Detection of Public Sentiment Analysis Model on the Implementation of PPKM in Indonesia

R P Henessa, M A Fisabilillah, W R Azizah

Politeknik Statistika STIS, Jl. Otista Raya No. 64C, Jakarta Indonesia

*Corresponding author's e-mail: 221911087@stis.ac.id

Abstract. Covid-19 pandemic which has been being serious problem in Indonesia indirectly force Indonesian government to issue policies in order to decrease the number of Covid-19 spread. One of the policies is the Implementation of Restrictions on Community Activities (PPKM) in Java-Bali region from January 11-25, 2021. Due to its continued implementation, this policy raises pros and cons in the community. This research's goal is to determine the best classification model and determine the effect of adding feature engineering in analyzing public sentiment on PPKM with scrapping data from Twitter so that with the best model, it is possible to classify public responses to PPKM automatically. The twitter scrapping dataset is preprocessed first, which includes case folding, tokenizing, filtering, stemming, and term weighting to clean the data. After preprocessing and through the analysis steps, it concludes that using feature engineering can increase the accuracy of the best selected four models. The logistic regression method with feature engineering with accuracy rate of 87.50% become the best method. In conclusion, the best suggested model to analyze public sentiment using Twitter scrappimg towards *PPKM* is by using the logistic regression.

1. Introduction

Pandemic Covid-19 which has been occurred for 18 months is being serious problem for government and society in Indonesia. The reason is that various types of restrictions on community activities are still being implemented by the government in suppressing the increase in the spread of Covid-19. Starting from the Indonesian government's policy as Large-Scale Social Restrictions (*PSBB*). However, this policy has not been able to completely suppress the spread of Covid-19 because thousands of *PSBB* violations are still being found. According to Traffic Directorate of Polda Metro Jaya, there are more than 23,000 people who violate traffic rules related to PSBB and a number of other areas that have implemented PSBB are still operating normally in traditional markets [1].

The violations of PSBB triggered the government to issue a new policy, namely the Implementation of Restrictions on Community Activities (PPKM) in Java-Bali region from January 11-25, 2021 with the aim of only limiting community activities without having to do lockdown the area as in PSBB policy [2]. According to BPS, the results of community behavior survey related to emergency PPKM conducted on July 13-20, 2021, stated that 60% of respondents felt bored during the implementation of PPKM [3]. It could happen because the implementation of PPKM continues to be extended every period so that many people complain because of the decreasement of economy [4].

During the implementation of PPKM in limiting people's mobility, people spend a lot of time in their home. Whether doing work from home, online class, or many other activities. This provides free time for people to access social media more often than usual because work at home can be done

flexibly compared to work in the office. Social media is a place for users to share and interact in a virtual world using internet-based technology [5]. By using social media, people are free to express their opinion, complaints, or other things in response to events happens in today's digital world.

One of the most popular social media in Indonesia is Twitter. A lot of users interested in using Twitter because they can interact easily in asking questions, giving opinions, and replying each other [6].

Based on this problem, a method is needed to quickly and precisely classify public sentiment towards the implementation of PPKM during the Covid-19 pandemic in order to find out the community's actual response to government policies. Data mining, which is a process of collecting and extracting large data then processing it into information can be used as a basis for decision making, has a classification method that is able to handle this problem. By using data mining approach, Classification can be used to label a tweet whether it is a positive or negative automatically from built model so that conclusions and collecting information can be done quickly.

2. Data Collecting

The dataset that being used is public dataset by Moch Kholil uploaded in July 2021 on https://www.kaggle.com/mochkholil/ppkm-sentiment-classification/data?select=ppkm_dataset.csv [7]. The data is the result of scrapping on Twitter that has been exported in the form of a csv file with a total of two columns, class and comments, and a total record of 300 tweets. The labelling of the dataset consists of three categories: negative, neutral, and positive. This data is raw and unstructured, so it needs to been preprocessing to clean and bring the data in a certain format in order to ease the process of analysis and model-forming.

3. Theoretical Basis

3.1. Preprocessing

The preprocessing needs to be done to clean up data before doing deep analysis for data training and data testing. Instance that is cleaned up somehow an instance that being very diverge or commonly known as outlier [8]. Cleaning up process aims to converting unstructured inputted data to certain form. There are few steps of preprocessing to be done:

3.1.1. Case Folding. In preprocessing text, case folding is used to change every uppercase letter in the input document to lowercase letters [9]. This step is to similize text in document.

3.1.2. Tokenizing. Tokenizing is a step to separate text in a document into a sequence of tokens in form of words, terms, symbols, or other partitions [10]. Tokenizing is to explore words in a sentence [11]. The list of tokens that have been obtained from the tokenizing process will be used in text mining.

3.1.3. Filtering. From the list of tokens that have been obtained from the tokenizing process, the next step is filtering. Filtering is a process to remove unimportant words by using list of stopword removal [12]. In this paper, list of stopword that being used is class StopWordRemoverFactori from library Sastrawi.

3.1.4. Stemming. Stemming is a step of converting words into their basic form by removing their suffixes. This step has a significant effect on preprocessing performance [13]. This step has various analysis on each language because the main goal is to know the meaning of a word even though the form is different.

3.1.5. Term Weighting. The term weighting process is a step of calculating the weight of each word or term in each document to knowing the similarity of a word or term in the document [14]. In term weighting, there are two well-known assumptions: TF and IDF. TF assumption calculate how often a word or term appears in a document. IDF assumption calculate how a term can be widely distributed

or stated that a term that occurs infrequently is no less important than a term that occurs frequently [15]. These assumptions can be expressed in the following TF-IDF equation:

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot log \frac{|T_r|}{\#T_r(t_k)} \quad (1)$$

Where $\#T_r(t_k)$ shows the number of documents in T_r when t_k occurs at least once and

$$tf(t_k, d_j) = 1 + log\, \#(t_k, d_j) \;\; when\, \#(t_k, d_j) > 0 \quad (2)$$

Where $\#(t_k, d_j)$ shows how many times t_k occurs in d_j. $tf(t_k, d_j)$ and $log \frac{|T_r|}{\#T_r(t_k)}$ components from equation (1) stands each assumption of TF and IDF.

3.2. Feature Engineering

The development of features in a model is intended to support the improvement of the goodness of the model. Features that suitable with the built model will have a good impact on the prediction results. Feature engineering is the main process of preprocessing in machine learning. Feature engineering involves the application of transformation functions such as arithmetic and aggregate operators to produce something new [16]. In this feature, transformation helps to create feature scale or convert non-linear association between features and target classes into easy-to-learn linear association. Thus, in the process of sentiment analysis, it will be very helpful if between words that initially do not have any correlation are converted into linear association so that produce a decision in classification.

Feature engineering by utilizing NLP (Natural Language Processing) technique is a combination or interaction between computers and human ability of language and communication. By using NLP technique, the representation results expected to be closer to reality because it is based on assumptions that come from human ability in language. As a part of artificial intelligence, NLP technique needs to be tried and developed to produce a better representation in text classification [24].

3.3. Classification Method

3.3.1. Random Forest. Random forest is a method that "considered promising" for classifying data because of its algorithm's simplicity and its classification performance which is very prominent for high-dimensional data [17]. One of the popular random forest constructions is proposed by Breiman by randomly selecting the feature subspace at each node to grow a decision tree branch then using the Bagging method, a subset of training data is generated to build an individual tree then each individual trees are combined to form random forest model [18]. The implication is a random forest in its principle is an amalgamation of individual trees, the tree will also grow along with the increasement in the dimensions or size of the dataset.

*3.3.2. SVM. S*VM classifier is a method used to obtain an optimal hyperplane to separate observations with different target variable values. SVM aims to minimize the upper limit of generalization error by maximizing the margin between the separating hyperplane and the data [19]. SVM is a supervised learning method that is considered more effective in terms of concepts because it contains clear mathematical calculations compared to other classification methods. Commonly a data problem cannot be separated linearly in the input space or in other words SVM is not able to find the separator on hyperplane. Thus, an additional kernel is needed to transform to a higher dimension to get great accuracy and good generalization ability.

3.3.3. Logistic Regression. The regression function works on the association between independent variable and dichotomous dependent variable [20]. The logistic regression method in classification process will analyze the association between input features and certain output opportunities produced. Logistic regression is widely chosen because of its high accuracy and fairly simple model with one of

the outputs or the goal is to predict the probability of an event occurring or not based on the calculated predictor value.

3.3.4. Naïve Bayes Method. Naïve Bayes method is a classification method based on the Bayes theorem which assumes that its features are independent classes [21]. Naïve Bayes is quite effective in conducting text-based classification, system work management, and diagnosing health problems. The Naïve Bayes classification method is optimal for the concept of two classes with nominal features that assign a value of 0 for the first category and 1 for the other category with probability 1 [22].

3.4. Sentiment Analysis

Sentiment analysis is a text data extraction technique to get responses or sentiments that is classified into positive, negative, or neutral [23]. The value obtained in this sentiment analysis can be adjusted according to needs. In this paper the dataset was selected with three labels: positive, negative, and neutral. With the development of information technology, it also triggers an increase in social media users. Therefore, sentiment analysis can be carried out by utilizing opinions and facts that are spread on social media and website media.

3.5. Model Evaluation

A good system is a system that can fulfil the purpose of itself. If the need is to classify, then the system is good system if it can provide the right classification results. However, it is very rare or even rare to find a system whose performance is absolutely 100% perfect because the system is a human-made product that may still have a bit or more human error in it.

Therefore, in model evaluation process there is an accuracy calculation to state how well the model has been built. A good model is certainly a model that has a high level of accuracy. However, a high the degree of accuracy tends to be relative depends on the analysis needs and also the availability of datasets that can be used in model formation. The equation of accuracy value of a model can be written as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Where the values of TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are obtained from the confusion matrix of model formation.

4. Research and Result

This research begins with a search for available scrapping from Twitter datasets on kaggle.com web related to the topic of sentiment analysis of *PPKM*. Then, the dataset is downloaded and cleaned up from unnecessary attributes through preprocessing.

The first step of preprocssing starts with case folding, which is the process of converting the letter format of the obtained tweets into lowercase letters. Next step is tokenizing, which is the process of cutting text into words, terms, symbols, and punctuation that are unnecessary. Then, punctuation marks that are considered unimportant will be removed at this tokenizing step. Next step is filtering, the process of choosing important words from the tokenizing results which are capable to represent the document. Next is the stemming step, which is the process of converting words extraction into the basic form by removing their suffixes. The last step to do in preprocessing is term weighting, which is to remove words whose frequency of occurrence is less than five times.

Besides the preprocessing steps described in the previous steps, this research also used several classification methods with slightly different processes. By using the same dataset, there are two kinds of tests were carried out, with and without using feature engineering. The feature engineering that being used is NLP (Natural Language Processing). NLP is a sub-section of artificial intelligence that learns the interaction between computers and human language according to experience. The use of feature engineering is by counting the number of question marks and exclamation points occurred in

the user's tweets and then inserting them into the dataset table by forming a new column. The calculation of question marks occurrence in tweets is assumed that user comments that lead to negative sentiments may not contain question marks while the calculation of exclamation points occurrence in tweets is assumed that user comments with exclamation marks are very likely to be contained in negative sentiments because the presence of these marks is usually related to someone's anger towards something or indicates a high and firm tone of speech. In addition, text transformation into vector is also carried out by calculating the similarity between comments. Similar comments will earn high marks. This calculation is also based on the assumption that a person's negative response to something is sometimes repeated several times in human-to-human communication. The goal is to find out how influential feature engineering is in increasing the accuracy of the obtained model. The results of calculating the accuracy of several classification models with and without using feature engineering are presented in table 1:

Table 1. The Accuracy of Classification Models With and Without Using Feature Engineering (%).

Method		Accuracy
With Feature Engineering	Logistic Regression	87.5
	Naïve Bayes	85.41
	SVM	81.25
	Random Forest	81.25
Without Feature Engineering	Logistic Regression	75
	Naïve Bayes	68.33
	SVM	76.67
	Random Forest	80

From table 1, by using feature engineering to the four classification methods, the level of accuracy *is* increased. In processing without using feature engineering, the classification method with the highest accuracy level *is* the random forest method with an accuracy rate of 80%. Meanwhile, after adding engineering features to the calculation, the highest level of accuracy changed to the logistic regression method, which is 87.50% followed by the Naïve Bayes method in the second position with accuracy of 85.41% followed by SVM and random forest with an accuracy of 81.25%.

From the four methods, if you look at the evaluation value of the model, the application of feature engineering has a positive impact because it can increase the accuracy value. This indicates that the application of feature engineering has been appropriate for these four methods.

5. Conclusion

From the research and result above, it can be concluded that based on the evaluation of the model, the best model is obtained is logistic regression with feature engineering which has accuracy rate of 87.50%. Thus, from the four classification methods that have been tested, the logistic regression method is the right method used to classify public sentiment for Twitter related to the *PPKM*. In addition of conclusion, feature engineering can improve the goodness of the classification model which being tested. Therefore, good knowledge is needed in the selection of appropriate features to increase the accuracy of a model.

In addition, this research does not determine whether the majority of people's sentiments towards the *PPKM* program are whether the majority of people like it, don't like it or are neutral with it because this paper only focused on testing the formation of several models which were ultimately adopted with high accuracy models so that in future research this model can be used to take into account the majority of public sentiment towards *PPKM* on different occasions.

References

[1] Callistasia W. Virus corona : rencana pelonggaran PSBB, 'apa yang mau dilonggarkan? ini sudah longgar sekali' [Internet]. BBC News Indonesia. 2020 [cited 09 September 2021]. Available from : https://www.bbc.com/indonesia/indonesia-52631514

[2] CNN Indonesia. Habis PSBB terbitlah PPKM, apa bedanya? [Internet]. CNNIndonesia. 2021 [cited 09 September 2021]. Available from : https://www.cnnindonesia.com/nasional/20210107807043820-590992/habis-psbb-terbitlah-ppkm-apa-bedanya

[3] Septian D. Survei BPS: 60 Persen Masyarakat Merasa Jenuh Selama PPKM [Internet]. Liputan6. 2021 [cited 09 September 2021]. Available from : https://www.liputan6.com/bisnis/read/4621951/survei-bps-60-persen-masyarakat-merasa-jenuh-selama-ppkm

[4] A Aziz Said. Dampak PPKM darurat, BI pangkas proyeksi ekonomi tahun ini jadi 3,8% [Internet]. Katadata. 2021 [cited 09 September 2021]. Availeble from : https://katadata.co.id/yuliawati/finansial/60ec52dd70190/dampak-ppkm-darurat-bi-pangkas-proyeksi-ekonomi-tahun-ini-jadi-3-8

[5] Cahyono AS. Pengaruh media sosial terhadap perubahan sosial masyarakat di Indonesia. Jurnal Publiciana. 2016;**9(1):**140-57.

[6] Nurjanah WE, Perdana RS, Fauzi MA. Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN. 2017;**2548**:964X.

[7] Kholil, Moch (2021, July). ppkm_sentiment, Version 2. Retrieved September 09, 2021 from https://www.kaggle.com/mochkholil/ppkm-sentiment-classification/data?select=ppkm_dataset.csv

[8] Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data preprocessing for supervised learning.Internationaljournal of computer science. 2006 Jan;**1(2)**:111-7.

[9] Rosid MA, Fitrani AS, Astutik IR, Mulloh NI, Gozali HA. Improving text preprocessing for studentcomplaint document classification using sastrawi. InIOP Conference Series: Materials Science and Engineering 2020 Jun 1. **874(1)**. 012017. IOP Publishing.

[10] Gupta G, Malhotra S. Text document tokenization for word frequency count using rapid miner (taking resume as an example). Int. J. Comput. Appl. 2015;**975**:8887.

[11] Kannan S, Gurusamy V, Vijayarani S, Ilamathi J, Nithya M, Kannan S, Gurusamy V. Preprocessing techniques for text mining. International Journal of Computer Science & Communication Networks. 2014 Oct;**5(1):**7-16.

[12] Amalia A, Lydia MS, Fadilla SD, Huda M. Perbandingan Metode Klaster dan Preprocessing Untuk Dokumen Berbahasa Indonesia. Jurnal Rekayasa Elektrika. 2018 Apr 27;**14(1)**:35-42.

[13] Can F, Kocberber S, Balcik E, Kaynak C, Ocalan HC, Vursavas OM. Information retrieval on Turkish texts. Journal of the American Society for Information Science and Technology. 2008 Feb 1;**59(3)**:407-21.

[14] Sabbah T, Selamat A, Selamat MH, Al-Anzi FS, Viedma EH, Krejcar O, Fujita H. Modified frequency-based term weighting schemes for text classification. Applied Soft Computing. 2017 Sep 1;**58**:193-206.

[15] Debole F, Sebastiani F. Supervised term weighting for automated text categorization. InText mining and its applications 2004. 81-97. Springer, Berlin, Heidelberg.

[16] Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga DS. Learning Feature Engineering for Classification. InIjcai 2017 Aug 19. 2529-35.

[17] Xu B, Guo X, Ye Y, Cheng J. An Improved Random Forest Classifier for Text Categorization. J. Comput.. 2012 Dec 1;**7(12)**:2913-20. dan Pembobotan Jumlah Retweet. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN. 2017;**2548**:964X.
[18] L Breiman. Random forest. 2001. Machine Learning. **45**. 5-32.
[19] Amari SI, Wu S. Improving support vector machine classifiers by modifying kernel functions. Neural Networks. 1999 Jul 1;**12(6):**783-9.
[20] Ramli, Desi Yuniarti dan Rinto Goejantoro. Perbandingan metode klasifikasi regresi logistik dengan jaringan saraf tiruan. May, 2013. EKSPONENSIAL Journal. **4**. 17-24.
[21] Rish I. An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4;**3(22)**:41-6
[22] Rish I, Hellerstein J, Thathachar J. An analysis of data characteristics that affect naive Bayes performance. IBM TJ Watson Research Center. 2001 Jun 1;**30**:1-8.
[23] Sari FV, Wibowo A. Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer. 2019 Nov 29;**10(2):**681-6.
[24] Scott S, Matwin S. Feature engineering for text classification. InICML 1999 Jun 27;**99**:379-388.