in which they create, share, and exchange information and ideas, not just broadcast channels or sales and marketing tools. That is why vaccination misinformation on social media can potentially decrease public confidence or trust in the effectiveness of vaccines.

Based on these problems, we aimed to analyze the sentiment for COVID-19 vaccination programs; it is used to gain insights into how customers feel about specific topics and detect urgent issues in real-time before they spiral out of control. A data processing method is required. The model is derived based on the analysis of a series of training data and can be used to predict the class label of an object with an unknown label (Han., Pei., & Kamber, M, 2012). The analysis used the method called Naïve Bayes Classifier. Naïve Bayes is one of the data classification algorithms with simple probabilistic classifications that calculates a set of probabilities by calculating the frequency and combination of values in a particular data set (Tsangaratos & Ilia, 2016). Millions of opinions are posted every second on several media that help to know how the public responds to this vaccination program.

Rachman and Pramana wrote the previous journal (2020) and discussed sentiment analysis about the COVID-19 vaccination on social media. This journal used public opinion data collected from Twitter from October 25th to November 3rd, 2020. The conclusion was that more people had positive sentiments about the Covid-19 vaccine then. In addition, sentiment analysis was carried out by classifying tweets using the lexicon-based method or the positive-negative dictionary created by Liu, Hu, & Cheng (2005). The sentiment analysis carried out in this study will focus on public opinion after the vaccination program is running. In addition, this study will also use a different validated method considered to have good accuracy.

Sentiment analysis is often used to understand the target audience's thoughts and learn what the audience needs. This research can direct an effective campaign or promotion to create successful governance aimed at vaccination programs. The government can also immediately anticipate fake news that is still believed by the public and find out what is happening in the public environment.

B. LITERATURE REVIEW

Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning to identify helpful information and related knowledge from various large databases (Kadafi, 2018). Text mining (text analytics) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform free (unstructured) and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms (Kolluri., Razia., & Nayak, 2020). Text mining differs from data mining, which focuses on discovering interesting patterns from large databases rather than textual information. In contrast, a text dealing with certain features is relatively unstructured and usually requires preprocessing (Anggraini., Harahap., & Kurniawan, 2021). Sentiment classification is the one branch of the field of Text Mining that identifies opinions and labels them as positive, negative, or neutral (Hermansyah & Sarno, 2020). Opinions mean people's emotions that can be expressed through any tools, such as social media, about issues currently being discussed. That is why sentiment classification can be crucial in evaluating some problems. The main purpose of sentiment classification is to determine the polarity of positive, negative, and neutral sentiments.

One of the algorithms used in classification is Naive Bayes, a probabilistic machine learning model used for classification tasks. The Naive Bayesian algorithm uses the probability of the events for its purpose, which assumes no interdependence amongst the variables (Shah., Patel., Sanghvi., & Shah, 2020). The crux of the Classifier is based on the Bayes theorem. The flow of the Bayes method is described in the equation below (Asmiati & Fatmawati, 2020).

$$P(d) = \frac{P(d|c)\,P(c)}{P(d)}$$

Where c is the next category to be classified, and P(c) is the prior probability of the text category c. At the same time, d is a text document representing a set of words.

To feed an outstanding model and get output in probabilities, measuring the effectiveness of our model, confusion matrices are used to come into the limelight. The confusion matrix itself is a performance measurement for machine learning classification. The performance measurement for classification problems where output can be two or more classes, through this matrix, the accuracy, error rate, precision, and recall values can be known (Massy, 1964). The confusion matrix is shown as follows.

Table 1. Confusion Matrix of Model Result

Class	Positive Classified	Negative Classified
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True negative)

Source : (Luque., Carrasco., Martín., & de las Heras., 2019)

The Confusion Matrix is calculated with TP, FN, FP, and TN. TP (True Positive) is the amount of positive data with an actual truth value. In contrast, FN (False Negative) is the

amount of negative data the system considers to have a false truth value. FP (False Positive) is the amount of positive data that the system considers to have the truth value false. While TN (True Negative) is the amount of negative data, the system finds true truth value.

In conducting sentiment analysis, Permadi (2020) on their research used the Naive Bayes method to analyze data from restaurant reviews in Singapore. Based on their study, Naive Bayes shows an accuracy of 73,33% to classify satisfied or unsatisfied judgments. Meanwhile Ikasari & Widiastuti (2021), based on Twitter data, research using the Naïve Bayes Classifier method shows the results of tweet classification have a positive or negative tendency so that the system's accuracy in sentiment analysis of tweets on Twitter MRT Jakarta is 95.88%. The other research related to the Naive Bayes algorithm shows that Naive Bayes can be used in classifying majors found in *SMA Negeri 1 Kampar Timur*, namely Natural Sciences and Social Sciences, respectively at 70% and 30%, producing high accuracy of 96.19% (Mustakim et al., 2018). The Naive Bayes Classifier is proven to have good model accuracy even with small data (Feng., Guo., Jing., & Sun, 2015).

Since the vaccination program rolled out, various local groups have rejected the vaccination. The rejection of COVID-19 vaccinations reveals the negative sentiments in specific contexts such as socio-cultural backgrounds, including personal beliefs like religion. In Indonesia, a country with the world's largest Muslim population, the rejection is mostly because of the *halal* (permissible in Arabic) status concerned. However, structural problems such as a lack of trust in the arrangement for developing and distributing vaccines globally and political views also contributed.

The sentiments can influence the progress toward herd immunity which plays a role in ending the COVID-19 pandemic. People like expressing sentiment. Happy or unhappy. Like or dislike. Praise or complain. Good or bad. That is, positive or negative. People want to share their opinions through social media because the platform is free to share, and people can exchange ideas with anyone. YouTube is also one of the platforms that people often use to voice their opinions on various current issues (Möller., Kühne., Baumgartner., & Peter, 2019). Many people give their feedback through the comment column on the uploaded video. Some videos or news about Covid are uploaded on YouTube.

The sentiment analysis of public opinion on the Covid-19 vaccination has also been carried out several times. Three of them conducted a sentiment analysis of public opinion on Twitter before the vaccination program started. The research done by Rachman and Pramana (2020) used a Lexicon-based approach that assesses polarity based on assigning a value to each word meaning in the sentence, where one is for words with positive meanings and -1 for words with negative meanings (Kolchyna., Souza., Treleaven., & Aste., 2015). The result said public sentiment when vaccination had not started was more favourable (Rachman & Pramana, 2020).

Meanwhile, two other compared studies were written by Yulita, Nugroho, & Algifari (2021) and Firmasyah & Puspitasari (2021) that conducted sentiment analysis using the Naive Bayes Classifier like this study. However, both used public opinion data from Twitter. Meanwhile, this study used public opinion in its response to the issue of vaccination that was uploaded on YouTube. Yulita, Nugroho, & Algifari (2021) did the classification using training data from Kaggle, not mining data directly from Twitter. The result said that the public responded positively to the vaccination policy with a model accuracy rate of 93% (Yulita, Nugroho, & Algifari, 2021). In contrast, Firmasyah & Puspitasari (2021) used data from Twitter directly, as many as 1000 tweets on July 6-11, 2021. Labelling of training data was done manually using the SentiStrength library in Python. However, they stated that there were some tweet data whose sentiment values did not match the result of labelling using the library (Firmasyah & Puspitasari, 2021).

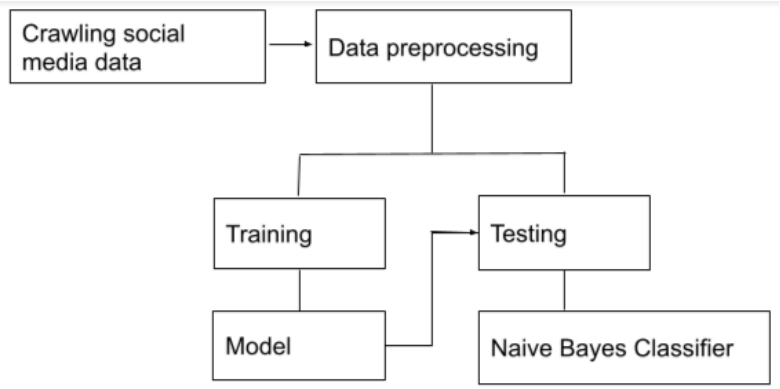
This research will use different data sources with other data labelling techniques. Data labelling will also be done manually using the Text blob library in Python. Opinion data will

be cleaned first by removing some "noise" words using certain techniques to improve the accuracy of manual labelling. Furthermore, the opinion sentences will be translated into English to match the library, which can only assess sentences in English. With this method's improvement, sentiment analysis results are expected to be more accurate and provide powerful insight.

C. METHOD

In this research, the device utilized is Rapid Miner, a tool with a faster data classification algorithm processing speed and accuracy superior to other tools (Faid., Jasri., & Rahmawati, 2019). RapidMiner is an easy-to-use visual environment for predictive analytics because no programming is required (Madyatmadja et al., 2021). Meanwhile, the data used in this study is obtained from the crawling process, taking opinion sentences from people's opinions in the comment columns on YouTube Channel. The scraping process data can be done using PHP and HTML (Viny Christant., Walda, & Sutrisno, 2020). In this research, gathering data from YouTube is obtained using a platform called Coberry. It is an exporting tool that scrapes comments of a video on YouTube based on a specific YouTube link and exports them as CSV (Hagemann, 2020).

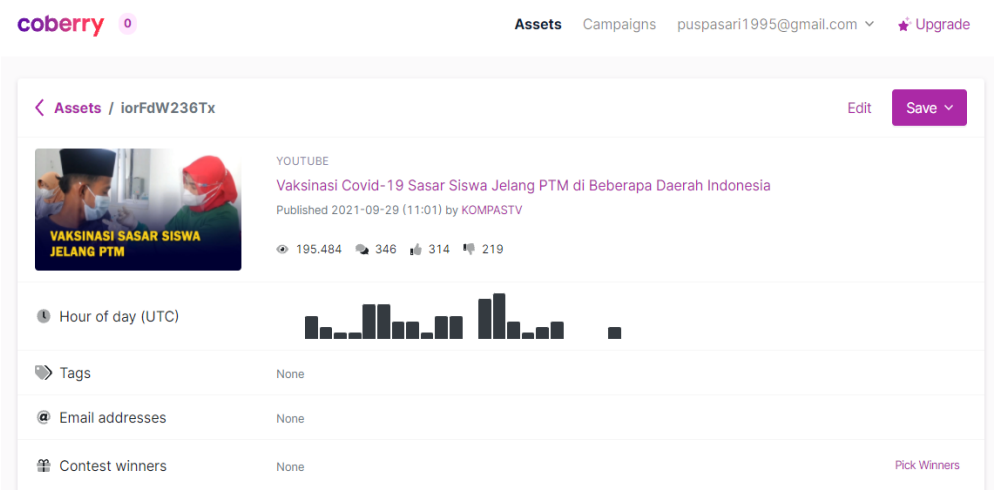
Several processes are carried out in this research; firstly, we collect people's opinions in the comments columns on YouTube Channel by using a crawler, followed by labelling, and then processing the data and text using the Naive Bayes method (Hermanto., Ziaurrahman., Bianto., & Setyanto, 2018). The method used in this study is an experimental research method (Fig. 1), which consists of (1) data collection, (2) initial data processing, (3) the proposed model, (4) testing the model, (5) Evaluation and Validation the model (Hofmann & Klinkenberg, 2014).



Source : Hofmann & Klinkenberg (2014)

Figure 1. The Steps of This Research

People like to share their opinions through social media because the platform is free to share, and people can exchange ideas with anyone. YouTube is also one of the platforms that people often use to voice their opinions on various current issues. Many people give their feedback through the comment column on the uploaded video. In some videos or news about Covid that are uploaded on YouTube, an analysis can be done regarding the tendency of people's sentiments towards the video (Nofriani, 2018). For this reason, using Coberry's help, several videos related to the latest covid news were selected, and then we crawled the comments to analyze the people's sentiments.



Source: Result of collecting YouTube data by Coberry, 2021

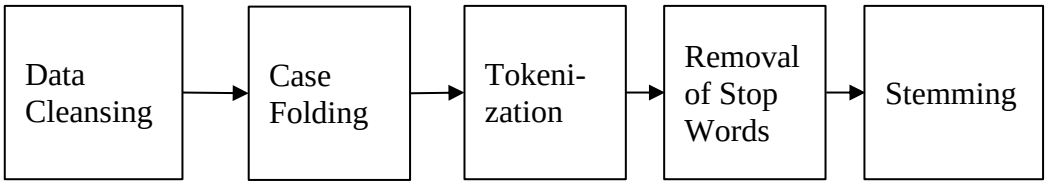
Figure 2. Overview of The Coberry Platform

In this case, the videos are selected based on news videos from several official channels that have had a high engagement in the last eight months. High engagement is assessed from the ratio of the number of viewers to the number of likes, dislikes, and comments, such as the viewers number more than 100,000 with response viewers rate more than 1% or the number of viewers less than 50,000 with a response rate more than 2% for the last four months. The official channels observed are official news portals such as CNN Indonesia, Kompas TV, and Metro TV News. In that period, nine videos were selected based on the keywords "*vaksinas*i" or "*vaksin*" that managed to attract the public's attention to express their opinion in the comment column. The comments column on that video will be crawled using the Coberry tool and then analyzed using the RapidMiner tool. The following is a list of videos and their channel origins.

"*Momen Saat Presiden Jokowi Disuntik Vaksin Covid-19 Untuk Kedua Kalinya*" from Kompas TV channel.

- "*Mengenal Efek Samping Usai Vaksinasi* Covid-19" from CNN Indonesia.
- "*Menolak Vaksinasi, Apa yang Akan Terjadi?*" from Metro TV News channel.
- "*Terbaru, 3 Orang Diduga Meninggal karena Astrazeneca, Ini Faktanya* - ROSI" from Kompas TV channel.
- "*Vaksinasi Berhadiah Kambing dan Ayam Ramai Peminat*" from CNN Indonesia channel.
- "BREAKING NEWS - *Pernyataan Presiden Jokowi Terkait Vaksinasi* Covid-19 *untuk Anak*" from Kompas TV channel.
- "*Stok Terbatas, Capaian Vaksinasi* Covid-19 *Timpang*" from CNN Indonesia channel.
- "*Vaksinasi* Covid-19 *Sasar Siswa Jelang PTM di Beberapa Daerah Indonesia*" from Kompas TV channel.
- "*Gerak Cepat Vaksinasi* Coron*a di Indonesia*" from Kompas TV channel.

The abundance of data that is automatically produced by social media, such as YouTube, often confuses users in understanding information. Therefore, data preprocessing can be an important factor in significantly impacting data quality. Before the data can fit a model, it needs to be transformed to a format the model can understand; the preprocessing stages in sentiment analysis are generally the same. Before review data is converted into numerical vectors, the text preprocessing sets are carried out, which include case folding, tokenization, and removal of stop words. The steps of data that will be processed as explained in the diagram below.



Source: The Steps Designed by the Authors, 2021

Figure 3. The Steps of Preprocessing Data

Data preprocessing is a catch-all term used to undertake to get the documents ready to be analyzed, and cleansing refers to standardizing the text, removing characters that aren't relevant, repairing the errors in spelling, and deleting trivial comments (Pokharel & Bhatta, 2021). The data generated from Coberry will go through the data cleaning stage. Data cleaning is the stage to identify and resolve corrupt, inaccurate, and irrelevant data, to obtain accurate and accountable analysis results, such as comments that only use emojis or symbols and out-of-context advertising comments. In this phase, duplicate row identification is also made to reduce redundancy. The comments in Indonesian will be translated into English. The comment data must have a variety of font forms. The folding case will convert all characters in the document to all uppercase or lowercase to speed up comparisons during the indexing process. The data processing involves the separation of the document created in lower case. Computers will often treat capitalized versions of words as being different to their lowercase, so this can cause problems during analysis. Then, turning all text into lowercase can solve these problems.

Tokenizing aims to break up the raw text into smaller chunks. Tokenization breaks raw text into words called tokens. This token helps in understanding the context or developing a model for NLP. Tokenization helps interpret the text's meaning by analyzing the order of the words. The following process is stopped word removal, which is the cessation of deleting common words that dominate, such as "*yang*", "*di*", "*dengan*" which do not provide much insight into the document. The texts are filtered before being analyzed.

The last process is stemming, a core natural language processing technique for efficient and effective information retrieval (Frakes & Baeza-Yates, 1992).

Table 2. The Example of Preprocessing Step

Description	Input	Output
Data Cleansing	"*Mantap* 👍 *Terimakasih atas penjelasannya* 🙏 *Sekarang yang paling penting adalah Imun/Antibodi Kita harus kuat* 🙏 "	"*Mantap Terimakasih atas penjelasannya Sekarang yang paling penting adalah ImunAntibodi Kita harus kuat* "
Case Folding	"*Mantap Terimakasih atas penjelasannya Sekarang yang paling penting adalah Imun/Antibodi Kita harus kuat* "	"*mantap terimakasih atas penjelasannya sekarang yang paling penting adalah imunantibodi kita harus kuat* "
Tokenizing	"*mantap terimakasih atas penjelasannya sekarang yang paling penting adalah*	[*'mantap', 'terimakasih', 'atas', 'penjelasannya', 'sekarang', 'yang', 'paling', 'penting',*

Description	Input	Output
	imun/antibodi kita harus kuat “	*'adalah', 'imunantibodi', 'kita', 'harus', 'kuat'*]
Removal Stop Words	[*'mantap', 'terimakasih', 'atas', 'penjelasannya', 'sekarang', 'yang', 'paling', 'penting', 'adalah', 'imun/antibodi', 'kita', 'harus', 'kuat'*]	[*'mantap', 'terimakasih', , 'penjelasannya', 'sekarang', 'paling', 'penting', 'imunantibodi', 'kuat'*]
Stemming	[*'mantap', 'terimakasih', , 'penjelasannya', 'sekarang', 'paling', 'penting', 'imun/antibodi', 'kuat'*]	[*'mantap', 'terimakasih', , 'jelas', 'sekarang', 'paling', 'penting', 'imunantibodi', 'kuat'*]

Source: Result of analyzing data by authors, 2021

The next step is to classify text by labelling a text by the representative words for each class, and unlabeled document sets are then used to build a Classifier. For this research, we attempt to label an unlabeled dataset using a python library called Text blob, which is made for Natural Language Processing data and preprocessing (Heronius, 2019). The text blob itself is not working in Indonesian, so the data will be translated into English before the text blob is used. Furthermore, data processing is needed to extract numerical features from text content (Duong & Nguyen-Thi, 2021). Next, the dataset is split for training and testing to check how well the model has performed, using RapidMiner software for the Naive Bayes Classifier Model.

In the modelling process, the problem often faced is the imbalance of data classification. This significantly affects the model's accuracy, which tends to follow the class with the highest number of classes (Tallo & Musdholifah, 2018). Therefore, we need a method to balance the data. The method used is the Synthetic Minority Oversampling Technique (SMOTE) which is one of the effective and easy methods to overcome data imbalance. However, the amount of data will be reduced.

The distribution of the Naive Bayes model classification provides an overview of public sentiment about vaccination. Through this, we can evaluate the policies that have been implemented. Further evaluation can be done by analyzing word frequency. This research can also be compared with other similar studies with different data collection processes to see the most recent public response and the policy actions to be taken.

D. RESULT AND DISCUSSION

The process of Naive Bayes in the RapidMiner is first to shuffle the dataset, split it into a training and testing dataset, and then develop the model with three classes (positive, negative, and neutral) (Kurniawan & Kriestanto, 2016). Text classification techniques can be grouped into supervised, semi-supervised, and unsupervised learning (or clustering). All the methods are related but significantly different from all these approaches (Riyanah & Fatmawati, 2021). In supervised learning, a set of labelled training data for every class is used by a learning algorithm to build a classifier (Feldman & Sanger, 2007).

This paper collected YouTube reviews from the eight videos above through web scraping tools called Coberry. It obtained 6,170 comment rows in total from all videos. All data was displayed in *Bahasa* that needed some treatments to conform to the method. Then the data is

exported to CSV for further identification. The YouTube reviews or any online data that is scraped are likely to contain "noise" like repetitive or extraneous text, emojis, or other text and symbols that are irrelevant and can affect our analysis. Identification is needed to cleanse the datasets. Columns of data that are not used in sentiment analysis will be omitted. Rows of data that do not provide meaningful information and duplicate rows were also deleted. After the cleaning, 5,510 comment rows remained.

The clean data had to be labelled into three types of sentiment, that is positive, negative, and neutral. The labelling used a text blob method on Python to grab the sentiment. An example of the result is shown below which the text blob told two values: polarity and subjectivity.

Table 3. Result of Text Blob Method for Polarity and Subjectivity

Original Content (In *Bahasa*)	Translated Content (In English)	Sentiment (Polarity, Subjectivity)
otoriter pemaksaan pengancaman	Authoritarian coercion of threat	(0.000, 0.000)
dari sisi ham menolak vaksin itu hak rakyat	In terms of human rights, reject the vaccine th...	(0.000, 0.100)
jualan vaksin rakyat jadi korbangila pakai san...	Selling People's Vaccines So *Korbangila* Use San...	(0.000, 0.000)
pertama vaksin astrazeneca demam tinggi 2 hari...	First AstraZeneca vaccine high fever 2 days sh...	(-0.047, 0.443)
daftar vaksin online habis terus geblek	online vaccine list continues to *geblek*	(0.000, 0.000)
pertanyaan dan jawaban tida menyambung apa paj...	The question and answer did not connect what t...	(0.259, 0.509)

Source: Result of Analyzing Data by Authors, 2021

The output of Text Blob's polarity will be in the form of floating points in the range of [-1.0,1.0]. When the polarity value is less than 0, then the sentiment of the statement is negative. If the polarity value is greater than 0, then it is positive. If it is equal to 0, then the statement is neutral. This happens when subjectivity and objectivity fall in the range [0.0,1.0] when 0.0 represents a very objective sentence, and 1.0 is subjective. The next step is to define a function that calculates subjectivity and polarity and give it a label.

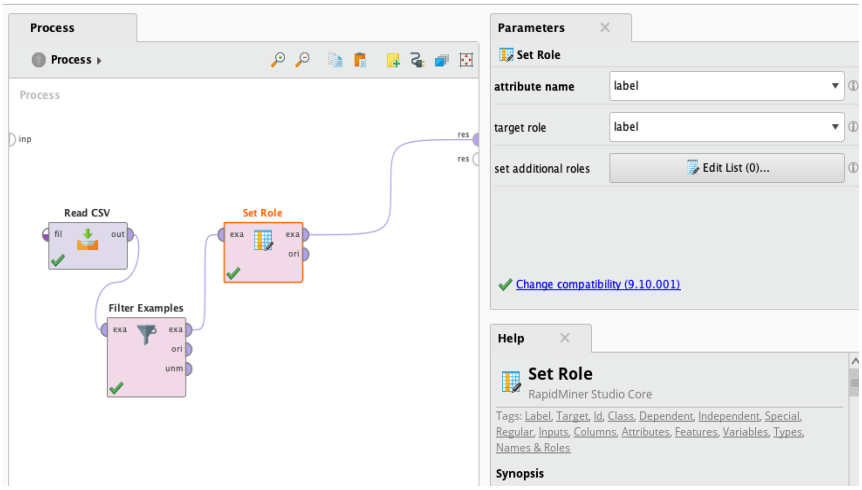
Table 4. Result of Text blob Method for Sentiment Labeling

Original Content (In *Bahasa*)	Translated Content (In English)	Sentiment Label
otoriter pemaksaan pengancaman	Authoritarian coercion of threat	Neutral
dari sisi ham menolak vaksin itu hak rakyat	In terms of human rights, reject the vaccine the...	Neutral
jualan vaksin rakyat jadi	Selling People's Vaccines So	Neutral

Original Content (In *Bahasa*)	Translated Content (In English)	Sentiment Label
korbangila pakai san...	*Korbangila* Use San...	
pertama vaksin astrazeneca demam tinggi 2 hari...	first AstraZeneca vaccine high fever 2 days sh...	Negative
daftar vaksin online habis terus geblek	online vaccine list continues to *geblek*	Neutral
pertanyaan dan jawaban tida menyambung apa paj...	The question and answer did not connect what t...	Positive

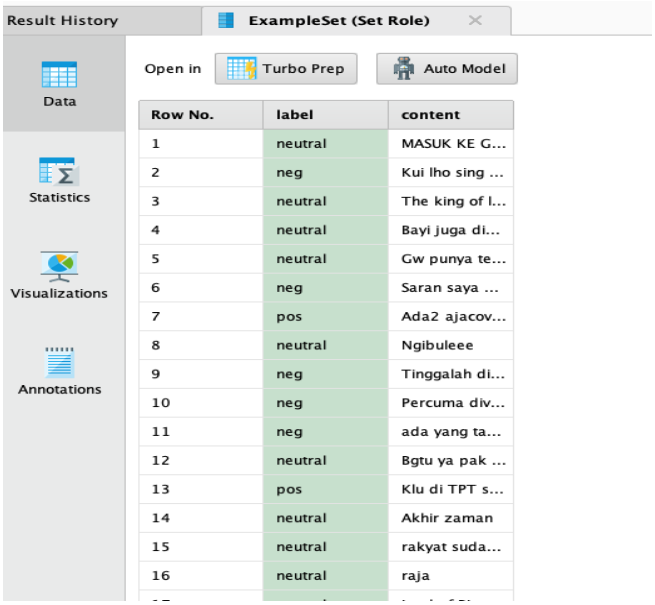
Source: Result of analyzing data by authors, 2021

The dataset on the CSV file was imported to RapidMiner. The dataset consists of two attributes, content and label, which were previously labelled. The attribute's label contains three values, which are positive, negative, and neutral. After that, the dataset will be filtered again to ensure no missing values. Then, the attributes of the regular role and the label role attribute will be selected using the set role operator. In this case, content a regular role and label a label, the predicted class for the modelling.



Source: Process of the Modelling on RapidMiner, 2021

Figure 4. Import Data Process on RapidMiner

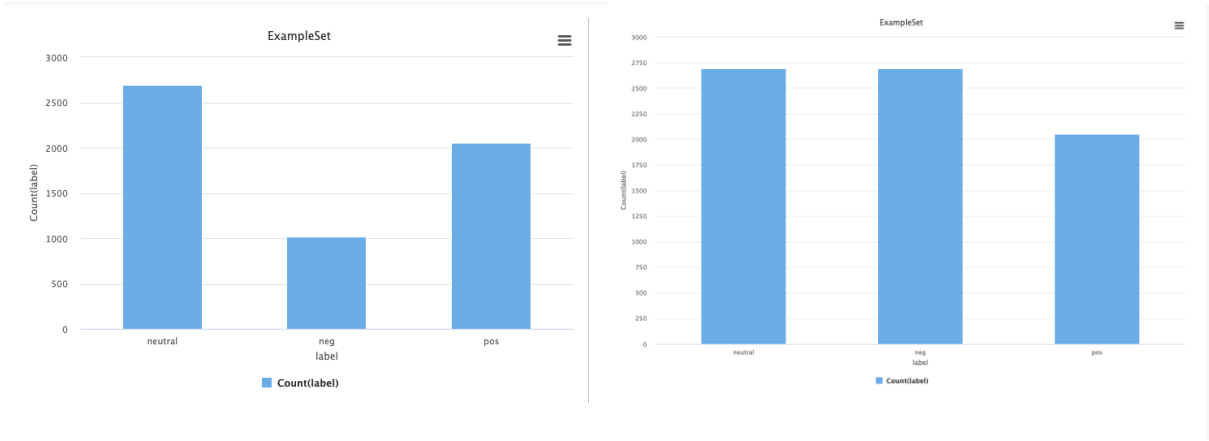


Row No.	label	content
1	neutral	MASUK KE G...
2	neg	Kui lho sing ...
3	neutral	The king of l...
4	neutral	Bayi juga di...
5	neutral	Gw punya te...
6	neg	Saran saya ...
7	pos	Ada2 ajacov...
8	neutral	Ngibuleee
9	neg	Tinggalah di...
10	neg	Percuma div...
11	neg	ada yang ta...
12	neutral	Bgtu ya pak ...
13	pos	Klu di TPT s...
14	neutral	Akhir zaman
15	neutral	rakyat suda...
16	neutral	raja

Source: Result of the modelling on RapidMiner, 2021

Figure 5. Import Data Process on RapidMiner

Before splitting the data into training and test, check the performance of the dataset model on test data, and based on this, we can see that the dataset seems to be unbalanced. In contrast, sentiment neutral looks more dominant than others. Unstable data classification is a crucial problem in machine learning and data mining. Data imbalances poorly impact classification results where minority classes are often misclassified as majority classes. In this study, we used the Synthetic Minority Over-Sampling Technique (SMOTE) to solve the class imbalance problem on the dataset. SMOTE focused on generating synthetic data between each sample of the minority class and its nearest neighbours. That is, for each one of the samples of the minority class, its "k" nearest neighbours are located (by default k = 5). New synthetic data are generated between the pairs of points generated by the sample and each of its neighbours. The figure below shows the dataset before and after SMOTE was applied.



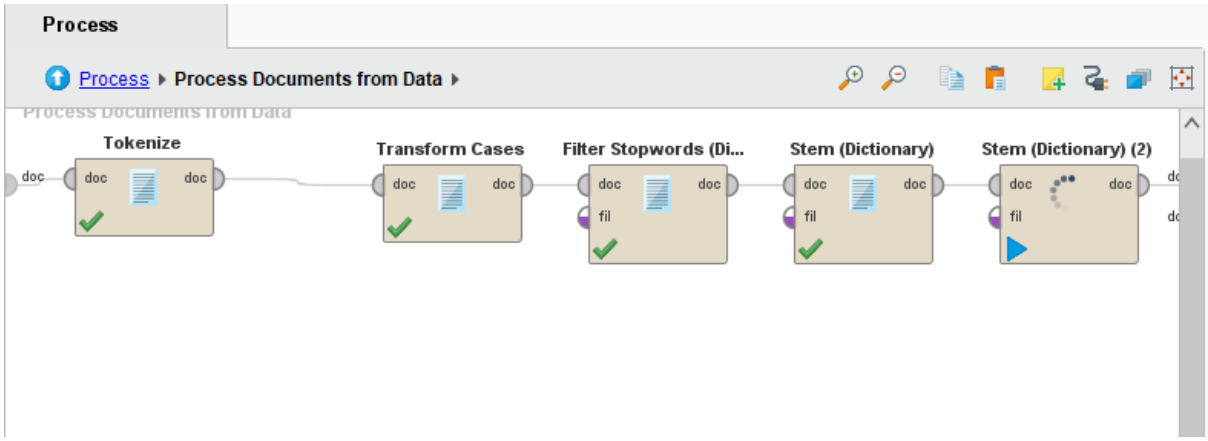
(a) (b)

Source: Result of the Modeling on RapidMiner, 2021

Figure 6. Comparison of Unbalanced Data (a) and Balanced Data Using SMOTE (b)

Unlike in the context of using text for a category as a nominal, it is important to bring the text into RapidMiner Studio as nominal so that it can do any further text processing, so we use

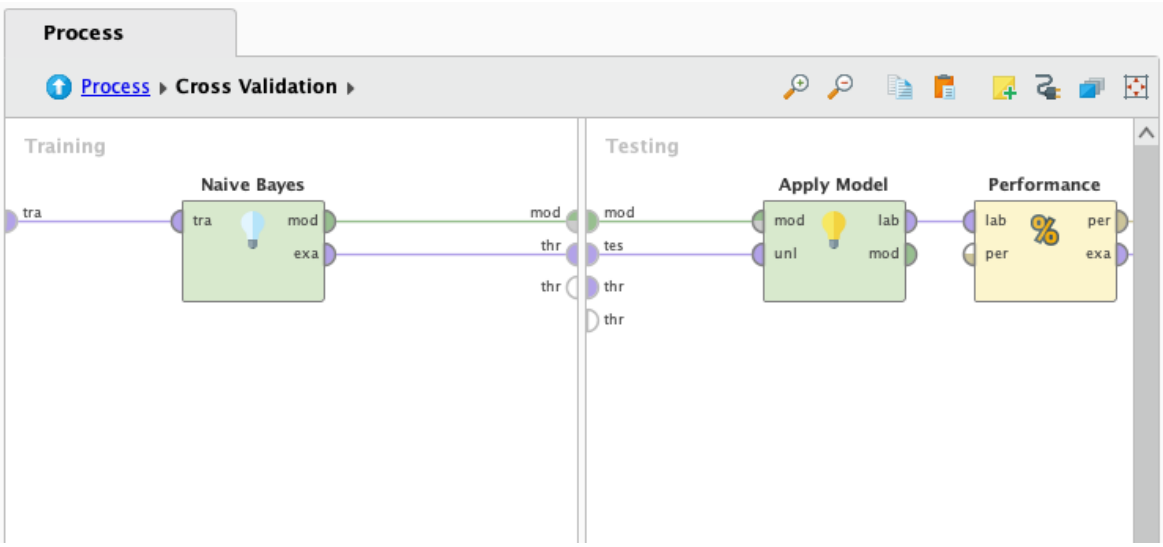
nominal to text. The nominal to text operator converts all nominal attributes to string attributes. Each nominal value is simply used as a string value of the new attribute. The new value will also be missing if the value is missing in the nominal attribute. Then, the preprocessing data can be started after converting nominal to text. In this phase, the data will have at least five steps: case folding, tokenization, filter tokens, removal of the stop words, and stem. Figure 7 shows the contents of the operator "Process Documents", used by operators "transform cases", "tokenize", "stop word filters", and "stem" in Indonesian. The preprocessing data is just like the picture below.



Source: Result of the Modelling on RapidMiner, 2021

Figure 7. The Preprocessing Data Phase on RapidMiner

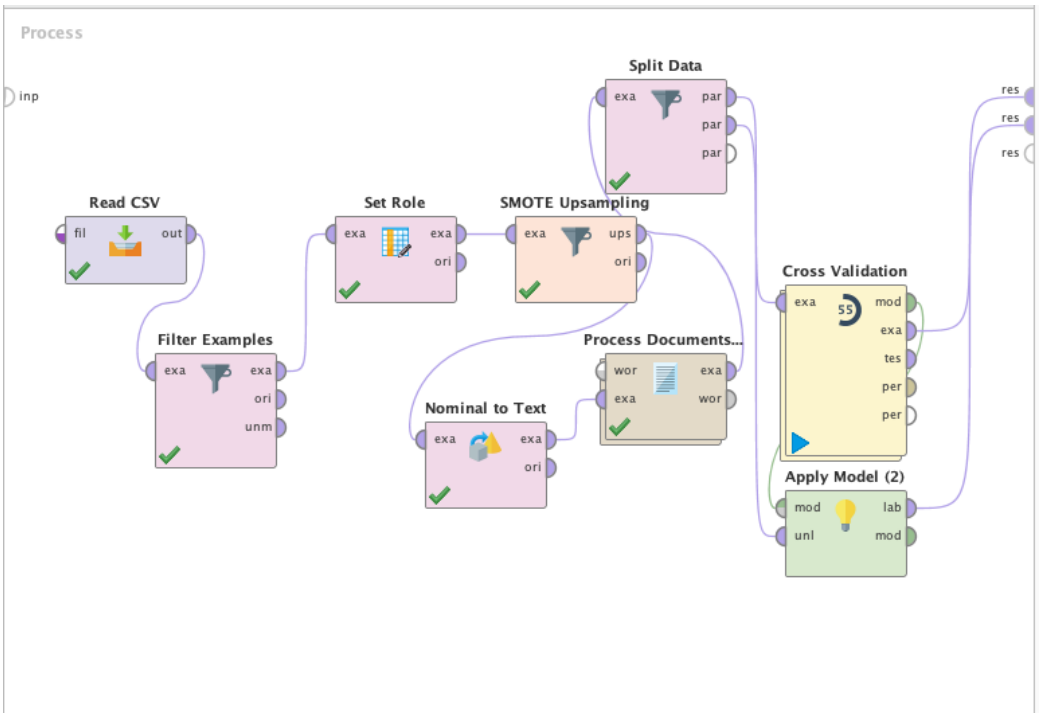
Succeeding preprocessing data, further to this stage, is providing training and implementing various data mining algorithms. Figure 8 shows the contents of the "Cross Validation" operator in the RapidMiner application. This step included the Naive Bayes operator, the Apply Model operator, and the Performance operator. Those connections will show the model's result, including the model's accuracy, after the process is executed. The data was split into two partitions, with 80% data for model building and validation and the rest for testing.



Source: Result of the Modeling on RapidMiner, 2021

Figure 8. The Process of Training a Naive Bayes Model

Add another Apply Model operator outside the Split Validation operator and deliver the model. When the above process is run, the confusion matrix and ROC curve for the validation sample should be generated. After the sentiment analysis classification process is complete, one more step is needed to determine the quality of the process that has been carried out, namely evaluating the results. Calculations performance will be tested using accuracy, precision, and recall parameters. The whole process of a RapidMiner can be described as follows.



Source: Result of the Modelling on RapidMiner,2021

Figure 9. The Whole Process of Sentiment Analysis on Rapidminer

accuracy: 56.32% +/- 1.47% (micro average: 56.32%)

	true Negative	true Neutral	true Positive	class precision
pred. Negative	1756	926	631	53.00%
pred. Neutral	111	830	345	64.54%
pred. Positive	171	282	594	56.73%
class recall	86.16%	40.73%	37.83%	

Source: Result of the Modelling on RapidMiner, 2021

Figure 10. The Accuracy of Naive Bayes Classifier Model Result

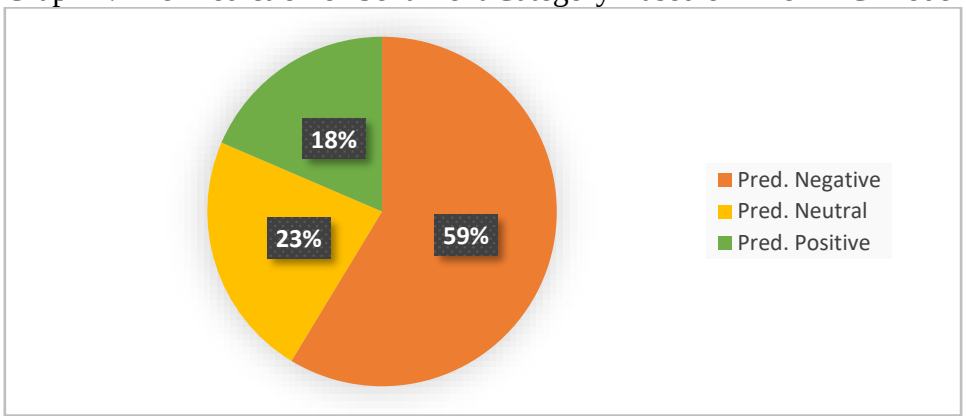
From the image above, we can conclude that the overall accuracy was 56.32%. Accuracy was calculated by taking the percentage of correct predictions over the total number of examples. Correct prediction means examples where the value of the prediction attribute was equal to the value of the label attribute. The weighted mean recall was calculated by taking every class's average recall. As we have seen in the last row of the resultant matrix in the Results Workspace, class recall for the neutral class was 40.73%, class recall for the negative class was 86.16% and class for the positive class was 37.83%. Thus, the weighted mean recall was calculated by taking the average of these class recall values, 54.91%.

Meanwhile, the weighted mean precision is calculated by taking the average precision of every class. In the last column of the resultant matrix in the Results Workspace, class precision

for neutral prediction was 64.54%. Class precision for negative prediction was 53.00%, and class precision for positive prediction was 56.73%. Thus, weighted mean precision was calculated by taking the average of these class precision values, 57.09%. Overall, the obtained model has an accuracy value of more than half.

The precision for the neutral prediction of correctly predicted neutral predicted out of all predicted 64.54%, so it is approximately 64.54% of the neutral that our predictor classifies as neutral. On the other hand, the recall for neutral is the number correctly predicted as neutral out of the number of neutrals which was 40.73%. This means that our classifier classified 2/5 of the neutral predicted as neutral. Similarly, it can calculate the precision and recall for the other classes, which are negative and positive. The image explains more about the performance of model naive bayes. The prediction of the sentiment score is shown in the graph 1.

Graph 1. The Prediction of Sentiment Category Based on The NBC Model

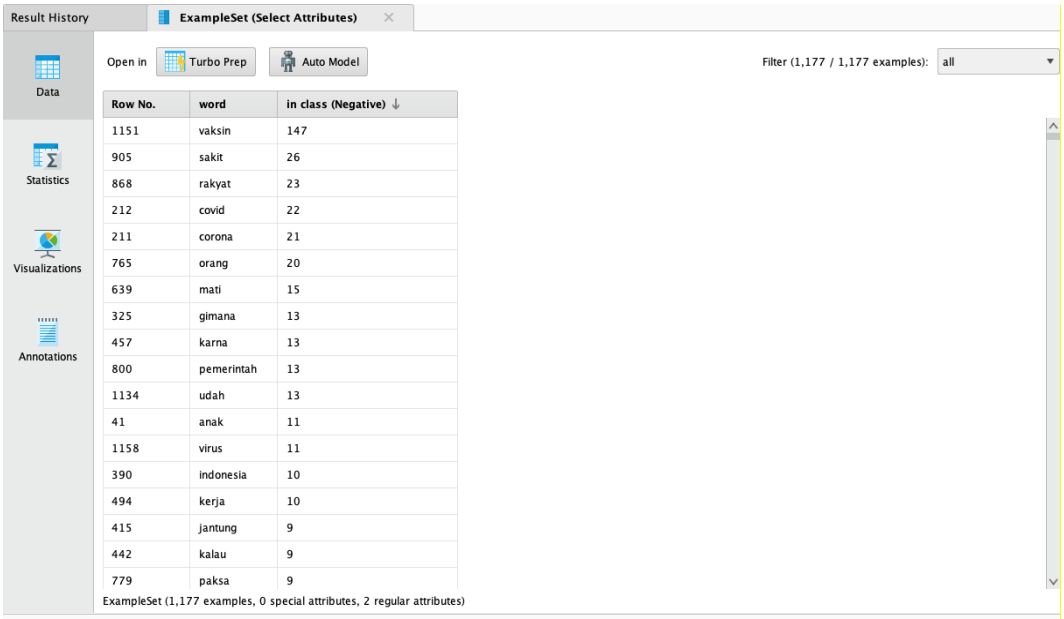


Source: Result of the Modelling on RapidMiner, 2021

The data described that neutral and negative sentiments dominated the comments. These results differed from those of three journals in similar studies discussed previously. This reveals that differences in social media sources greatly affect the sentiment results. In addition, in this study, public comments were taken based on news that raised the issue of a particular vaccination program so that the community's impulsive response is more visible to the planned decisions taken by the government.

In addition to the reasons for spontaneity, differences in sentiment conclusions also occur because this study collected commentary data from nine different videos in eight months since the vaccination program was first implemented. Meanwhile, in the three journals being compared, commentary data were collected only in fewer than two weeks at any given time, whether vaccination was being discussed or not. So that the data contained in this study better describe the development of the community's response during the vaccination program, where each video selected has its problems.

Further analysis was conducted to uncover why people had more negative than positive things about the case. The reason will be applied through the frequency of negative sentiment words.



Source: Result of the Modelling on RapidMiner, 2021

Figure 11. The Frequency of Words for Negative Sentiments

Based on the word frequency analysis in the collected data, it was found that several keywords made the response fall into the category of negative sentiment. The keywords referred to include "*sakit*", "*mati*", "*gimana*", "*karna*", "*pemerintah*", "*kerja*", and "*jantung*". If we look more closely, some of these keywords reflect doubts about the vaccination program, which remains an obstacle for the public. The keywords "*sakit*", "*mati*", "*karna*", "*jantung*" reflected that people have doubts about vaccines. On the contrary, they believe the vaccine can make them sick or even die. The keyword "*kerja*" implied people's fear of losing their jobs and the difficulty of finding new employment. It was known that the number of open unemployment in Indonesia increased sharply during the pandemic. And two other keywords, namely "*gimana*", "*pemerintah*", were an expectation from the community that the government was expected to act quickly and have a strategic plan to stop the pandemic and restore the country's economic stability.

Based on the analysis above, the government needs to transform public policy by utilizing mass media, especially digital-based ones, to voice the positive impact of the COVID-19 vaccination program. The government can also use the power of influencers to convey the vaccine program's importance in restoring economic activity in Indonesia. The government also needs to continue monitoring the community's response to this program to fulfil their aspirations and determine policies that are by the community's wishes. The selection of data carried out by this research can be the basis for gathering people's aspirations by utilizing digital media to raise certain issues and seeing the public's response to the issues raised. Furthermore, the collected data can be further developed with more advanced analytical methods for sharper results.

E. CONCLUSION

Public policy is one of the manifestations of the results of efforts to understand and interpret what the government should do about a problem for the benefit of the public. Sentiment analysis is one way to determine the public's response to a problem as a reference for the government to formulate public policies to answer the problem. Collecting public response data is a crucial key in the process of sentiment analysis on an issue. Different sources

of public response data and timing can provide different sentiment conclusions, providing different inputs for public policy formulation.

This study conducted a sentiment analysis based on data collected from nine videos with different issues on YouTube. It resulted in the dominance of negative sentiment from the public towards the implementation of the COVID-19 vaccination program. The negative sentiment was triggered by public doubts about the side effects of vaccines, the limited number of jobs in a pandemic, and the government's lack of follow-up to deal with the pandemic and restore the country's economy. Several limitations in the data processing process can focus on further improvements to produce more accurate and sharp input in formulating policies related to handling COVID-19 by the government. The main limitations encountered in the data cleansing process are slang, regional languages, and some emojis and symbols. The use of a more suitable model also needs to be improved in future research.

Acknowledgement

The manuscript has been presented at the 3rd ICoGPASS 2021.

REFERENCES

Anggraini, N., Harahap, E. S. N., & Kurniawan, T. B. (2021). Text Mining - Analisis Teks Terkait Isu Vaksinasi COVID-19. *Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi*, *23*(2), 141–153.

Asmiati, N., & Fatmawati. (2020). Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Pengaruh Negatif *Game Online* bagi Remaja Milenial. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *2*(3), 141–149. https://doi.org/10.35746/jtim.v2i3.102

Duong, H.-T., & Nguyen-Thi, T.-A. (2021). A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis. *Computational Social Networks*, *8*(1), 1. https://doi.org/10.1186/s40649-020-00080-x

Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer dalam Algoritma Klasifikasi. *Teknika*, *8*(1), 11–16. https://doi.org/10.34148/teknika.v8i1.95

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Feng, G., Guo, J., Jing, B.-Y., & Sun, T. (2015). Feature Subset Selection Using Naive Bayes for Text Classification. *Pattern Recognition Letters*, *65*, 109–115. https://doi.org/10.1016/j.patrec.2015.07.028

Firmasyah, Z., & Puspitasari, N. F. (2021). Analisis Sentimen Masyarakat terhadap Vaksinasi COVID-19 Berdasarkan Opini pada Twitter Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Informatika*, *14*(2), 171–178.

Frakes, W., & Baeza-Yates, R. (1992). *Information Retrieval, Data Structures and Algorithms*. Pearson.

Hagemann, M. (2020). *How to Export Youtube Comments*. Https://Coberry.Com/Blog/How-to-Export-Youtube-Video-Comments.

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining*. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Hardy, F. R. (2020). Herd Immunity Tantangan New Normal Era Pandemi Covid 19. *JURNAL ILMIAH KESEHATAN MASYARAKAT : Media Komunikasi Komunitas Kesehatan Masyarakat*, *12*(2), 55. https://doi.org/10.52022/jikm.v12i2.70

Hermansyah, R., & Sarno, R. (2020). Sentiment Analysis about Product and Service Evaluation of PT Telekomunikasi Indonesia Tbk from Tweets Using TextBlob, Naive Bayes & K-NN Method. *2020 International Seminar on Application for Technology of Information*

and Communication (ISemantic), 511–516. https://doi.org/10.1109/iSemantic50169.2020.9234238

Hermanto, D. T., Ziaurrahman, M., Bianto, M. A., & Setyanto, A. (2018). *Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier*. 12037. https://doi.org/10.1088/1742-6596/1140/1/012037

Hernikawati, D. (2021). Kecenderungan Tanggapan Masyarakat terhadap Vaksin Sinovac Berdasarkan Lexicon Based Sentiment Analysis. *Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi*, *23*(1), 21–31.

Heronius, A. (2019). *Twitter Sentiment Analysis Bahasa Indonesia dengan TextBlob*. Retreived from https://Medium.Com/@albertusheronius/Twitter-Sentiment-Analysis-Bahasa-Indonesia-Dengan-Textblob-F34e1ffdcdaa.

Hofmann, M., & Klinkenberg, R. (2014). *RapidMiner : Data Mining Use Cases and Business Analytics Applications*. Chapman and Hall/CRC.

Ikasari, D., & Widiastuti, W. (2021). Sentiment Analysis Review Novel "Goodreads" Berbahasa Indonesia Menggunakan Naïve Bayes Classifier. *Seminar Nasional Riset Dan Inovasi Teknologi (SEMNAS RISTEK)*, *5*, 760–765.

Kadafi, A. R. (2018). Perbandingan Algoritma Klasifikasi untuk Penjurusan Siswa SMA. *Jurnal ELTIKOM*, *2*(2), 67–77. https://doi.org/10.31961/eltikom.v2i2.86

Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*.

Kolluri, J., Razia, S., & Nayak, S. R. (2020). Text Classification Using Machine Learning and Deep Learning Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3618895

Kurniawan, D. A., & Kriestanto, D. (2016). Penerapan Naive Bayes untuk Prediksi Kelayakan Kredit. *JIKO (Jurnal Informatika Dan Komputer)*, *1*(1), 19–23. https://doi.org/10.26798/jiko.2016.v1i1.10

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. Proceedings of International Conference on World Wide Web (WWW-2005).

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The Impact of Class Imbalance in Classification Performance Metrics based on the Binary Confusion Matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/J.PATCOG.2019.02.023

Madyatmadja, E. D., Sembiring, D. J. M., Angin, S. M. B. P., Ferdy, D., & Andry, J. F. (2021). Big Data in Educational Institutions using Rapid Miner to Predict Learning Effectiveness. *Journal of Computer Science*, *17*(4), 403–413. https://doi.org/10.3844/jcssp.2021.403.413

Massy, W. F. (1964). *Confusion Matrices as Measures of Audience Similarity*. Stanford University.

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2019). Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube. *Social Science Computer Review*, *37*(4), 510–528. https://doi.org/10.1177/0894439318779336

Mustakim, M., Hidayat, A., Efendi, Z., Aszani, A., Novita, R., & Lestari, E. T. (2018). Algorithm Comparison of Naive Bayes Classifier and Probabilistic Neural Network for Water Area Classification of A Fishing Vessel in Indonesia. *Journal of Theoretical and Applied Information Technology*, *96*(13), 4114–4125.

Nofriani, N. (2018). Analysis On Internet Pattern of Youtube Browsing in Indonesia Using Web Crawling and Unsupervised Learning (Analisis Pola Minat Tayangan Youtube DI Indonesia dengan Web Crawling dan Supervised Learning). *JURNAL IPTEKKOM : Jurnal Ilmu Pengetahuan & Teknologi Informasi*, *20*(2), 93. https://doi.org/10.33164/iptekkom.20.2.2018.93-106

Permadi, V. A. (2020). Analisis Sentimen Menggunakan Algoritma Naive Bayes terhadap Review Restoran di Singapura. *Jurnal Buana Informatika*, *11*(2), 140. https://doi.org/10.24002/jbi.v11i2.3769

Pokharel, R., & Bhatta, D. (2021). Classifying YouTube Comments Based on Sentiment and Type of Sentence. *ArXiv Publication*.

Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Indonesian of Health Information Management Journal*, *8*(2), 100–109.

Riyanah, N., & Fatmawati, F. (2021). Penerapan Algoritma Naive Bayes untuk Klasifikasi Penerima Bantuan Surat Keterangan Tidak Mampu. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *2*(4), 206–213. https://doi.org/10.35746/jtim.v2i4.117

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1), 12. https://doi.org/10.1007/s41133-020-00032-0

Tallo, T. E., & Musdholifah, A. (2018). *Modifikasi Metode Synthetic Minority Oversampling Technique (SMOTE) Menggunakan Algoritma Genetika untuk Menangani Masalah Imbalanced Dataset*.

Tsangaratos, P., & Ilia, I. (2016). Comparison of a Logistic Regression and Naïve Bayes Classifier in Landslide Susceptibility Assessments: The Influence of Models Complexity and Training Dataset Size. *CATENA*, *145*, 164–179. https://doi.org/10.1016/J.CATENA.2016.06.004

Viny Christanti, M., Walda, & Sutrisno, T. (2020). Comments Scraping Application For Review Youtube Content. *IOP Conference Series: Materials Science and Engineering*, *852*(1), 012167. https://doi.org/10.1088/1757-899X/852/1/012167

Yulita, W., Nugroho, E. D., & Algifari, M. H. (2021). Analisis Sentimen terhadap Opini Masyarakat tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Data Mining Dan Sistem Informasi*, *2*(2), 1–9.