Jurnal Elektronik Ilmu Komputer Udayana
Volume 9, No 3. February 2021

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Sentiment Analysis Of Tribal, Religion, And Race With LIWC

Prasetyo Adi Utomo[a1], AAIN Eka Karyawati[a2]

[a]Informatics Department, Faculty of Math and Science, Udayana University
South Kuta, Badung, Bali, Indonesia
[1]pras.au404@email.com
[2]eka.karyawati@unud.ac.id

Abstract

During this pandemic, social media has become a major need as a means of communication. One of the social media used is Twitter by using messages referred to as tweets. In Indonesia itself, there are various tribes, religions, and races in their society so the use of these names is also become commonly used. However, sometimes, the use of the name is followed by negative sentiment that used to insult and aimed at an individual or group. To filter that kind of tweets, a sentiment analysis was performed with LIWC method that divides tweets into 3 classes of positive, neutral, and negative. From the sentiment analysis that has been performed, the average score for precision is 69.62%, recall is 70%, and f-measure is 69.81%.

Keywords: *Sentiment Analysis, Tweet, LIWC, Indonesia, Religion*

1. Introduction

The need for social media has become part of Indonesian society. Moreover, in present that the current Covid-19 pandemic happen where social media are used as the main means to communicate due to social restrictions to prevent the spread of viruses. One of the social media that used is the Twitter where users communicate using tweet as a message. Messages in tweets can contain such as congratulations messages, event descriptions, or can be issues and opinions about a person, politics, or government regulations. In Indonesia, there are many tribes, religions, and races where people often use the name of a tribe, religion, or race to indicate something like to congratulate someone or something else. In using tweets, people also often use those words. However, there are also tweets that use the name of a tribe, religion, or race and contain negative sentiment or bad words to insult an individual or a group. The use of such tweets can lead to fights between tribes, religions, or races. To prevent this, it is necessary to select or filter tweets that contain an insult or bad wording of tribal, religion, or race.

To find out if a tweet contains a tribe, religion, or race and know the sentiment of the tweet whether it is positive, neutral, or negative can be done by applying sentiment analysis to the tweet. Sentiment analysis is a field of science that analyzes opinions, attitudes, evaluations, and assessments of an event, topic, organization, or individual [3]. In sentiment analysis itself, approaches that can be used are machine learning-based and lexicon-based approach. The example of machine learning-based is sentiment analysis using Naïve Bayes method which have been carried out by [7]. The research conduct a study about classification of snack review and the performance in their research is 80.5% for the average accuracy score. Lexicon-based sentiment analysis method that can be used is Linguistic Inquiry Word Count or LIWC for short. Using the LIWC method, researcher analyzed the sentiment to determine the sentiment of the tweet which can be positive, neutral, or negative sentiment. To find out how the performance of the sentiment analysis is performed, the scores of precision, recall, and f-measure are used as the performance values of the analysis. Research of LIWC have been carried out by [2]. The research conducted a study about comparing text classification method where LIWC is one of them. LIWC method in their research have performed and getting accuracy about 43% to 62.5%. The other research about LIWC have carried out by [1]. The research conducted a study

about comparing the performance of LIWC_2007pt and LIWC_2015pt using Brazilian Portuguese lexicon. The result that can get from their research is LIWC_2015pt outperforms LIWC_2007pt. Based on previous research, author hope this study can classify Indonesian tweet using LIWC really well and have a good result.

2. Research Method

2.1 Research Stage

The research is divided into several stages. These stages are the data collection stage, the preprocessing stage, the sentiment analysis stage with LIWC method, and the results evaluation phase. Here is the flowchart that show how the flow of the research done.

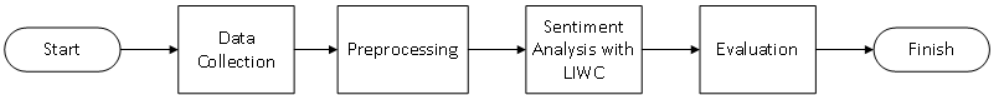

Figure 1. Research Stage

2.2 Data Collection

The data that being used is an indonesian tweet data. Data collection is done using the Twitter API and tweepy library in Python. Data searching are performed by searching for specific words such as religion or race names such as "kristen" and "madura" and filtering for retweets. The amount of the data that have been collected is 300. The data is divided into 3 classes namely positive data, neutral data and negative data with the amount of data as much as 100 data each. Data labeling is done manually by the author and assisted by colleagues.

2.3 Preprocessing

Preprocessing is a process to convert data that still does not have a meaning into data that has meaning and can be processed. The preprocessing stage is done to make the data "clean" so that errors in data processing can be reduced and make the process more efficient. Here are the flowchart that show how the stages in preprocessing.

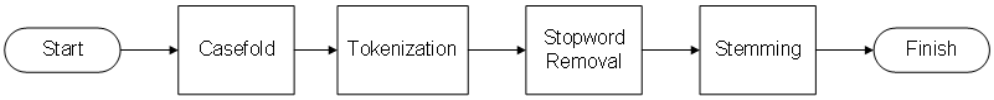

Figure 2. Preprocessing

Casefolding is a process to create the same form of data that contains only lowercase letters. Casefolding is done so that the existing data is equal. Tokenization is the process for creating tokens from the initial data. Tokens are a smaller part of the initial data [5]. Tokenization is done so the process of sentiment analysis with LIWC can be done because LIWC method is using the words for analysis. In this tokenization process also carried out the calculation of the number of words on the tweet which will be used in the process of sentiment analysis with LIWC. Stopword removal is a process for removing words that are very commonly used and have no meaning in performing sentiment analysis [5]. Stopword removal is done to make the process run more efficiently. Stemming is the process of removing the prefix or suffix in the data so that it turns into a basic form. Stemming is done to equate data that has different writing.

2.4 Linguistic Inquiry Word Count

LIWC is a text analysis application developed with the aim of analyzing the emotions, cognitive, and structural components of a text [6]. Using LIWC, a tweet analysis is performed so that the sentiment of the tweet is known. LIWC works by searching for each word in text and matching it to a word in the lexicon, and adding a percentage ratio value of that category if a category is found [6]. Therefore, in process

p-ISSN: 2301-5373
e-ISSN: 2654-5101

of sentiment analysis on an Indonesian tweet with LIWC, a lexicon or dictionary is required to contain the words and sentiments of the word. Lexicon or Dictionary creation is done manually [2]. There only 2 class in lexicon, positive and negative. When the word categorization process on a tweet is complete and the category percentage results have been obtained, it can be compared to which category presentation is larger to determine the class of the tweet. The stages of sentiment analysis with LIWC are as follows.

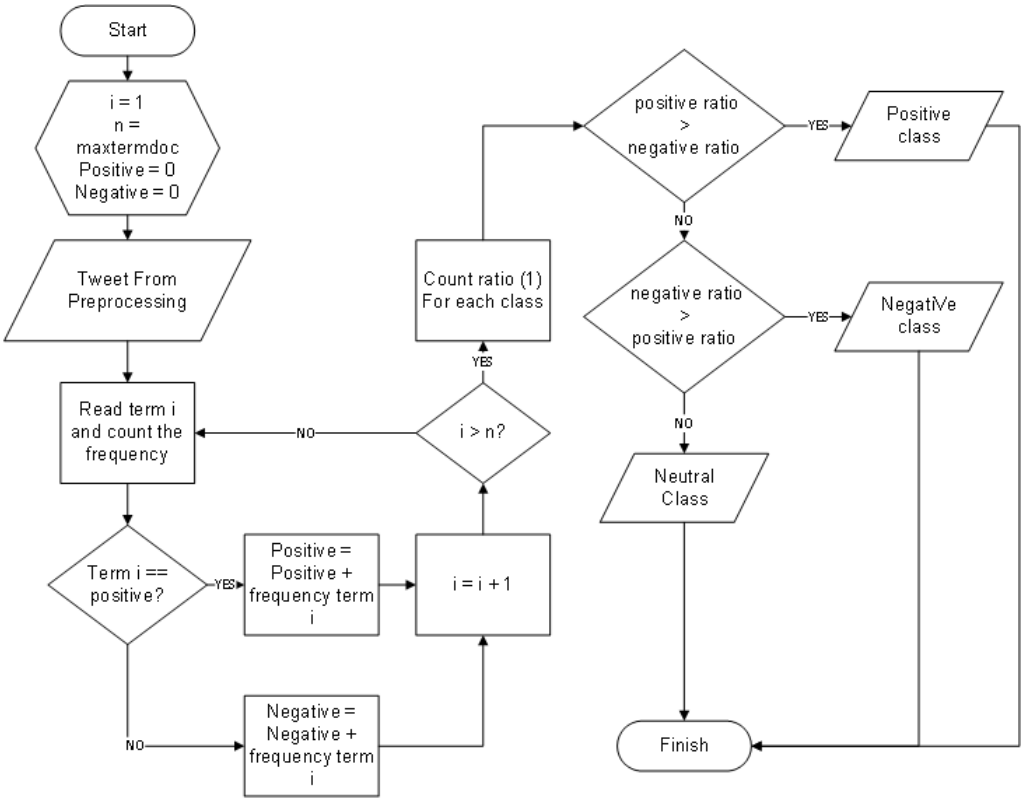

Figure 3. Linguistic Inquiry Word Count

a. Read all the terms in tweet from preprocessing
b. Check if the term is on the lexicon
c. If available, calculate the frequency and calculate the ratio as follows [2]:

$$Rt = \frac{f_t}{N} \quad (1)$$

With Rt is the ratio of the term t, ft is the frequency of the term t, and N is the total number of words on the document.

d. Sum ratios into class categories according to lexicon
e. Perform step b through d until all terms have been checked and calculated
f. Compare ratio between classes. If the negative ratio is greater than the positive then the tweet is negative, and vice versa. If the ratio of both classes is equal or "0" then the tweet is neutral.

2.5 Evaluation

The evaluation stage is carried out to find out how the sentiment analysis is performed. Evaluation is done by calculating the score of precision, recall, and f-measure of each class and calculate the average. Here's how to find precision, recall, and f-measure values for each class [4].

$$\text{PrecisionK} = \frac{TPk}{TPk + FPk} \quad (2)$$

TPk is used for the sum of the tweet that have correct prediction of class k. FPk is used for the sum of prediction of class k that wrong. The sum of TP and FP is total of all the prediction of class k.

$$\text{RecallK} = \frac{TP}{TP+FN} \quad (3)$$

FN is used for the sum of real data that have wrong prediction of class k. The sum of TP and FN is total of all real data of class k.

$$\text{F-MeasureK} = \frac{2*PrecisionK*RecallK}{PrecisionK+RecallK} \quad (4)$$

The average calculation is done by summing the precision, recall, or f-measure scores of all classes and divided by 3 because the amount of data of each class is already the same.

3. Result and Discussion

From sentiment analysis with LIWC that has been done, the prediction results are obtained as follow.

Table 1. Prediction Result

	Positive (Predict)	Neutral (Predict)	Negative (Predict)	Total (Real)
Positive (Real)	81	15	4	100
Neutral (Real)	26	48	26	100
Negative (Real)	10	9	81	100
Total (Predict)	117	72	111	300

From sentiment analysis on tweets, predictions are obtained as in the Table 1. Where the correct prediction result is 81 data for positive class, 48 data for neutral class, and 81 data for negative class. For prediction errors that occur in sentiment analysis, the amount is 19 data for positive classes, 52 data for neutral classes, and 19 data for negative classes. From the prediction results, precision, recall, and f-measure scores for each class are calculated and obtained as follows.

Table 2. Precision, Recall, F-Measure

	Positive	Neutral	Negative	Average
Precision	0.6923	0.6667	0.7297	0.6962
Recall	0.81	0.48	0.81	0.7
F-Measure	0.7465	0.5581	0.7678	0.6981

From the result of sentiment analysis's performance that have been calculated, for positive class, the score of precision is 0.6923 or 69.23%, recall's score is 0.81 or 81%, and f-measure' score is of 0.7465 or 74.65%. For neutral class, the score of precision is 0.6667 or 66.67%, recall's score is 0.48 or 48%, and f-measure's score is 0.5581 or 55.81%. And for negative class, the score of precision is 0.7297 or 72.97%, recall's score 0.81 or 81%, and f-measure's score is 0.7678 or 76.78%. After the average being calculated for each precision, recall, and f-measure and get an average score of 0.6962 or 69.62% for the average precision score, 0.7 or 70% for the average recall score, and 0.6981 or 69.81% for the average f-measure score.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

4. Conclusion

From the research that has been done, that is to analyze sentiment in tweets containing tribals, religions, and races of 300 data (each class of 100 data) to find out the sentiment in the tweet whether it is positive, negative, or neutral sentiment using linguistic method inquiry word count obtained an average result of precision score of 69.62%, recall score of 70% , and the f-measure score is 69.81%. It also can be seen that the predictive evaluation score of the data with the negative class is greater than the other two classes. So it can be concluded that the application of linguistic methods of inquiry word count to perform sentiment analysis on tweets containing tribal, religion, and race names can be applied and obtained good results, especially to analyze negative sentiments from tweets.

Referensi

[1] F. Carvalho, R. G. Rodrigues, G. Santos, P. Cruz, L. Ferrari, and G. P. Guedes, “Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks,” pp. 24–34, 2020, doi: 10.5753/brasnam.2019.6545.
[2] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, “Comparing automated text classification methods,” *Int. J. Res. Mark.*, vol. 36, no. 1, pp. 20–38, 2019, doi: 10.1016/j.ijresmar.2018.09.009.
[3] Liu, B., 2012. Sentiment Analysis and Opinion Mining. In: Chicago: Morgan & Claypool Publisher.
[4] M. Ahmad, S. Aftab, and I. Ali, “Sentiment Analysis of Tweets using SVM,” *Int. J. Comput. Appl.*, vol. 177, no. 5, pp. 25–29, 2017, doi: 10.5120/ijca2017915758.
[5] Manning, C., Raghavan, P. & Schütze, H. (2009). An Introduction to InformationRetrieval. Cambridge: Cambridge University Press.
[6] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas.
[7] I. G. C. P. Yasa, N. A. Sanjaya ER, and L. A. A. R. Putri, “Sentiment Analysis of Snack Review Using the Naïve Bayes Method,” *JELIKU*, vol. 8, no. 3, pp. 333–338, 2020.

This page is intentionally left blank