Classification of Emotions on Twitter using Emotion Lexicon and Naïve Bayes

Dhiya Fairuz Ray Dzahabiyyah[a,1], Novi Yusliani *[b,2], Kanda Januar Miraswan[b,3]

[a,b] Artificial Intelligence Laboratory, Faculty of Computer Science, Sriwijaya University
[1] fairuzdhiyay@gmail.com; [2] novi_yusliani@unsri.ac.id; [3] kandajm@yahoo.com

ARTICLE INFO

Article history
Received 30 Mei 2021
Revised 5 Juni 2021
Accepted 12 Juni 2021

Keywords
Emotion
Emotion Lexicon
Classification
Naïve Bayes
Twitter

ABSTRACT

Social media is a means of interaction and communication. One of the social media that is often used is Twitter. Twitter allows its users to express many things, one of which is being a personal media to provide various kinds of expressions from its users such as emotions. Users can express their emotions and sentiments through writing on the status of their social media posts. One method to find out the emotion in the sentence is using the Emotion Lexicon. However, the lexicon-based method is not good at classifying data because not every word contains emotion. So, there's a need to combine it with other classification method such as Naive Bayes. Naïve Bayes relies on independent assumptions to obtain a classification through the probability hypothesis that each class has. The results of the classification test with Emotion Lexicon alone have 46% accuracy, 45% precision, 51% recall and 36% f-measure. While the results of the classification test with Emotion Lexicon and Naïve Bayes resulted in an accuracy of 65%, precision of 77%, recall of 55%, and f- measure of 59%.

1. Introduction

The growth of social media users increased rapidly during recent years. Therefore, social media is now a new trend for people to interact and communicate. One of the current most popular and widely used social media is Twitter [1]. Twitter is a micro blogging social networking website where users can express their opinion in a short and simple with posts called tweets [2]. Tweets can contains a variety of things, starting from a description of an information to personal media which ultimately can contain human behavior including emotions [3].

Emotions are feelings that come from certain circumstances, moods, or one's relationship to another. Often referred to as pattern complex changes that reflect behavioral reactions, appear as a response to a situation that is felt personally significant by a person [4].

In daily life, most people use the method communication to express emotions about various events to every little thing that happens around them, and the most common way to expressing emotions is through speech and facial expressions [4]. However, since human behavior can actually be captured from emotions in the form of text, it allows people tend to express feelings that are felt through their social media posts [1]. Where emotions play an important role in social media and they are used to represent 1000's of other language forms that are spoken and understood universal [2].

Emotion Lexicon is a dictionary that contains a list of words for indicates the type of emotion associated with the word [5]. The Emotion Lexicon was chosen because it is useful for identifying emotions that evoked by a word [6] with a focus on divided Plutchik's eight basic emotions such as anger, fear, anticipation, trust, surprise, sadness, joy, and disgust [7]. Emotion Lexicon cannot be used as an opinion or a specific result, because not every word on sentences contain emotions [8]. Classification using Lexicon Based is very dependent on the dictionary (lexicon), if a sentence does not have words that contain emotion, then the sentence will be classified as "unknown", this causes poor accuracy results. Therefore there is a need of combining classification methods to increase the accuracy results, such as Naïve Bayes [9].

Naive Bayes has often been applied in several other classifications because it is easy to implement, the process is fast and can work well [9]. The advantage of Naïve Bayes itself is on the independent assumption which helps to obtain a quick classification through the probability hypothesis that obtained as the probability that each class has, meaning that each of the words in the sentence will be categorized according to their class. So even if a sentence does not contain emotion, every word in the sentence will categorized by their emotional class [2].

In this research, Emotion Lexicon and Naïve Bayes will be used for emotion classification on Twitter. Emotion Lexicon is used to know the meaning of each word in a tweet. Naïve Bayes will derive a hypothesis probability for the classification results.

2. Literature Study

In this section contains the theoretical foundation and some research that has been done by previous researchers. This was made to strengthen the reasoning and rationality of the involvement of several variables in this study. It also functions as a scientific opinion that is integrated with the results of a literature review to build a researcher's mindset in relation to the problem being studied. Research conducted by [2] because of an increase in research on understanding of human emotions that play an important role in life. Then Twitter as a micro blogging social networking website skyrocketed because people can express their emotions in short and simple. By using Twitter content, emotions can be classified, and used Naïve Bayes as a classification method that works with good at text classification. Of the 500 tweets, emotion extraction (sad, joy,trust) resulted that the average accuracy, precision and recall of 99.31% for every emotion class.

Research conducted by [5] the rapid development of social media so it concerns the analysis of sentiments and emotions as outlined in the post in text form. The emotional lexicon used in this study developed using the Indonesian Thesaurus book and focus describe the process of developing the emotional lexicon, especially in Indonesian language. The development of the lexicon consists of two processes, seed selection words and lexicon expansion where each word will be given a binary weight 1 or 0. This study contains the initial results of the development of the lexicon with produced 1165 emotion lexicon words in Indonesian.

Research conducted by [10] for the development of human-computer interaction by detecting emotions on Facebook social media which is used as a tool for communication, a place to express opinions in everyday life using lexicon approach and Natural Language Processing. Emotional Lexicon test managed to get 55.45% or 15,357 of 27,696 words from user status Facebook. Natural Language Processing is used to improve text comes from status updates. The results of these improvements are matched with the lexicon which has been made to know the label of their emotions. Then from 26 status Facebook has detected an emotional label of 61.53%.

3. Methodology

The data used in this research is comes from the *kaggle* site which is a collection of tweets in *csv* format. Data in the form of text that will go through text pre-processing, training the data to create classifier model (Emotion Lexicon and Naïve Bayes) and testing the data, all stages are illustrated in Fig 1.

3.1. Preprocessing

The pre-processing stage or data pre-processing is a process for preparing raw data before other processes are carried out [11]. The pre-processing includes case folding, tokenization, and stemming. In this research the case folding process will change all the letters in the document into lowercase letters. Furthermore, the sentence will be broken up into pieces of words or known as tokenizing. The last step is stemming the process of reducing words to basic forms after filtering, where the algorithm used is the Porter library for stemming English words.

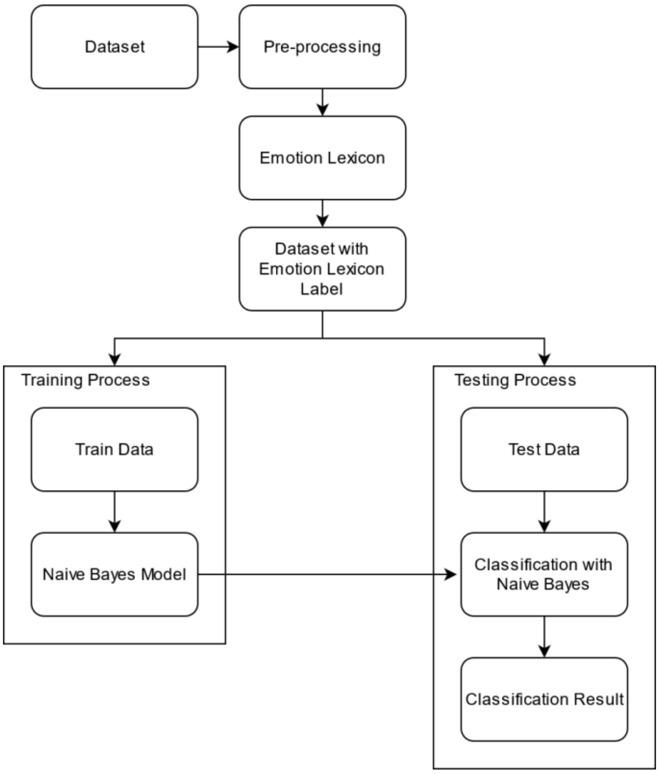

Fig. 1. Framework Diagram

3.2. Emotion Lexicon

Emotion Lexicon are method that use lexicon from NRC which contains words with their emotions. All words have their own weight according to their emotion. One word can have multiple weight which means one word can have one or more emotions [12].

3.3. Naïve Bayes

Naïve Bayes was proposed by the British scientist Thomas Bayes, which contains the theory of probability that predicts future opportunities based on experience in the past [9]. Naive Bayes is a method used in statistics to calculate the probability of a hypothesis, which assumes all attributes are not interdependence or independent is given by the value on the class variable by calculating the highest probability [13]. The basis of the Naive Bayes theorem used in this study is the following Bayes formula (1):

$$P\,(H|X) = \frac{P(H).P(H)}{P(X)} \tag{1}$$

ISSN 2807-2391

The Naive Bayes theorem explains that the classification process requires some pointers that will assign a class fit to the sample that analyzed as stated in equation (2):

$$P\ (C|F_1, \ldots, F_n) = \frac{P(C)P(F_1, \ldots, F_n|C)}{P(F_1, \ldots, F_n)} \tag{2}$$

Based on equation (2), the variable C is identified as a class, while the variables F_1, ..., F_n show the characteristics needed to classify. The formula in equation (2) explains that the probability of entering a sample of certain characteristics in class C (Posterior) is the probability of the appearance of class C (previously called prior), multiplied by the probability of occurrence of sample characteristics in class C (also called likelihood), divided by the probability of occurrence of characteristics - characteristics of the sample globally (also called evidence). Therefore, the formula is expressed as equation (3):

$$Posterior = \frac{Prior\ x\ likelihood}{evidence} \tag{3}$$

The value of evidence will remain consistent for each class. The value of the posterior will be compared with the values of the other posterior classes to determine which class a sample will be classified in. Further decomposition $(C|F_1, \ldots,)$ by means of multiplication as in equation (3):

$$\begin{aligned} \mathrm{P}(\mathrm{C}|F_1, \ldots, F_n) &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1, \ldots, F_n|\mathrm{C}) = \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2, \ldots, F_n|\mathrm{C}, F_1) \\ &= \mathrm{P}(\mathrm{C})\mathrm{P}((F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C}, \mathrm{F1})\mathrm{P}(F_3, \ldots, F_n|\mathrm{C}, F_1, F_2) \\ &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C}, F_1)\mathrm{P}(F_3|\mathrm{C}, F_1, F_2), \mathrm{P}(F_4, \ldots, F_n|\mathrm{C}, F_1, F_2) \\ &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C}, F_1)\mathrm{P}(F_3|\mathrm{C}, F_1, F_2) \ldots \mathrm{P}(F_n|\mathrm{C}, F_1, F_2, F_3, \ldots, F_{n-1}) \end{aligned} \tag{4}$$

The results of the decomposition as in equation (4) produce many complex factors that affect the probability value, which is almost impossible when analyzed one by one. This makes the calculation difficult to do. Here the assumption of very high (naive) independence is used, that each clue $(\boldsymbol{F_1}, \ldots,)$ is independent of each other. With these assumptions, it is shown in equation (5):

$$P(P_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \tag{5}$$

for i ≠ j

$$P(F_i|C, F_j) = P(F_i|C)$$

Based on the equation (5), it can be concluded that the assumption of naive independence makes the probability conditions simple, so that the calculation becomes possible. Then ($\boldsymbol{C|F1}$, …, $\boldsymbol{Fn}$) can be shown as in equation (6):

$$P\ (C|F_1, F_2, F_3, \ldots, F_n) = P(C) \overset{n}{\underset{i=1}{\mathrm{G}}} P(F_i|C) \tag{6}$$

Which can be described based on equation (6):

$$P(C \mid F) = P(F_1 \mid C)\, P(F_2 \mid C)\, P(F_3 \mid C) \ldots P(F_n \mid C) P(C) \tag{7}$$

The equation (7) is a model of the Naive Bayes theorem which will be used in the classification process.

4. Result and Discussion

The tests were carried out using the Emotion Lexicon and Naïve Bayes methods to produce a Confusion Matrix table. From the results of the table, the values of accuracy, precision, recall, and f-measure will be obtained. The data used for this study were 300 tweets with details of 210 tweets for training data and 90 tweets for test data.

Table 1. Tweet Test Classification Result

No	Tweet	Actual	Predict
1	looking forward to sitting out in the garden for lunch at a friend's house	Anticipation	Joy
2	... and finally, I have corrected colors on my screen Yeah	Joy	Anger
3	Just arrived in Paris. ADORABLE hotel room. Surprise bouquet of flowers on the desk for me	Anticipation	Joy
…	…	…	…
90	So glad I made it through work - with an extra hour too and my paycheck. Still waiting on the one I lost though	Anticipation	Anticipation

Furthermore, the process of calculating the percentage of accuracy shown in table 2 is the result of classification testing data using Emotion Lexicon and Naïve Bayes for each emotion.

Table 2. Evaluation Table in Confusion Matrix

Class	Precision	Recall	F-Measure
Anticipation	0.62	0.96	0.76
Joy	0.54	0.4	0.46
Sadness	0.8	0.33	0.47
Disgust	1	0.62	0.76
Surprise	1	0.33	0.5
Trust	0.5	0.83	0.625
Anger	1	0.375	0.54
Fear	0.75	0.6	0.66

The percentage of the calculation results of the accuracy of the results of the classification of documents with Emotion Lexicon without Naïve Bayes and Emotion Lexicon with Naïve Bayes, shown in Table 3.

Table 3. Comparison of Document Classification Test Results

Algorithm	Accuracy
Emotion Lexicon	45%
Emotion Lexicon + Naïve Bayes	65%

Analysis of the results of the study explained the comparison of the results of the document classification test with Emotion Lexicon without Naïve Bayes and Emotion Lexicon with Naïve Bayes, which is shown in Figure 2 of the system classification comparison graph.

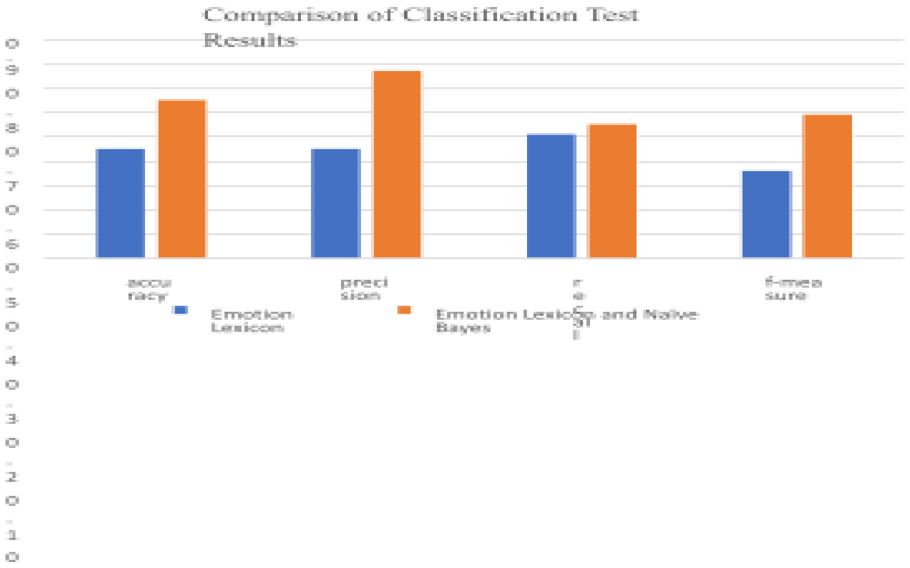

Fig. 2. Comparison of Classification Test Results

Figure 2 shows that all metrics, namely accuracy, precision, recall, f-measure on the Emotion Lexicon with Naïve Bayes are greater than the metric values on the Emotion Lexicon without Naïve Bayes. So, it can be concluded that the results of the classification of emotions using the Emotion Lexicon and Naïve Bayes are better than the results of the classification of emotions using the Emotion Lexicon without Naïve Bayes. This proves that Naïve Bayes has succeeded in reducing the problem of data that cannot be classified by the Emotion Lexicon alone.

5. Conclusion

Based on the research results, classification emotion in tweet using Emotion Lexicon yields 45% on accuracy while classification emotion using hybrid method, Emotion Lexicon and Naïve Bayes, yields 65% on accuracy. The result proves that using Naïve Bayes on Emotion Lexicon yields better classification result than without Naïve Bayes.

Further research is expected to apply other methods such as ensemble learning (Random Forest, Bagging, AdaBoost) which has advantages in data classification that has noise/outlier (low bias and high variance).

References

[1] M. S. Saputri, R. Mahendra, and M. Adriani, “Emotion Classification on Indonesian Twitter Dataset,” *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, no. November, pp. 90–95, 2019.

[2] H. Krishnan, “Emotion Detection of Tweets using Naïve Bayes Classifier,” *Int. J. Eng. Technol. Sci. Res.*, vol. 4, no. 11, pp. 457–462, 2017.

[3] I. M. D. Ardiada, M. Sudarma, and D. Giriantari, “Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour,” *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 55, 2019.

[4] E. Safaie *et al.*, “Emotion and Sentiment Analysis from Twitter Text,” no. 3, pp. 1–13, 2018.

[5] J. Bata, Suyoto, and Pranowo, “Leksikon Untuk Deteksi Emosi Dari Teks Bahasa Indonesia,” *Semin. Nas. Inform. 2015 (semnasIF 2015)*, vol. 2015, no. November, pp. 195–202, 2015.

[6] S. M. Mohammad and P. D. Turney, “Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon,” *CAAGET ’10 Proc. NAACL HLT 2010 Work. Comput. Approaches to Anal. Gener. Emot. Text*, no. June, pp. 26–34, 2010.

[7] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon,” *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[8] A. F. Anees, A. Shaikh, A. Shaikh, and S. Shaikh, “Survey Paper on Sentiment Analysis : Techniques and Challenges,” *EasyChair Prepr.*, vol. 2389, 2020.

[9] Bustami, “Penerapan Algoritma Naive Bayes,” *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[10] A. N. Rohman, E. Utami, and S. Raharjo, “Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing,” *Eksplora Inform.*, vol. 9, no. 1, pp. 70–76, 2019.

[11] S. Mujilahwati, “Pre-Processing Text Mining Pada Data Twitter,” *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[12] E. D. Liddy, “Natural Language Processing. In Encyclopedia of Library and Information Science,”

Marcel Decker, Inc., pp. 1–15, 2001.

[13] O. Somantri, “Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB),” *J. Telemat.*, vol. 12, no. 01, 2017.