COMPILER
Vol. 11, No. 2, November 2022, pp. 45~50
ISSN: 2252-2403 (Print), ISSN: 2549-2403 (Online)
Accredited Fourth Grade by Kemenristekdikti, Decree No:200/M/KPT/2020
DOI: 10.28989/compiler.v11i2.1356

Sentiment Analysis on the Centralized Isolation Policy for Covid-19 Response in Bali Province

I Made Ari Santosa[1,*], Luh Putu Ayu Prapitasari[2], I Putu Ramayasa[3]
[1,2]Computer Systems Study Program, Institute of Technology and Business STIKOM Bali
[3]Information Systems Study Program, Institute of Technology and Business STIKOM Bali

Article Info

Article history:

Received October 10, 2022
Accepted November 25, 2022
Published December 31, 2022

Keywords:

Sentiment Analysis
Centralized Isolation
Covid-19
Bali Province

ABSTRACT

Coronavirus disease (Covid-19) is an infectious virus caused by SARS-CoV-2 virus, which has caused a global pandemic. The global pandemic is declared by the World Health Organization (WHO) on March 11th, 2020. Until late of April 2022, Covid-19 has caused the death of the total of more than 6 million people, and in Indonesia it has caused more than 150 thousand of death. The pandemic does not only impact health area, but also economy, politics, cultures, etc. In order to end the pandemic, WHO has set out standard policies to be implemented, including by Indonesian government. Besides, each local goverment in Indonesia has also their own policies and regulations, including the government of Bali Province, for example the Community Activities Restrictions Enforcement or CARE (Indonesian: *Pemberlakuan Pembatasan Kegiatan Masyarakat*, commonly referred to as the *PPKM*), improvement in health care facilities and services, and isolation policy for the people whom are infected by SARS-CoV-2 virus. To prevent the increase in death, Bali Province's COVID-19 Handling Task Force along with the district and city governments have prepared a centralized isolation policy and facilities for those whom are found to be Covid-19 positive without having any symptoms and for those with mild symtomps. This research has been conducted in order to get the public sentiment about this centralized isolation policy in Bali Province by using Tweet data from Twitter social media, based on Naive Bayes classification. From the experiments, the results of data Tweet classifications have accuracy more than 90%, which have shown that the public opinion is neutral. The results also mean that there is no contradition found in the application of centralized isolation policy in the province of Bali-Indonesia.

Corresponding Author:

I Made Ari Santosa,
Program Studi Sistem Komputer
ITB STIKOM Bali
Jl. Raya Puputan No. 86 Denpasar, Bali-Indonesia, 80116
Email: arisantosa@stikom-bali.ac.id

1. INTRODUCTION

The disease known as Covid-19, which is caused by a viral infection that attacks the respiratory system called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1], has caused a world pandemic since it was officially announced by the World Health Organization (WHO). According to WHO, in early 2020 and until now, the Covid-19 pandemic is still not under control. Various handling measures have been set by WHO as a world health institution which were then implemented by all countries in the world, including the Indonesian government, coupled with various local policies with the same goal, namely to block the spread of the virus, but it seems the world is still struggling to end the pandemic.

SARS-CoV-2 is a type of virus that is very easy to mutate and continue to produce new virus variants, such as several well-known ones, namely: Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Theta, Iota, Kappa and Omicron variants. Each variant infects with almost the same general symptoms, for example those identified by WHO are fever, cough, feeling tired, and loss of the sense of taste and smell [2].

SARS-CoV-2 can attack anyone regardless of race, age or gender and others. The level of danger caused by this viral infection depends on the immune of the infected body, age and conditions with pre-existing medical diseases (comorbidities) also greatly determine the recovery rate of infected people. SARS-CoV-2 infection can cause mild, moderate or severe symptoms [3]. For moderate and severe symptoms, it can even cause lung infections, up to death. The impact of the Covid-19 pandemic is not only in the health sector, but also in the economic, political, cultural, and other fields. In the context of handling the Covid-19 pandemic, world health institutions, in this case WHO, have a very important role, for example providing accurate information, advocating for health services, including setting health standards and guidelines for all countries in the world, including Indonesia. In addition, each country and regional leaders also have their own policies. The Bali Provincial Government itself has implemented various efforts to deal with the Covid-19 pandemic, including implementing restrictions on community activities (PPKM), improving health facilities and services, as well as isolation policies for Covid-19 survivors. In addition to self-isolation, a centralized isolation policy is also implemented in places that have been determined by the government, especially for Covid-19 survivors without symptoms (OTG) and mild symptoms (GR).

However, behind the centralized isolation policy implemented by the government, there is controversy in the community from parties who have opposing opinions regarding the centralized isolation policy. Therefore, to obtain information regarding how public sentiment towards the isolation policy is centralized in the province of Bali, which will later be able to assist policy makers in taking appropriate steps to avoid the debate caused by the controversy over the policy, in this study we propose a method sentiment analysis on social media, especially Twitter with the Tex Mining approach.

Text mining is a process for extracting information from a very large collection of text. Text mining not only explores but also identifies interesting patterns and relationships in the text data [4]. Text mining itself is a multi-disciplinary field, which is a combination of data mining, natural language processing, machine learning, and information retrieval. Meanwhile, sentiment analysis, also known as opinion mining, is an inseparable part of natural language processing itself [5]. Sentiment analysis is very necessary to know. The main reason is that the results will be very helpful in decision making, whether carried out by individuals, institutions, companies, or other organizations [6]. Until now, many studies on sentiment analysis have been carried out, including publications [7, 8, 9, 10].

According to survey results from [11], the Naive Bayes algorithm is the simplest and most commonly used classification algorithm in sentiment analysis research. Naive Bayes begins with the Bayes theorem, which is a logic-based approach that gets the probability of a hypothesis based on the latest evidence [12]. There are several studies that report classification results with Naive Bayes which provide an accuracy rate of more than 70%, as in [13, 14, 15]. Accordingly, in this study, sentiment analysis will be processed using the Naive Bayes algorithm.

2. RESEARCH METHOD

In accordance with the flowchart presented in Figure 1, it can be seen that, in this study, to conduct sentiment analysis, the first stage to be carried out is web scrapping, which is to collect text related to the centralized isolation policy in handling Covid-19 implemented by the government of Bali Province. After all the data has been collected, the next step is to carry out the pre-processing stage, including labeling, so that the data is ready for further processing. The next step is to prepare training data and testing data to be included in the classification stage using the Naive Bayes method. After the classification stage is completed, it will be continued with the evaluation stage. From this evaluation stage, the results will be seen and analyzed, which will then become the output of this research.

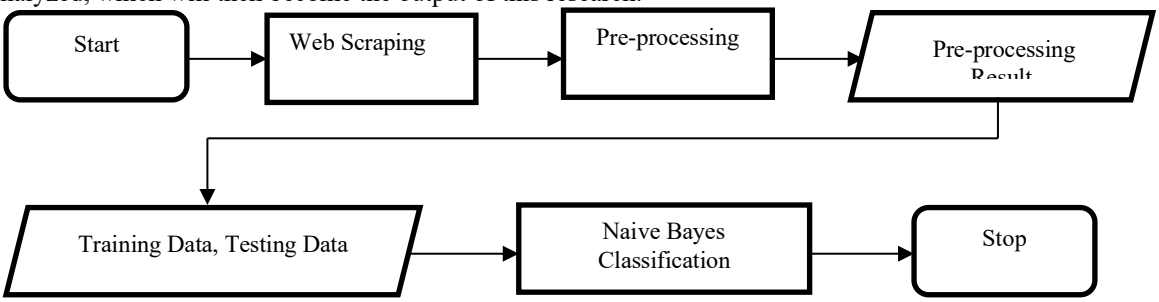

Figure 1. Research Flow

In general, it can be explained that there are 3 main stages in this research that raises the topic of Sentiment Analysis, namely: 1) Web Scrapping, 2) Pre Processing, with data labeling previously done, and 3) Classification using Naive Bayes. Each of these stages can be explained as follows :

Web Scraping

Web scraping or web harvesting, also known as web data extraction, is an activity to extract data from web pages. The web scraping process in this study uses the SNScrapper library. This library is a scrapper tool used for social networking services (SNS). The data that can be extracted are in the form of: user profile data, hashtags, and also relevant posts according to search keys.

Labeling

After the data source is obtained, the next step is labeling. From this stage, each tweet data will have its own sentiment label. However, before the labeling process, the data must be prepared in advance by doing text cleaning, tokenization, stopwords removal, case folding and data labeling. For labeling purposes, we use a Python library called Sastrawi, which is a library that can be used to determine the basic words of an affixed word (both prefixes, suffixes, infixes and confixes) in Indonesian. The use of this library is very important because the tweet data used in this study are Indonesian-language tweets.

The first stage in labeling is text cleaning which aims to remove punctuation marks, such as: ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ `{ | } ~ , including clearing text of excess spaces. Next is the tokenization stage. At this stage, a sentence is divided into tokens or words. For example, the sentence "*Isolasi terpusat di provinsi Bali*" will be divided into 5 tokens, namely: "*Isolasi*", "*terpusat*", "*di*", "*provinsi*", "*Bali*". We perform tokenization on all tweet data obtained.

The next stage is Stopwords Removal, at this stage the Sastrawi library is used to remove words that are included in stopwords, for example: *"yang", "at", "ke", "para",* and so on. The last step before labeling the data is to do case folding, which is changing the capital letters to lowercase letters. This step is carried out to facilitate the process at the next stage.

In this study, we use the polarity score to perform data labeling, which is based on the number of words that contain negative, positive or neutral meanings. With the polarity score, we determine that if the score is greater than or equal to 0.05, then the sentiment label is "Positive", if it is less than or equal to -0.05, then the sentiment label is "Negative" and outside of these two criteria will be labeled "Neutral".

Preprocessing

Before entering the classification stage, after each sample data has a label, the first thing to do is to calculate the number of sentiments for each label. Then proceed with term-weighting, namely the weighting of terms/texts/words. The method we use in this research is Term Frequency-Inverse Document Frequency (TF-IDF), where this method will provide information on how relevant a term in a document is. The calculation of TF-IDF is done by the formula in equation (1).

$$TFIDF = TF \times IDF \tag{1}$$

Where, Term Frequency and Inverse Document Frequency are respectively described as follows:

a) Term Frequency

Term Frequency (TF) is a measurement of a word's frequency within a manuscript (d). The frequency (TF) of a word in a text is expressed as a percentage of all the words in the text [16]. In Term Frequency, the frequency of a term in a document will be calculated using formula (2).

$$TF_{ij} = \frac{f_d(i)}{\max_{j \in d} f_d(j)} \tag{2}$$

b) Inverse Document Frequency

Inverse Document Frequency (IDF) is a measurement of the importance of a word. The IDF value serves to reduce the weight of a term, if the frequency of occurrence is too high in the entire document set [16]. IDF is calculated using formula (3).

$$IDF(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3}$$

N is the total number of documents, and N=|D|. From this formula, it is possible that if a term does not appear in the entire document, this condition will result in division by zero, where the result is undefined. To avoid this, the denominator can be added with 1, thus |{d∈D:t∈d} |=df(t), then the formula is as in formula (4) below:

$$IDF(t, D) = log\left(\frac{N}{df(t)+1}\right) \tag{4}$$

Classification with NaiveBayes

Naive Bayes is a simple but high-accuracy method used to classify data based on Bayes' Theorem. The Naive Bayes algorithm itself can be described as follows:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (5)$$

Formula (5) which can then be written as follows:

$$posterior = \frac{prior \times likelihood}{eveidence} \quad (6)$$

3. RESULTS AND ANALYSIS

In this study, the extracted data and will be used as a source for conducting Sentiment Analysis are user posts on Twitter social media (tweet data) related to the centralized isolation policy in the province of Bali. One example of the tweet data in question is as shown in Figure 2.

Figure 2. An example of a tweet uploaded on Twitter related to the centralized isolation policy in Bali Province area.

From the results of web scrapping using the SNScrape library, we got a total of 248 tweets related to the centralized isolation policy in Bali Province. The first tweet was published on March 27, 2020 and the last tweet was published on July 21, 2022.

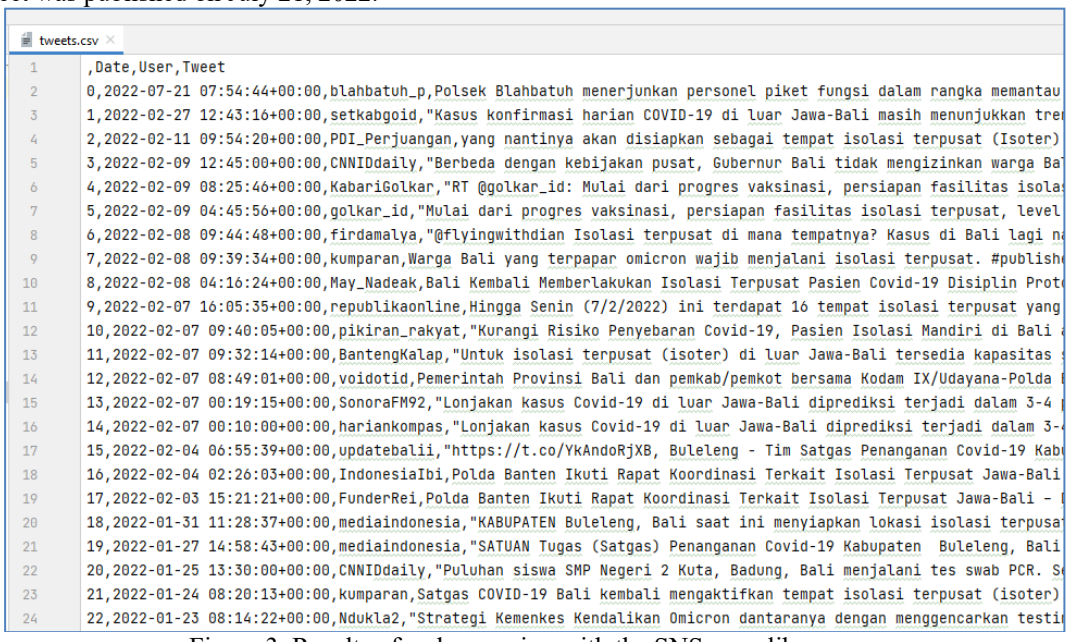

Figure 3. Results of webscrapping with the SNScrape library

Meanwhile, the results of the labeling data on all tweet data used in this study are as shown in Figure 4, with 237 neutral sentiments, 8 positive sentiments and 3 negative sentiments. From these results, it can be

seen that the sentiment is mostly neutral, which indirectly indicates that the centralized isolation policy in the province of Bali has not received a positive or negative response from the community. This possibility can be confirmed from the name of the user who posted the tweet with neutral sentiment, which is a user from the mass media as well as from official government institutions and institutions. This also indicates that the public, especially in the province of Bali, have not shown any rejection or significant controversy with this policy on Twitter.

```
   Unnamed: 0.1  Unnamed: 0  ... Compound Sentiment
0             0           0  ...      0.0   Neutral
1             1           1  ...      0.0   Neutral
2             2           2  ...      0.0   Neutral
3             3           3  ...      0.0   Neutral
4             4           4  ...      0.0   Neutral

[5 rows x 14 columns]
Neutral     237
Positive      8
Negative      3
```

Figure 4. Results of labeling data.

In this study, the Naive Bayes algorithm is implemented with the Multinomial Naive Bayes library, where previously the data was divided into two groups, namely data for training as much as 80% of the total amount, and the rest is data for testing as much as 20% of the total data. The results of the implementation of the Naive Bayes method shown in the confusion matrix are as shown in Figure 5. While the results of the accuracy score are shown in the Precision and Recall matrix with a value of 94%.

```
Confusion Matrix
[[ 0  1  0]
 [ 0 47  0]
 [ 0  2  0]]

Precision and Recall Matrix
              precision    recall  f1-score   support

    Negative       0.00      0.00      0.00         1
     Neutral       0.94      1.00      0.97        47
    Positive       0.00      0.00      0.00         2

    accuracy                           0.94        50
   macro avg       0.31      0.33      0.32        50
weighted avg       0.88      0.94      0.91        50

Accuracy score:  0.94
```

Figure 5. Confusion and Precision & Recall matrix classification results using the Naive Bayes method

Due to the limited number of samples with positive and negative labels, researchers conducted training and testing with the same percentage of data, namely 50% data for training and 50% data for testing. The results are as shown in Figure 6 below:

```
Confusion Matrix
[[  0   1   0]
 [  0 119   0]
 [  0   4   0]]
Precision and Recall Matrix
              precision    recall  f1-score   support

    Negative       0.00      0.00      0.00         1
     Neutral       0.96      1.00      0.98       119
    Positive       0.00      0.00      0.00         4

    accuracy                           0.96       124
   macro avg       0.32      0.33      0.33       124
weighted avg       0.92      0.96      0.94       124
```

Figure 6. Confusion and Precision & Recall matrix with the percentage of training data and testing data are 50% each.

It can be seen in Figure 6, with experiments using 50% of training and testing data, the accuracy score shown in the Precision & Recall matrix is 96%. A very high accuracy score and almost reaching 100% indicates that the resulting model will provide excellent results for any given input.

4. CONCLUSION

From the results of research that has been done, it can be concluded several things as follows:

1. Social media such as Twitter can be used to detect or read public sentiment against government policies through tweets posted by users.
2. From a total of 248 tweets obtained through data scrapping, the Bali provincial government's policy regarding centralized isolation for Covid-19 survivors was responded to neutrally by the public, including the media and institutions/agencies in the province of Bali.
3. With the NaiveBayes classification method, a high accuracy rate of more than 90% was obtained in two experiments with a ratio of training data and testing data of 80:20, and 50:50.
4. From the tweet data obtained and used as a source of training and testing the classification model, the number of positive and negative sentiments is very small. This also greatly affects the diversification of the data and the resulting classification model. Ideally, to find the best model, inputs with the same sentiment percentage are obtained for all sentiment labels used.

REFERENCES

[1] FAQ. Kementerian Kesehatan Republik Indonesia. (n.d.)., Retrieved February 13, 2022, from https://www.kemkes.go.id/folder/view/full-content/structure-faq.html
[2] World Health Organization. (n.d.)., Coronavirus. World Health Organization. Retrieved February 13, 2022, from https://www.who.int/health-topics/coronavirus#tab=tab_3
[3] Syauqi, A. Jalan Panjang Covid19. JKUBS, 1(1), 1-19., 2020
[4] Feldman, R., & Sanger, J., The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data., Cambridge University Press, 2007.
[5] Mejova, Y., Sentiment Analysis: An Overview. University of Iowa, Computer Science Department, 2009.
[6] Feldman, R., Techniques and Applications for Sentiment Analysis. Communications of The ACM, 56 (4), 82-89, 2013.
[7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J., Sentiment Analysis of Twitter Data. In Proceedings of The Workshop on Language in Social Media (LSM 2011) (pp. 30-38), June 2011.
[8] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., & Belyaeva, J., Sentiment Analysis in The News. arXiv Preprint arXiv:1309.6202, 2013.
[9] Kiritchenko, S., Zhu, X., & Mohammad, S. M., Sentiment Analysis of Short Informal Texts. Journal Of Artificial Intelligence Research, 50, 723-762, 2014.
[10] Ortigosa, A., Martín, J. M., & Carro, R. M., Sentiment Analysis in Facebook and Its Application to E-Learning. Computers in Human Behavior, 31, 527-541, 2014.
[11] Medhat, W., Hassan, A., & Korashy, H., Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113, 2014.
[12] Berrar, D. Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403, 2018.
[13] Herdhianto, A., Sentiment Analysis Menggunakan Naïve Bayes Classifier (NBC) PADA Tweet Tentang Zakat (Bachelor's thesis, Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta).
[14] Harijiatno , S.D., "Analisis Sentimen pada Twitter Menggunakan Multinominal Naive Bayes". Skripsi Sarjana Fakultas Sains dan Teknologi Universitas Sanata Dharma, Yogyakarta., 2019.
[15] Susianti, I., Ningsih, S. S., Al Haris, M., &Utami, T. W., Analisis Sentimen Pada Twitter Terkait New Normal Dengan Metode Naïve Bayes Classifier. EDUSAINTEK, 4., 2020.
[16] Jayaswal, V., Text Vectorization: Term Frequency — Inverse Document Frequency (TFIDF), Retrieved October 14, 2022, from https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d