ICIC Express Letters
Volume **13**, Number **8**, August **2019**

ICIC International ©2019 ISSN 1881-803X
pp. **705–709**

TEXT EMOTION DETECTION FOR ASIAN PARA GAMES TWEET IN INDONESIA

David David[1], Wayan Suparta[2], Agung Trisetyarso[2]
Bahtiar Saleh Abbas[3] and Chul Ho Kang[4]

[1]Computer Science Department, School of Computer Science
[2]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[3]Industrial Engineering Department, Faculty of Engineering
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ david01; atrisetyarso; bahtiars }@binus.edu; drwaynesparta@gmail.com

[4]Electronics and Communication Engineering Department
Kwangwoon University
20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea
chkang5136@kw.ac.kr

Received January 2019; accepted April 2019

Abstract. *Everyone is free to express their opinions through social media. This study aims to get the level of emotion towards the Asian Para Games event in Indonesia. This research begins by collecting data contained on Twitter using language filters and hashtags. Each data taken will be preprocessing, where the process was done by removing punctuation, hashtags and conversion of several emojis to text. By using the polyglot library, determining an emotion from each term can be done more easily because the library can recognize various languages, including Indonesian. This experiment achieves 55.41% accuracy. With so many non-standard word, abbreviation and unknown emojis, this research can still be developed by making several word changes into standard words, as well as converting all emojis or emoticons to the appropriate expression.*
Keywords: Emotion detection, Twitter, Hashtag, Asian Para Games, Indonesia

1. **Introduction.** In this era of developed communication technology, where everyone can freely express their opinions through social media. In Indonesia, freedom of speech makes it easy for everyone to express their opinions both verbally and in writing. One of the media that is often used to express opinions is Twitter, where one can create a status about everything. Every text that is made can express the emotional level of someone. This emotion can be used to make decision in business such as in luxury merchandise, emotion aspect as brand, sentiment analysis for business [1]. There are six types of emotions detected in text [2], namely joy or happiness as happy feeling or success to achieve goal, sadness as feeling sad or fail to achieve goal, anger as dissatisfaction, fear as threat for self, and surprise as daze or feeling not believe. By labeling and detecting each term, a document can produce emotion data, which can produce someone's emotional data when publishing the article on Twitter. The level of emotion on someone's writing can be used to find out someone's opinion about an event that occurs, whether many people feel happy about the event, or vice versa.

Polyglot is a tool that can be used for various things such as word separation or word grouping using various languages [3]. In this study, the language that will be focused is Indonesian because there are many people in Indonesia who use Twitter as one of the media to express their opinions freely. Determining the emotions of a tweet is not an easy

DOI: 10.24507/icicel.13.08.705

job, where a tweet can include a number of things such as hashtags, retweets, replies, emoticons, and conjunctions that are not related to writing.

In this paper, we proposed emotion detection method in Indonesian Language on Asian Para Games event, where in this event, the resulting tweets can vary because everyone can express their emotions when one athlete who competes experiences a win or loss in a match. This research focuses on the detection of one's emotions so that it can produce emotional levels from ongoing events.

2. **Literature Review.** The classification process of a text document in Indonesian includes pre-processing, extraction and selection features, weighting, learning and classification, and evaluation and can be seen in Figure 1 [2]. In the pre-processing process, folding, tokenizing, filtering, and stemming are used. Eliminating some affixes in each word can increase the classification results in which in Indonesian, some words are another form of the basic word [4]. The rules used to get the basic words are: particles (-lah, -kah, -tah, and -pun), possessive pronouns (-ku, -mu, and -nya), first level prefix (meng-, di-, ter- and ke-), second level prefix (per- and ber-), and suffix (-i, -kan, and -an). K-Nearest Neighbour (K-NN) is used one classification section and the query result is based on class majority.

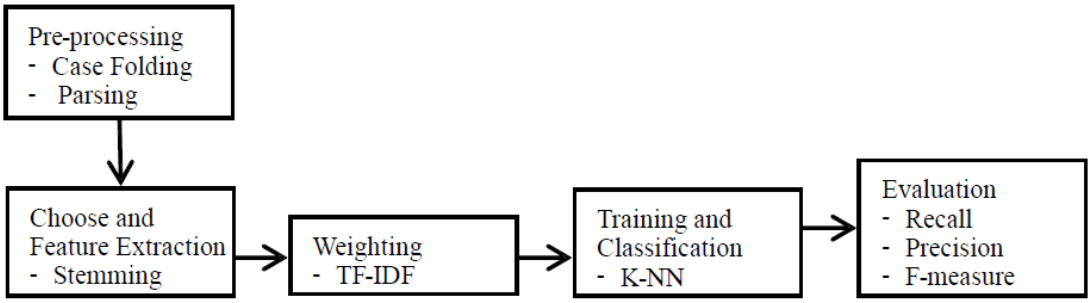

FIGURE 1. Classification process overview

In doing emotion detection, there are several challenges, namely collection of data, features choices, labeling of emotion, and machine learning classifier [1]. Problems with data collection include what data should be used as a feature, and how the data can always be relevant. Problems in feature choices include any indicators that can be used for emotional determination. Labeling of emotion includes any emotions that are included in a text, especially in some combinations of words. And the last is a machine learning classifier where the question that arises is how to determine the best classifier for classification. The methods used in emotion detection include keyword-based, where each sentiment is determined by each keyword. There is also a vector space model, where each word will be mapped using the weights of each term.

When it comes to detection emotion in other languages, the same method cannot be used because each language has a different word, structure, and meaning [5]. In Japanese language, compound sentences, double negation sentences, modifiers as adverb, and emoticons were combined to achieve the better result in recognizing each sentence. The research shows result and analysis for every method and each method get the different result. Compound sentences only achieve 60% accuracy, and the double negation only achieves 52%. However, with combining the two algorithms, the accuracy increases to 78%. On adverb processing experiment, the accuracy was 82.3%, and using emoticons words achieves only 36.4%.

SVM also can be used in sentiment analysis [6]. SVM classifies data into two classes. SVM uses $g(x)$ function as the discriminate function where the formula of $g(x)$ is as follows:

$$g(x) = w^T f(x) + b \tag{1}$$

where w is weight vector, $f(x)$ is nonlinear mapping from input space to high dimensional space, and b is bias. The variable w and b was learned automatically on learning process using the following formula:

$$Min1/2\, w^T w + C \sum_{i=1}^{N} ci \tag{2}$$

where N is the slack variables and C is the penalty coefficient. The result shows that the method works well and the algorithm was used in Bandung City Government.

3. **Research Method.** This research begins with collecting data in the form of tweets related to the event in Indonesian. The proposed method can be seen in Figure 2, where the process starts with data retrieval, then preprocessing and emotion detection are performed. The library used is a tweepy library, where the library is used to retrieve posts/tweets that are on Twitter. For the type of post taken is the hashtag #AsianParaGames where the hashtag is for the ongoing event in 2018. For the type of language taken only covers in Indonesian. For each post obtained through re-tweeting it will not be taken because it is included in data duplication. All data obtained will be stored in a json file which will be preprocessed.

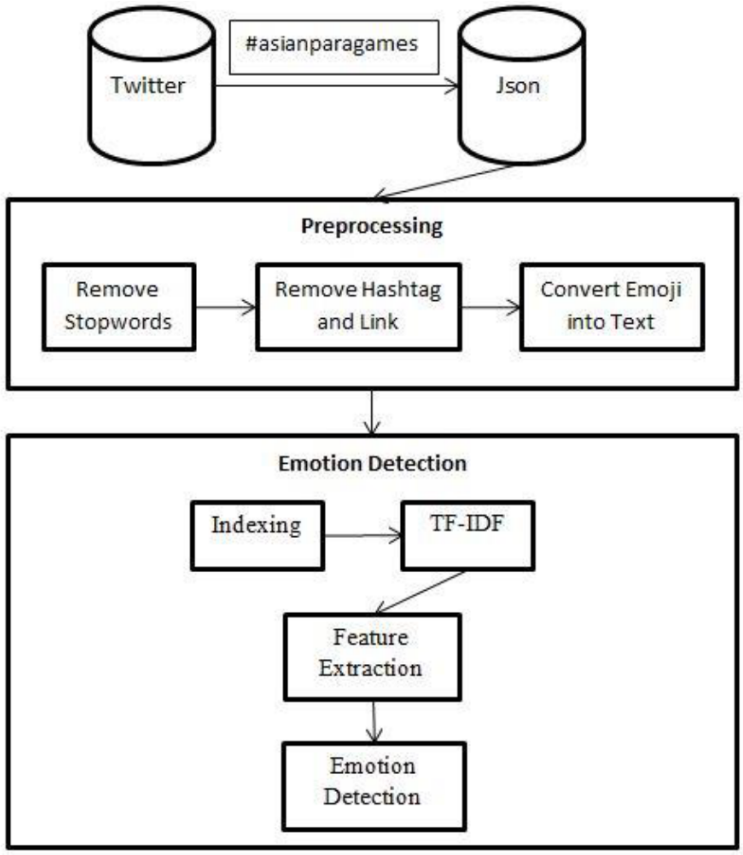

FIGURE 2. Proposed method

The preprocessing process begins by reading all the files in the json database that were created before. For each document that exists, all hashtag keywords can be removed. Hashtags are usually used to express the topic being discussed, so it is not needed in determining emotional levels. All punctuation except apostrophe was omitted because it does not affect the results in determining emotions. In a tweet someone usually has a link to direct the reader to the intended post. The link was also removed because it was not needed. At the end of the regular sentence there are several emojis to express emotions in the tweet. The emoji is converted into a text so that it can be recognized. Each emoji will

be manually converted first by replacing some symbols with the appropriate emotional keywords. There are repetitive emojis that will be eliminated so that word redundancy does not occur.

All documents taken will be inserted into the vector model so that classification can be carried out for each term. Each term will carry out a classification process to determine the type of emotion in that term. We use the polyglot library to get modules for term separation, as well as weighting for each term with TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF was used to get some important words in document. This method will rank all the words in document. High rank means the word was important and can be used as feature, while low rank can be discarded because it cannot be used as a feature. Polyglot libraries can be used for various languages, including Indonesian [7]. With different language support, polyglot also can be used in sentiment analysis to get the positive or negative aspect on each word or sentence.

For each emoji, the conversion is done by replacing emojis with keywords that match the emoji. Some emojis that are manually converted can be seen in Table 1.

TABLE 1. Emoji conversion

Expression	Emoji	Converted text
Joy	:), :-), :D, :p	Senang
Sadness	:(, :-(	Sedih
Fear	><, >.<	Takut
Angry	:X, :/	Marah
Surprise	:o, :-o	Kaget

4. **Result and Discussion.** After extracting all existing features, an emotion detection process will be carried out, where each feature will be adjusted to Russell's circumplex model of affect [8] (Figure 3). The value that best fits the model determines the emotions of the document. For happy emotions found if the value obtained is positive with a normal activation level, while sad emotions are the opposite of happy emotions with a low activation level. For emotions of anger and fear have similar parts/quadrants, where emotions are afraid of having a negative value and a higher level of activation compared to angry emotions.

In an experiment conducted on the hashtag #AsianParaGames, an experiment was conducted on 90 diverse tweets within 7 days. Every tweet, whether positive, negative or

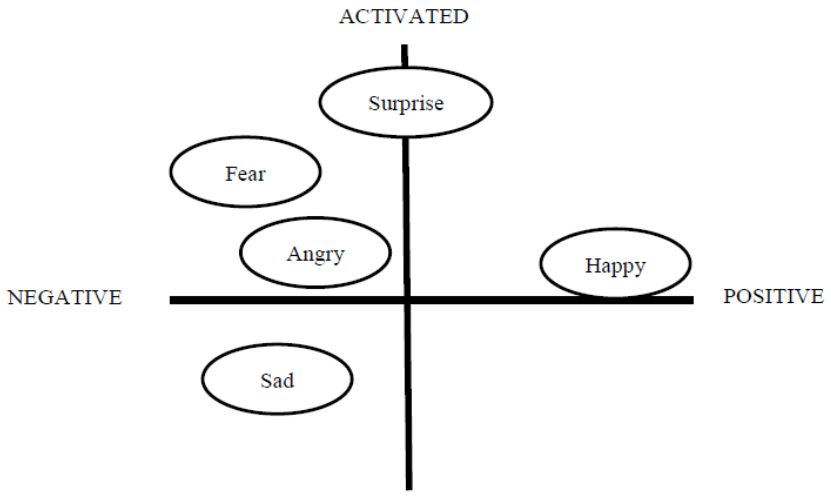

FIGURE 3. Circumplex model of affect based on Russell's theory

Table 2. Emotion detection result

Expression	Correct	Actual emotion	Percentage (%)
Joy	20	27	74.07
Sadness	3	7	42.85
Fear	3	5	60
Angry	7	18	38.88
Surprise	10	15	66.67
Others	9	18	50

neutral, will be mapped to the model and matched manually to get the accuracy of the document. The results of the accuracy obtained are 55.41% where the highest accuracy is happy emotions, which is equal to 74.07%. Detailed emotional detection results can be seen in Table 2.

From the table above, it can be seen that there are others, where the part is a document that is neutral or does not fall into 5 predetermined categories. If a document does not fall into these five categories, it will be included in the other category. The lowest accuracy results were obtained from anger and sad emotions, which amounted to 38.88% and 42.85%. Both emotions fall into the category of negative sentiment and are in similar areas, namely negative and activated regions.

5. **Conclusion.** The preprocessing process is very influential on the results achieved, where if the pre-processing is not done correctly, then the results obtained will not be appropriate because of the many words or sentences that are not related. From the results of the experiments conducted, there are various types of emojis and emoticons in existing tweets. The use of emojis as data for the determination of emotions is very useful, but because of the use of emojis and emoticons that are very diverse, there are still difficulties to convert all emojis according to the appropriate expression on the emoji. With the conversion of emojis or emoticons better, it is expected to increase the accuracy of emotion detection. The use of non-standard language also makes it difficult to determine the level of emotions in a sentence. There are various types of sentences that combine Indonesian with regional languages. This research can be applied to various parts such as games or chatbots. This research can still be developed by replacing several non-standard languages into standard, and more complete emoji conversion can improve results and accuracy in emotional detection.

REFERENCES

[1] V. Ramalingam, A. Pandian, A. Jaiswal and N. Bhatia, Emotion detection from text, *National Conference on Mathematical Techniques and Its Applications*, 2018.

[2] Arifin and K. E. Purnama, Classification of emotions in Indonesian texts using K-NN method, *International Journal of Information and Electronics Engineering*, vol.2, no.6, pp.899-903, 2012.

[3] Y. Chen and S. Skiena, Building sentiment lexicons for all major languages, *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.383-389, 2014.

[4] F. Z. Tala, *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*, 2003.

[5] R. Rzepka, K. Araki, M. Ptaszynski and J. Vallverdu, From words to emoticons: Deep emotion recognition in text and its wider implications, *International Journal of Computational Linguistics Research*, vol.9, no.1, pp.10-26, 2018.

[6] Asniar and B. R. Aditya, A framework for sentiment analysis implementation of Indonesian language tweet on Twitter, *International Conference on Computing and Applied Informatics*, 2016.

[7] R. Al-Rfou, B. Perozzi and S. Skiena, Polyglot: Distributed word representations for multilingual NLP, *Proc. of the 17th Conference on Computational Natural Language Learning*, pp.183-192, 2013.

[8] J. A. Russell, A circumplex model of affect, *Journal of Personality and Social Psychology*, vol.39, no.6, pp.1161-1178, 1980.