Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter

R L Mustofa*, B Prasetiyo
Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Indonesia

*Corresponding author: refolabib@students.unnes.ac.id

Abstract. Back In 2020, World Health Organization (WHO) has announced COVID-19 as a pandemic. From the many public responses, especially those on Twitter regarding the #newnormal campaign, a sentiment analysis process needs to be carried out to find out the perceptions that exist in society through social media. In this study, data were obtained through the crawling process on Twitter using the Twitter API. The method used in the sentiment analysis process is lexicon-based. The lexicon-based method works by labeling words containing sentiments based on a lexicon dictionary that already has weight on each word or doesn't have weight on words in a lexicon dictionary. The classification results using lexicon-based are also used to make training data in the testing process using the naive Bayes classifier algorithm. In general, the research stages in this sentiment analysis include data crawling, text preprocessing, feature extractions, and the classification process. The sentiment analysis process results showed that the percentage of social media users on Twitter about #newnormal was 33.19% containing negative sentiments and 66.36% containing positive sentiments. Meanwhile, for testing the naive Bayes classifier algorithm in the sentiment analysis process got an accuracy of 79.72%.

1. Introduction
In 2020, World Health Organization (WHO) had announced COVID-19 as a pandemic and had not been able to fully control the pandemic because it is waiting for a vaccine to cure it [1]. The stipulation of COVID-19 as a pandemic led the Indonesian government to issue various policies related to handling COVID-19 by campaigning for new normal.

From the many responses from the public, especially those on social media regarding the #newnormal campaign, it is necessary to process or mine data mining to find out the perceptions that exist in the community. From the many responses from the public, especially those on social media regarding the #newnormal campaign, it is necessary to carry out a data mining process to find out the perceptions that exist in the community. The data mining process often used in research on social media is about sentiment analysis which is part of text mining. The text mining process that is often used in research on social media is about sentiment analysis. This sentiment analysis will classify responses or opinions into several categories: very positive, positive, neutral, negative, and very negative [2]. Previous studies that examined sentiment analysis included [3,4,5] and sentiment analysis that focused on the Covid-19 case [6].

In the sentiment analysis classification process, there are several methods. Two of them are machine learning and lexicon-based. Machine learning methods train text classifiers on labeled training datasets. Sentiment classification based on machine learning can be formulated as learning based on a problem...
In comparison, the lexicon-based approach involves calculating the document orientation for a semantic word or phrase in a document [8]. One of the word dictionaries that have been made is the lexicon dictionary from Chen and Skiena [9] which researched making lexicon sentiments in 136 main languages, including Indonesian, using the graph propagation approach. Techniques for analyzing sentiment are usually carried out by researchers, namely by using several classification algorithms such as the Support Vector Machine (SVM), Naive Bayes Classifier (NBC), and Decision Tree.

The Naive Bayes classifier is a simple probability classifier that calculates a set of probabilities by adding up the frequencies and value combinations from a given dataset. The algorithm uses the Bayes theorem and assumes that all attributes are independent or are not interdependent given by the value of the class variable[10].

The advantage of using the naive Bayes classifier is that the naive Bayes classifier only requires a small amount of training data to determine the estimated parameters needed in the classification process [11]. As one example in the journal belonging to [12]. Researchers in the sentiment analysis process use the naive Bayes classifier algorithm by grouping Facebook user sentiments into two categories: positive sentiment and negative sentiment. The sentiment analysis process used the naive Bayes classifier with the unigram baseline feature in this study resulted in an accuracy of 70%. In addition, the naive Bayes classifier is considered to work very well compared to other classifier models because the naive Bayes classifier has a better level of accuracy[13].

Goel [14] makes a study by comparing the naive Bayes classifier algorithm with several algorithms such as support vector machines and decision trees. The accuracy results obtained from the naive Bayes classifier and support vector machine in this study affect the datasets. Meanwhile, the resulting accuracy of the decision tree algorithm is not affected by the number of datasets.

Based on the description of the problems above, this study will try to focus on analyzing the lexicon-based method's use as labeling for each sentence in the sentiment analysis. Naive Bayes classification algorithm used as machine learning to predict sentiment analysis classification on social media, Twitter.

2. Methods
In this study, a sentiment analysis process was carried out on the #newnormal hashtag using the lexicon-based method with the naive Bayes classifier algorithm. The whole process of sentiment analysis can be seen in Figure 1.

2.1 Data collecting
The dataset used is tweet data about the #newnormal hashtag on social media Twitter. The data collection technique uses a crawling technique using the Twitter API with RStudio. The keyword #newnormal hashtag is used to get tweet data on Twitter. In this study, we are crawling with the target of retrieving 1000 tweet data. The lexicon dictionary used is the Indonesian lexicon dictionary based on the research results of Bing [15] and Wahid (2016) [16], which contains 1182 positive words and 2402 negative words. Collection of stopword removal words obtained from research Tala et al. (2003)[18].

2.2 Data preprocessing
Sentiment analysis is text-based research so that the data to be used must go through the text preprocessing stage. The preprocessing stage used in this study includes cleansing, case folding, and stopword removal. Cleansing performs cleaning on tweets that contain unnecessary characters such as symbols, numbers, and punctuation marks. After going through the cleansing stage, the text is carried out with a case-folding process. In the case folding process, all letters that are still capitalized will be converted into lowercase letters. The last step in preprocessing is stopword removal. Stopword removal aims to remove meaningless words such as abbreviations and conjunctions.
2.3 Lexicon based feature

Lexicon Based Features is a word feature with positive sentiment and negative sentiment based on the word dictionary or lexicon. Lexicon is a collection of sentiment words that have been known and collected [18]. There are several steps in lexicon-based, such as determining the polarity of words, handling negation, and giving each tweet entity a score. The formula for calculating the entity score can be seen in equation 1.

$$\text{Score}(e) = \sum_{wi: wi \in L \cup \text{words}} \frac{wi \cdot so}{\text{dist}(wi, e)}$$ (1)

2.4 Machine learning model

In this experiment, a machine learning model was created to carry out the sentiment analysis process. Machine learning will work by predicting tweets containing positive or negative sentiments. In this experiment, the machine learning model uses two algorithms as a comparison, the two algorithms, namely the naive Bayes classifier algorithm.
2.5 Confusion matrix
To determine the machine learning model's accuracy from the naive Bayes classifier and the percentage of sentiment on the #newnormal hashtag, we use an accuracy test using the confusion matrix method. The confusion matrix table can be seen in Table 1.

| Predicted Positive | Actual Positive |          | Actual Negative |
|--------------------|-----------------|----------|-----------------|
| Predicted Negative |          |          |                  |

Calculation of accuracy, precision, and recall can be seen in the equation 2, 3 and 4.

\[
\text{precision} = \frac{TP}{TP + FP} \tag{2}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}
\]

3. Results and Discussion
From the crawling process results, the data obtained were 984 observations (rows) and 90 variables (columns). This study focuses on the sentiment analysis process so that the variable data taken is only tweet data taken from the text variable. Then some of the other variables that are taken as a support are the created_at and screen_name variables. Besides, tweet data taken in the crawling process in this study are tweets that use Indonesian. Sample data from the crawling process can be seen in Table 2.

| Created_at        | Screen_name         | Text                                                                 |
|-------------------|---------------------|----------------------------------------------------------------------|
| 2020-07-17 03:15:37 UTC | oktanonymous28      | Harusnya masih #dirumahaja \nTapi pemerintah terlalu memaksakan untuk #NewNormal \nYah gini jadinya,, Korban covid-19 makin banyak\nHmm,,, |
| 2020-07-16 09:37:48 UTC | HanGabriella        | Era new normal membuat banyak orang semakin cemas, yang pada akhirnya mengakibatkan stres. Apa kamu salah satu yang merasakannya juga? #NewNormal #Viral #Tips #Stress #ericko https://t.co/3UwaWjacIK |
| 2020-07-18 11:50:32 UTC | klinovasi           | Memulai gaya hidup baru dengan membawa bekal dari rumah untuk menghindari penyebaran virus dan bakteri\nGlasslock dapat menjadi pilihan yang tepat bagi Anda, untuk memulai gaya hidup sehat di fase #newnormal \U0001f917\n#adglasslock #glasslockindonesia #glasslockfoodstorage https://t.co/zZdDVReqXS |

The data processing process then carries out the data obtained on the data collection stage. In this study, the data processing stage includes the preprocessing stage and feature extractions using the lexicon-based method. The results of preprocessing tweet data after going through the cleansing, case folding, and stopword removal stages can be seen in Table 3.
**Table 3. Text preprocessing result**

| Input                                                                 | Output                                                                 |
|-----------------------------------------------------------------------|------------------------------------------------------------------------|
| Memulai gaya hidup barumu dengan membawa bekal dari rumah untuk menghindari penyebaran virus dan bakteri \textbackslash{n}Glasslock dapat menjadi pilihan yang tepat bagi Anda, untuk memulai gaya hidup sehat di fase \#newnormal \textbackslash{U0001f917} \#glasslock \#glasslockindonesia \#glasslockfoodstorage | gaya hidup barumu membawa bekal rumah menghindari penyebaran virus bakteri glasslock pilihan gaya hidup sehat fase newnormal glasslock glasslockindonesia glasslockfoodstorage |

Era new normal membuat banyak orang semakin cemas, yang pada akhirnya mengakibatkan stres. Apa kamu salah satu yang merasakannya juga? \#NewNormal \#Viral \#Tips \#Stress \#ericco https://t.co/3UwaWjacIK

Memulai gaya hidup barumu dengan membawa bekal dari rumah untuk menghindari penyebaran virus dan bakteri \textbackslash{n}Glasslock dapat menjadi pilihan yang tepat bagi Anda, untuk memulai gaya hidup sehat di fase \#newnormal \textbackslash{U0001f917} \#glasslock \#glasslockindonesia

| Word                          | Frequency |
|-------------------------------|-----------|
| newnormal                     | 988       |
| covid                         | 329       |
| kesehatan                     | 148       |
| virus                         | 134       |
| protokol                      | 121       |
| corona                        | 113       |
| normal                        | 109       |
| sindangkasihfm                | 108       |
| new                           | 95        |
| cegahviruscoronomenyebar      | 93        |

In the lexicon-based feature stage, the value and sentiment label will be weighted. Each word will have a score value as well as a label, positive and negative. Lexicon-based works by matching data with existing lexicon dictionaries. Tweets that score more than zero will get positive label results, while those that score less than zero will get negative label results. The process of weighting words used the lexicon-
based feature can be seen in Figure 2, and the results of the lexicon-based feature can be seen in Table 5.

**Table 5. Lexicon based feature result**

| Text                                      | Score | Sentiment |
|-------------------------------------------|-------|-----------|
| gaya hidup barumu membawa bekal rumah     | 2     | Positive  |
| menghindari penyebaran virus bakteri      |       |           |
| glasslock pilihan gaya hidup sehat        |       |           |
| fase newnormal glasslock glasslockindonesia glasslockfoodstorage | | |
| era new normal orang cemas mengakibatkan  | -3    | Negative  |
| stres salah merasakannya newnormal viral tips | | |
| stress ericko dirumahaja pemerintah       |       |           |
| memaksakan                                | -2    | Negative  |
| newnormal yah gini korban covid hmm       |       |           |

This study's classification results include classification using lexicon-based, and classification using the naive Bayes classifier algorithm based on training data made with the classification method lexicon-based. Based on the feature extraction stage using lexicon-based, the sentiment percentage of the total tweet data about #newnormal can also be seen in Figure 2.

![Tweet #newnormal Lexicon based feature](image)

**Figure 2. Lexicon based feature result**

From the results of the analysis produced from previous research, the resulting accuracy of each algorithm in sentiment analysis is deeper in the use of the method in the labeling process and the variation in parameter values of each algorithm used in making machine learning models such as in the example of using the support vector machine algorithm with test parameters. They are using the degree value and the learning rate constant value. Whereas in this study, the most optimal accuracy is produced with a ratio of 70:30 and parameter values in the word vectorization process using the max_feature = 4000 value. The max_feature value in this study is determined based on the total words in the lexicon dictionary. Sentiment classification results data in machine learning model testing on the naive Bayes classifier algorithm can be seen in Table 6.
Table 6. The results of the naive Bayes classifier

| Prediction | True positive | True negative | Class precision |
|------------|---------------|---------------|-----------------|
| positive   | 192           | 6             | 96%             |
| negative   | 44            | 54            | 55%             |
| Class recall| 81%           | 90%           |                 |

At the stage of making a machine learning model, training sets of 688 data with a percentage of 70% of the total data were then trained using the naive Bayes classifier algorithm. The results of machine learning testing on the naive Bayes classifier algorithm show an accuracy of 78.68%.

4. Conclusion
The research results show that tweets with positive sentiment are more than tweets that contain negative sentiment. Of the total data, as many as 984 from the crawling process results, tweets containing positive sentiments of 66.36% and tweets containing negative sentiments of 33.19%. This percentage is obtained from the lexicon-based feature process for preprocessing datasets. Meanwhile, testing the machine learning model using the parameter comparison 70:30 and max_feature = 4000 for sentiment analysis using the naive Bayes classifier algorithm shows an accuracy of 79.72%.

References
[1] Bing L, Minqing H & Junsheng C 2005 Proc. 14th Int. Conf. World Wide Web 342.
[2] Chen Y and Skiena S 2014 Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. 2 383
[3] Larasati U I, Muslim M A, Arifudin R, and Alamsyah 2019 Sci. J. Inform. 6 138
[4] Muslim, M A 2020 J. Soft Comput. Explor. 1 8
[5] Fransiska S, Rianto R and Gufroni A I 2020 Sci. J. Inform. 7 203
[6] Efrilianda D A, Dianti E N, and Khoirunnisa O G 2021 J. Soft Comput. Explor. 2 1
[7] Desai M and Mehta M A 2016 Int. Conf. Comput. Commun. Autom. 149
[8] El Zowalaty M E and Järhult J D 2020 One Health 9 100124
[9] Goel A, Gautam J & Kumar S 2017 Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. 257
[10] Patil, T R., & Sherekar, S 2013 Int. J. Comput. Sci. Appl. 6, 256–261.
[11] Hailong Z, Wenyang G & Bo J 2014 Proc. 11th Web Inf. Syst. Appl. Conf. 262
[12] Liu B 2010 IEEE Intell. Syst. 25 76
[13] Xhemali, D, J Hinde, C, & G Stone, R 2009 Int. J. Comp. Sci. 4(1) 16–23.
[14] Pattekari S A and Parveen A 2012 Int. J. Adv. Comput. Math. Sci. 3 290.
[15] Troussas C, Virvou M, Espinosa K J, Llaguno K & Caro J 2013 4th Int. Conf. Inf. Intell. Syst. Appl. 198
[16] Turney P D 2002 Co R R, cs.LG/0212. http://arxiv.org/abs/cs/0212032
[17] Tala, F, Kamps, J, Müller, K, & M, R 2003. Int. J. Symbol. Logic. 30 271
[18] Wahid D H and Azhari 2016 Indones. J. Comput. Cybern. Syst. 10 207
[19] Provost, F, & Kohavi, R 1998 Glossary of terms. J. of Machine Learning 30 271