Applied Mathematics and Computational Intelligence
Volume 10, No.1, Dec 2021 [309-318]

Analysis of Public Sentiment on Covid-19 Vaccination Policy Based on Text Mining with The Naïve Bayes Classifier Approach

Rita Susanti[1], Alvito Aryo Pangestu[2], Haydar Arsy Firdaus[3], M. Fariz Fadillah Mardianto[4*]

[1,2,3,4]Statistics, Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

[*] Corresponding author: m.fariz.fadillah.m@fst.unair.ac.id

Received: 25 October 2021; Accepted: 08 November 2021; Available online: 21 December 2021

ABSTRACT

One of the goals in the SDGs, which is to ensure a healthy life and promote the welfare of all people of all ages, has become difficult to maintain since the emergence of Covid-19 in Indonesia. Thus, the Indonesian government has issued a policy regarding the procurement of vaccines and the implementation of vaccinations through Presidential Regulation Number 99 of 2020. Meanwhile, the public's perception of the Covid-19 vaccine that appears are varies and will affect the Covid-19 vaccination process in Indonesia, so a sentiment analysis needs to be carried out to free Indonesia from the Covid-19 pandemic. By using the text mining method, the primary data collected is in the form of public opinions from Twitter. With the Naïve Bayes Classifier approach, it is concluded that the model is consistent and good enough to be used to classify public sentiment regarding the Covid-19 vaccination policy.

Keywords: Naïve Bayes Classifier, Sentiment, Twitter, Text Mining

1 INTRODUCTION

The health issues in the Sustainable Development Goals (SDGs) are integrated into ensuring a healthy life and promoting welfare for all people of all ages. However, this has been difficult to maintain since the emergence of Coronavirus Disease 2019 (Covid-19) in Indonesia. The rapid and massive increase in positive cases of Covid-19 has led the Indonesian government to designate Covid-19 as a national disaster.

The Emergency Committee of the World Health Organization (WHO) has stated that the spread of Covid-19 can be stopped if protection, early detection, isolation, and fast treatment are carried out to create a strong system implementation to stop the spread of Covid-19 [1]. Therefore, the Indonesian government has issued a policy regarding the procurement of vaccines and the implementation of vaccinations in the context of the Covid-19 pandemic through Presidential Regulation Number 99 of 2020. The government has also issued a policy that provides vaccines for free to the public. The vaccination process begins by prioritizing early-stage vaccines for health workers, public service workers, and residents on the age of 18 - 59 years. This policy is problematic because the largest number of deaths due to Covid-19 in Indonesia is dominated by the age group over 60 years [2].

In addition, the policy of imposing criminal penalties in the form of imprisonment and fines for citizens who refuse to be vaccinated [3] is also inappropriate, especially for the long term because the narrative of the outbreak that develops at the community level is very diverse. The public thinks that the Covid-19 pandemic is not a health disaster, but an economic disaster, so that the government is better off immediately allowing economic activities to run normally or increasing the number of social assistance funds rather than investing in vaccines whose effectiveness is still doubtful.

This has led to different perceptions in the community regarding the Covid-19 vaccination policy. One of the things that are of concern due to these differences is the emergence of a movement against vaccines using provocative narratives, hashtags, and hoaxes on social media [4]. Therefore, the government seeks to provide an understanding of the importance of Covid-19 vaccination to the public so that the vaccination process runs smoothly. Thus, sentiment analysis is needed to determine the public's assessment, especially Twitter social media users, of the Covid-19 vaccination policy in Indonesia. Sentiment analysis is important to find out the aspects that need to be considered to formulate policies related to the issues raised.

The method used in this research is the text mining method. Text mining is an unstructured process of extracting data and turning it into useful information for decision-making [5]. Wongkar and Angdresey has classified the public sentiment on Indonesia's 2019 presidential candidates. This study compared the Naïve Bayes Classifier, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) which concluded that Naïve Bayes Classifier has the highest accuracy value of 75.58% [6]. Therefore, this study used the Naïve Bayes Classifier approach. The Naïve Bayes Classifier is one of many methods used for classifying sentiments [7]. The Naïve Bayes Classifier approach is used to classify public sentiment regarding the Covid-19 vaccination policy in Indonesia. Naïve Bayes Classifier is a statistical classification method used to predict the probability of members in a class [8].

The novelty of this research is to use text mining with the Naïve Bayes Classifier approach to determine public sentiment towards the Covid-19 vaccination policy. Covid-19 vaccination is a hot issue that is discussed by the community, especially starting in December 2020 when the Sinovac vaccine first arrived in Indonesia. The issue of Covid-19 vaccination has increased in popularity based on Google Trend. Therefore, this study uses data that is still actual, namely up to February 2, 2021. The author uses sentiment analysis of public opinion regarding the Covid-19 vaccination policy as one of the bases for providing solutions to unravel the polemic in this policy. Based on the sentiment that has emerged, the results of this study are expected to be used as a reference or reference for the government in convincing the public regarding the Covid-19 vaccination process towards a healthy Indonesia.

2 MATERIAL AND METHODS

2.1 Data

The research data is in the form of public tweets regarding the Covid-19 vaccination policy by the government taken from Twitter social media from August 4, 2020 - February 2, 2021, totaling 500 tweets. Out of those 500 data, 400 were used as training data and 100 were used as testing data. Each tweet was then tagged as a positive or negative sentiment.

The data was collected using the Website Crawling technique with Application Program Interface (API) using the Open Source Software-R (OSS-R). Based on the data collection, the variables used in this study consisted of response and predictor variable. The response is the sentiment classification consisting of positive sentiment and negative sentiment, while the predictor is the public's comments regarding the vaccination policy by Indonesia's government. The proportion between the data used as training and testing data is 80% and 20%. The distribution of sentiment on training and testing data is presented in Table 1.

Table 1: Sentiment Distribution on Training and Testing Data

Data Type	Sentiment		Total
	Positive	Negative	
Training	198	202	400
Testing	52	48	100
Total	250	250	500

2.2 Analysis Procedure

Based on Table 1, the number of positive data used as testing data was 20.8%, while 19.2% had negative sentiment. The steps used in this research are as follows:

i) Describe an overview of public sentiment regarding the Covid-19 vaccination policy.

ii) Preprocess the data, which includes spelling normalization, case-folding, tokenizing, filtering, and stemming.

iii) Create a bar chart to show frequently occurring words.

iv) To make Classify public comments regarding the Covid-19 vaccination policy based on sentiment categories using the Naïve Bayes Classifier algorithm and program on training and testing data. The Naïve Bayes Classifier classifies the public comments into positive or negative sentiment based on Naïve Bayes probability. Furthermore, the Naïve Bayes Classifier formula can be found in [9].

v) Calculate the sensitivity and specificity values with the following formula [10]:

$$Sensitivity = \frac{tp}{tp + fn} \times 100\% \quad (1)$$

$$Specificity = \frac{tn}{fp + tn} \times 100\% \quad (2)$$

with,

tp: number of true positive

fp: number of false positive

fn: number of false negative

tn: number of true negative

Then, calculate the chance of misclassification using an evaluation procedure called the Apparent Error Rate (APER). The APER value represents the value of the proportion of samples that were misclassified by the function [11]. The APER value is defined by (3):

$$APER(\%) = \frac{fp + fn}{tp + fp + tn + fn} \times 100\% \tag{3}$$

and the accuracy value in (4) is as follows [10]:

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \times 100\% \tag{4}$$

with,

tp: number of true positive

fp: number of false positive

fn: number of false negative

tn: number of true negative

vi) Measuring the stability of classification using Press' Q which is defined as follows:

$$Press'\ Q = \frac{[N - (nk)]^2}{N(k-1)} \tag{5}$$

with N is the total number of samples, n is the number of individuals with the right classification and k is the number of groups. The hypothesis test used is as follows:

H_0 : Unstable or inconsistent classification

H_1 : The classification is stable or consistent

Classification accuracy is said to be consistent if the value of $Press'\ Q > X^2_{(1,\alpha)}$ [12].

3 RESULTS AND DISCUSSION

3.1 Data Overview

The categories of positive and negative sentiments obtained are the same, namely 50% each. The equality of the number of sentiments illustrates that the strength of the people who support the Covid-19 vaccination policy is equal to those who refuse. Then, the pre-processing stage for public sentiment data regarding the Covid-19 vaccination policy was carried out in the steps shown in section 2.2.

Every public tweet related to the Covid-19 vaccination policy has words that often appear. These words will later be used in the classification stage. Bar charts are used to visualize the words. The following Figure 1 shows a bar chart of 10 positive sentiment words that often appear.

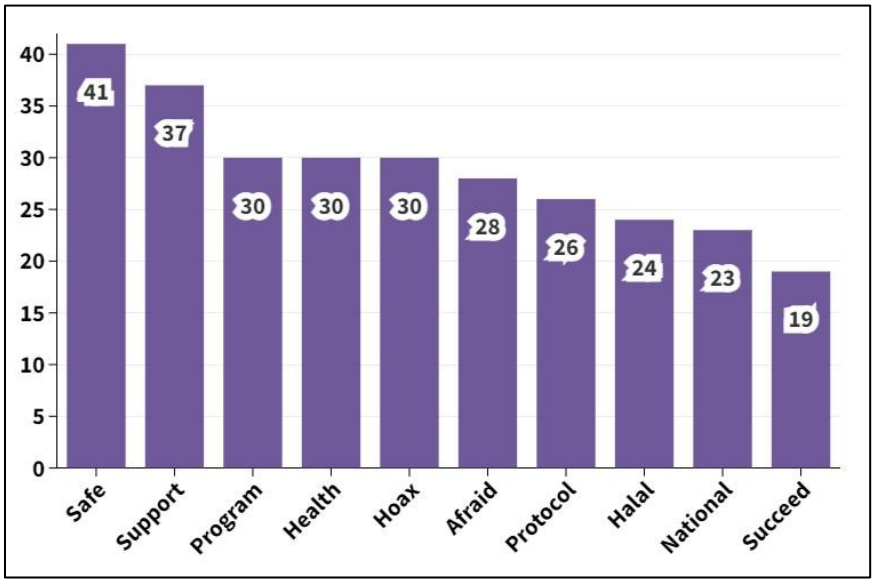

Figure 1: Bar Diagram for the Appearance of Words on Positive Public Sentiment regarding the Covid-19 Vaccination Policy

Figure 1 describes the 10 words with the highest frequency. Based on Figure 1, it can be seen that the word “safe” is the word that most often appears in positive public sentiment regarding the Covid-19 vaccination policy. This is influenced by the polemic of the safety of using the Covid-19 vaccine. The majority of people with positive sentiments believe in the safety of the Covid-19 vaccine so that people support the vaccination policy. Therefore, the words “safe” and “support” are in the first and second place in Figure 1.

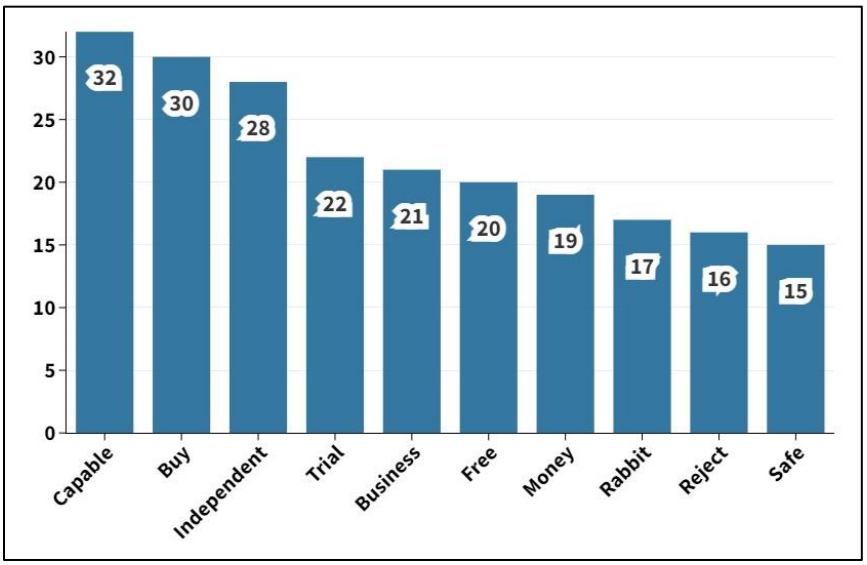

Figure 2: Bar Chart of Word Appearances on Negative Public Sentiments regarding the Covid-19 Vaccination Policy

Based on Figure 2, it can be seen that the word "capable" is the word that most often appears in negative public sentiment regarding the Covid-19 vaccination policy. Even though the word is categorized as a positive word, this word describes public's anxiety about the ability of the Sinovac vaccine in dealing with Covid-19 so that many choose to buy independent vaccines. Therefore, the words "capable", "buy", and "independent" are in the first to third place in Figure 2.

3.2 Sentiment Classification

There are 1540 words produced after the preprocessing stage. All words in the training data will be used to create a Naïve Bayes Classifier model.

3.2.1 Training Data Classification

Classification using the Naïve Bayes Classifier probability value in the training data is carried out to determine the value of the accuracy of the classification. The classification results can be seen in Table 2.

Table 2: Training Data Classification Results

Prediction	**Actual**		**Total**
	Positive	**Negative**	
Positive	198	29	227
Negative	0	173	173
Total	198	202	400

Based on Table 2, it can be seen that 14.35% of negative sentiments experienced prediction errors, which are classified as positive sentiments. Not a single positive sentiment has experienced a prediction error. The classification results in Table 2 need to know the percentage of sensitivity, specificity, APER, and accuracy. The calculation of these values for the training data is based on (1), (2), (3), and (4).

The sensitivity of 87.23% indicates that the model is very good for predicting positive sentiments, while the specificity of 100% indicates that the model is very good for predicting negative sentiments. Then, the accuracy percentage of 92.75% indicates that the probability value generated by the Naïve Bayes Classifier method on training data is very good for classifying public sentiment towards the Covid-19 vaccination policy. Furthermore, the stability of the classification can be determined by calculating the value of the Press' Q test statistic with (5). The hypothesis proposed is as follows:

H_0: The results of the classification of public sentiment regarding the Covid-19 vaccination policy on training data are unstable or inconsistent

H_1: The results of the classification of public sentiment regarding the Covid-19 vaccination policy on training data are stable or consistent

The calculation of the Press' Q value on the results of the training data classification is conducted by (5). The Press' Q value obtained is 292.41. This value is then compared with the statistical value of the Chi-square statistics of $\chi^2_{0.05(1)} = 3.841$. It can be seen that the Press' Q value is greater than 3.841 or is in a critical area. Thus, it can be concluded that the results of the classification of public sentiment regarding the Covid-19 vaccination policy on training data are stable or statistically consistent.

3.2.2 Testing Data Classification

The testing data classification is carried out by the same procedure. The results of the testing data classification can be seen in Table 3.

Table 3: Testing Data Classification Results

Prediction	**Actual**		**Total**
	Positive	**Negative**	
Positive	48	20	68
Negative	4	28	42
Total	52	48	100

Based on Table 3, it can be seen that 41.67% negative sentiment and 7.69% positive sentiment experienced prediction errors. This figure is greater than the training data. It is necessary to calculate

the sensitivity, specificity, APER, and accuracy. The calculation of these values for the testing data is based on (1), (2), (3), and (4).

The sensitivity of 70.59% indicates that the model is good enough to predict positive sentiment, while the specificity of 87.5% indicates that the model is very good for predicting negative sentiment. Then, the accuracy percentage of 76% indicates that the probability value generated by the Naïve Bayes Classifier method on training data is good enough to be used for classifying public sentiment towards the Covid-19 vaccination policy. Then, the value of the Press' Q test statistic is calculated based on (5) to determine the stability of the testing data classification. The hypothesis proposed is as follows.

H_0: The results of the classification of public sentiment regarding the Covid-19 vaccination policy on testing data are unstable or inconsistent

H_1: The results of the classification of public sentiment regarding the Covid-19 vaccination policy on testing data are stable or consistent

The calculation of the Press' Q value on the results of the training data classification is conducted by (5). The Press' Q value obtained is 27.04. This value is then compared with the statistical value of the Chi-square statistics of $\chi^2_{0.05(1)} = 3.841$. It can be seen that the Press' Q value is greater than 3.841 or is in a critical area. Thus, it can be concluded that the results of the classification of public sentiment regarding the Covid-19 vaccination policy on training data are stable or statistically consistent.

Although the Naïve Bayes Classifier method on testing data is stable or consistent, this method is good enough to be used to classify public sentiment regarding the Covid-19 vaccination policy, given the not too high level of classification accuracy, namely 76%. However, this value is higher than the previous study that conducted by Wongkar and Angdresey [6].

Based on Figure 2, the word with the highest frequency of negative sentiment is the word “capable”. Some people still doubt the ability of the Sinovac vaccine in dealing with Covid-19. The public thought that because the efficacy of the Sinovac vaccine in Indonesia is only 65.3% [13]. This makes some people think that the community is made a kind of guinea pig by the government. For that reason, the words “trial” and “rabbit” came in fourth and eighth place on the negative sentiment. However, the word “safe” in Figure 1 is the word that has the highest frequency of positive sentiment. This means that some people still believe that the Sinovac vaccine is safe to use.

In addition, the words “buy” and “independent” rank second and third in Figure 2. This was triggered by the government's plan to allow independent vaccination [14]. The community thinks that independent vaccination is safer because people can choose the Covid-19 vaccine that has much greater efficacy than the Sinovac vaccine. Apart from the independent vaccination policy, the government is also trying to work with the private sector to cut costs [15]. This is what makes people think that the vaccination policy is turning into a business. The number of people who think this way is not small, given that the word “business” is in fifth place on negative sentiment.

The Covid-19 vaccination policy is a government effort to free Indonesians from the Covid-19 pandemic. The existence of criticism due to deficiencies in its implementation can be used as material for evaluating the Covid-19 vaccination policy to make it better. This is closely related to the expectations of the community to leave the house immediately and carry out their daily activities as usual. However, this hope is difficult to realize if the Covid-19 vaccination policy itself raises a

polemic among the public. The government and society should be able to be open to each other in listening to and giving criticism regarding the Covid-19 vaccination policy to accelerate the end of the Covid-19 pandemic towards a healthy Indonesia which is in line with the SDGs achievement targets in the health sector.

4 CONCLUSION

The positive and negative sentiments regarding the Covid-19 vaccination policy obtained have the same proportion. In addition, the results of the classification of training and testing data were stable or statistically consistent with the Press' Q test statistical values of 292.41 and 27.04, respectively. The training data generated APER values and accuracy of 7.25% and 92.75%, respectively, while the sensitivity and specificity were 87.23% and 100%.

The testing data generated the APER values, accuracy, sensitivity, and specificity of 24%, 76%, 70.59%, and 87.5%, respectively. Thus, it is concluded that the model is good enough to be used for the classification of positive sentiment for training data and negative sentiment for testing data. Overall, the Naïve Bayes Classifier method is a method that is quite appropriate to use to classify public sentiment regarding the Covid-19 vaccination policy because the accuracy value in training and testing data is quite high.

ACKNOWLEDGEMENT

The author would like to thank all parties who played a role during the research process. In addition, the authors also thank the Faculty of Science and Technology for providing support in this research activity.

REFERENCES

[1] P. Sun, X. Lu, C. Xu, W. Sun, and B. Pan, "Understanding of Covid-19 Based on Current Evidence," *Journal of Medical Virology,* vol. 92, no. 6, pp. 548-551, 2020.

[2] B. Djaafara, F. R. Andiwijaya, F. Verisqa, I. Fadilah, K. Saraswati, and R. Maulida. "Indonesia's Decision to Prioritise Covid-19 Vaccination to Citizens Aged 18 – 59 Years Old Questionable." The Conversation. https://theconversation.com/indonesias-decision-to-prioritise-covid-19-vaccination-to-citizens-aged-18-59-years-old-questionable-153883 (accessed Feb. 10, 2021).

[3] The Jakarta Post. "Decision to Punish Anti-Vaxxers Under Regional Administrations: Task Force." The Jakarta Post. https://www.thejakartapost.com/news/2020/12/25/decision-to-punish-anti-vaxxers-under-regional-administrations-task-force.html (accessed Feb. 10,2021).

[4] The Jakarta Post. "COVID-19 vaccine hoaxes could impact other immunization programs." The Jakarta Post. https://www.thejakartapost.com/news/2020/12/07/covid-19-vaccine-hoaxes-could-impact-other-immunization-programs.html (accessed Feb. 15,2021).

[5] M. P. Bach, Z. Krstic, S. Seljan, and L. Turulja, "Text Mining for Big Data Analysis in Financial Sector: A Literature Review," *Sustainability*, vol. 11, no. 1277, pp. 1-27, 2019.

[6] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naïve Bayes Algorithm of The Data Crawler: Twitter," in *Proceeding – 2019 IEEE International Conference on Informatics and Computing*, 2019, pp. 1-5.

[7] N. S. M. Asri and N. Nordin, "A Classifying and Ranking Selection of Healthcare Tourism Services", *Applied Mathematics and Computational Intelligence*, vol. 9, pp. 9-20, 2020.

[8] W. B. Zulfikar, M. Irfan, C. N. Alam and M. Indra, "The Comparation of Text Mining with Naïve Bayes Classifier, Nearest Neighbor, and Decision Tree to Detect Indonesian Swear Words on Twitter," in *Proceeding – 2017 IEEE International Conference on Cyber and IT Service Management*, 2017, pp. 1-5.

[9] X. Feng, S. Li, C. Yuan, P. Zeng, and Y. Sun, "Prediction of Slope Stability Using Naïve Bayes Classifier," *KSCE Journal of Civil Engineering*, vol. 22, no. 3, pp. 941-950, 2018.

[10] N. M. Ranjan and R. S. Prasad, "Automatic Text Classification Using BPLion-neural Network and Semantic Word Processing," *The Imaging Science Journal,* vol. 66, no. 2, pp. 69-83, 2017.

[11] M. Hasyim, D. S. Rahayu, N. E. Muliawati, D. Hayuhantika, R. Puspasari, D. Anggreini, R. C. Hastari, S. Hartanto, and F. H. Utomo, "Bootstrap Aggregating Multivariate Adaptive Regression Splines (Bagging MARS) to Analyse the Lecturer Research Performance in Private University," *Journal of Physics: Conf. Series,* vol. 1114, no. 1, p. 012117, 2018.

[12] C. M. Navas, J. V. D. Bermejo, A. K. McLean, J. M. L. Jurado, A. B. B. R. D. Torres, and F. J. V. Gonzales, "Discriminant Canonical Analysis of the Contribution of Spanish and Arabian Purebred Horses to the Genetic Diversity and Population Structure of Hispano-Arabian Horses," *Animals,* vol. 11, no. 269, pp. 1-27, 2021.

[13] A. Syakriah. "Indonesia Allows Emergency Use of Sinovac Vaccine." The Jakarta Post. https://www.thejakartapost.com/news/2021/01/11/indonesia-allows-emergency-use-of-sinovac-vaccine.html (accessed Feb. 20, 2021).

[14] M. Handayani and D. Kurniawan. "Vaccination Remains Free, Companies Must Buy and Are Not Allowed to Cut Salaries." VOI. https://voi.id/en/news/28736/coordinating-minister-airlangga-independent-vaccination-remains-free-companies-must-buy-and-are-not-allowed-to-cut-salaries (accessed Feb. 21, 2021).

[15] A. Syakriah. "Indonesia on Track for Jan. 13 Vaccination Drive." The Jakarta Post. https://www.thejakartapost.com/news/2021/01/09/indonesia-on-track-for-jan-13-vaccination-drive.html (accessed Feb. 21, 2021).