Sentiment Analysis using Weighted Emoticons and SentiWordNet for Indonesian Language

Nur Maulidiah Elfajr, Riyananto Sarno
Department of Informatics, Faculty of Information and Communication Technology
Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
nur.maulidiah14@mhs.if.its.ac.id, riyananto@if.its.ac.id

***Abstract*- The large number of internet users caused increasing the number of social media users. Twitter is one of social media that have a large number of users in Indonesia. As a social media, twitter allows users to share information via status in a tweet. Due to the limitations of the use of text is only 280 characters, emoticons are commonly used in tweet. Emoticon can explain the condition or feeling which is described in a text-shaped punctuation mark. This paper will focus on creating emoticon dictionary and weighting of an emoticon. Emoticon dictionary contains a list of 384 emoticons describing a variety of feelings and emotions. The used dataset contains Indonesian language tweets from twitter API. We tried to analyze sentiment on existing datasets with reference scores in SentiWordNet. Weighting emoticons done under the assumption that the emoticons have more effect in a sentence than ordinary words. After that, we classify the results into three classes, namely sentiment positive, negative and neutral. We compared the results between the emoticon-based algorithm and without considering emoticons algorithm. Accuracy obtained on the emoticon-based using algorithm is 0.74.**

Keywords- Emoticons; Indonesian; Sentiment Analysis; SentiWordNet.

I. INTRODUCTION

Technological developments in the world growing rapidly affects the increase internet users. According to the 2017 Index of Tetra Pak launched last year, there are about 132 million internet users in Indonesia. Half of that number or 40% are users of social media and can be categorized as an addict. We can find users of social media in everyday life. Social media allow users to share information, share a moment or share their condition and feelings. In Indonesia, Twitter and Facebook social media are the most widely used by the community.

Twitter is one of social media that have text limit only 280 characters. So, users who want to share information in the form of status or tweet must ensure the status that they post are not more than 280 characters.

Therefore, the use of acronyms and the use of emoticons is very common in the Twitter status. Acronyms is the use of abbreviations in the text. While emoticons can describe the feeling or emotion's user without detailed explanation.

In addition, the use of emoticons in tweet can describes directly emotion's user. So it can easily identify sentiment of the tweet.

II. RELATED WORK

There are many studies have been conducted to analyze the sentiment by considering the composition of Lexical [1]. SentiWordNet as one of the common used lexicon in several studies [2] [3]. The approach used in SentiWordNet produce better accuracy than others [4].

In some experiments, the existence of the acronyms and emoticons are not considered. So if there is a sentence that contains acronym and emoticon, the sentence would not be valued. On paper [5] they tried to insert emoticons in the calculation to determine the sentiment. [6] create a dictionary of acronyms and classification for emoticons. However, the calculation is still not optimal, the emoticon will be considered if the condition of sentence is neutral. So it can reduce the number of neutral sentiment.

In this paper we combine some of the elements above. We're trying to build a lexical use SentiWordNet and take into consideration the existence of emoticons on t Indonesian language tweets. We will use the emoticons in the calculation so it will have a sentiment score on SentiWordNet.

SentiWordNet has limitation of language that is only available in English. Paper [6] tried implemented SentiWordNet into multilingual.

In the paper [7], they try to develop them using Indonesian. They interpret data from Indonesian into English so that it can be analyzed using SentiWordNet. To translate a specific language into the Indonesian language, paper [8] uses Bing as a media translator launched by Microsoft. While this paper will use Google translate translator which is considered pretty good at translating words and sentences. Detail of this method has been depicted in a fig.1 the proposed method.

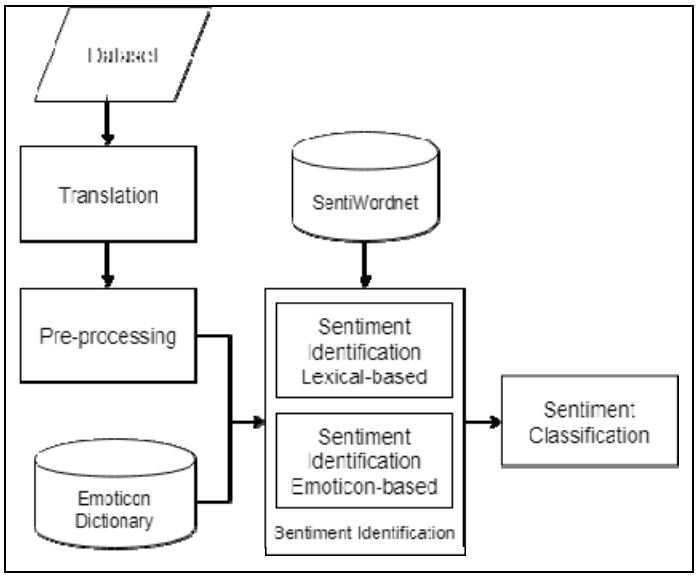

Fig. 1. Proposed Method

III. THE PROPOSED METHOD

A. Dataset

The dataset used data tweet Indonesian language. This data was taken with scrapling method using the twitter API called tweepy. Data tweet consists of id_user and tweet content. We labelling the data with the sentiment. The sentiments for each sentence are three classes, the which are positive, neutral, and negative. The dataset contains 100 sentences, the which are 32 positive class, 42 negative class, and 26 neutral class.

1. Example of positive sentence : Akhirnya selesai :))
2. Example of neutral sentence : Hari ini masuk sekolah :)
3. Example of negative sentence : Sedihnya ditinggal teman sekelas :(

B. Emoticon Dictionary

Emoticon is an interpretation of someone's feeling or emotion in text form icon. On each icon consists of ASCII and Unicode characters. Basically, emoticon consists of two types there are western and eastern.

1.) Western

Western emoticons type widely used in America and Europe. Western emoticons using all parts of the face are the eyes, the nose (not always) and mouth. emoticon western species are mostly written from left to right and then rotated 90 degrees in the direction of clockwise. Example emoticons:

:), :-(, : /, > :-), > _>, : '(

2.) Eastern

Western emoticons are widely used in Asia, especially Japan. Weastern emoticon type using eye as the main object. Eastern emoticons generally do not need rotation as in western

Example emoticons:

(^ _ ^;) (> _ <), ^ _ ^, (~ _ ~) , (* '∇`*)

We register all existing emoticon. according to Wikipedia sources [10] into list. We also consider it from paper [11] and paper [12]. This data consists of 384 eastern and western emoticons. From emoticons, we group similar emoticons depicting the moods or emotions together into one word. Thus forming an emoticon dictionary containing several words with each other's emotions. This grouping produce 46 emotions. Every emotion has its own list of members emoticon. For more details, see the Table. I

TABLE I. EMOTICON DICTIONARY

No.	Emotion	Emoticons
1.	Happy	':-)', ':)', ': -]', ':]', ': -3', ': 3', ': ->', ':>', '8-)', '8)', ': -}', ':}', ': o)', ': c)', ': ^)', '=]', '=)', '(? ^ o ^ ?) ',' (^ v ^) ',' (^ u ^) ',' (^? ^) ',' (^) o (^) ',' (^ O ^) ',' (^ o ^) ',' (^? ^) ',') ^ o ^ (',' (* ^? ^ *) ',' (????) '
2.	Laugh	'-D', ': D', '8D', '8D', 'xD', 'xD', 'XD', 'XD', '= D', '= 3' , 'B ^ D', '(* ^^) v', '(^^) v', '(^ _ ^) v', '(' - '*)', '(^ v ^)', '(^? ^)', '(·? ·)', '('? ')', '(???)'
3.	Excited	'\ (~ O ~) /', '\ (^ o ^) /', '\ (- o -) /', '? (^? ^)?', '? (^ O ^)?' '(* ^ 0 ^ *)'
4.	Cry	': \' (',': '- (', ':' (] ',' (T_T) ',' (; _;) ',' (; _; (; _ :) ',' (; O;) ',' (: _;) ',' (ToT) ',' (T? T) ',': _; ',' ;; ',': n; ',' ;; ' 'Q.Q', 'T.T', 'TnT', 'QQ', 'Q_Q'
5.	Sad	':-(' ':(', ': c', ': c', ': - <', '<', ': - [' ': [', ': - ||' ,

C. SentiWordNet

SentiWordNet a referral from WordNet [9]. WordNet is a dictionary word with different meanings according lexical structure for the purposes of text analysis. SentiWordNet contains a set of words from WordNet labeled sentiment. Type sentiment consists of 3 classes : positive, negative, and neutral. A sentiment score ranging from 0.0 to 1.0. The sum of the three is always worth 1.0.

D. Translation

Due to the limitations of the use of SentiWordNet, we have to translate the data from Indonesian into English. So it can be searched on SentiWordNet sentiment score. We use Google translate to translate it as a media translator. Google translate feels quite good at translating words. It has been able to consider the use of the word in a sentence.

In case of the text in the form of emoticons, emoticons that have been collected in one class also translated according to the emotion class.

E. Pre-processing

1.) Separation

This separation process is known as tokenization. The process started by separating the text into word tokens and symbols. This symbol is generally in the form of punctuation such as periods, commas and question marks. But in this case we try to identify symbols such as emoticons that exist in a sentence. We tried to separate between emoticons and said to then be used in the next process.

2.) POS Tagging

The process of POS (Part of Speech) tagging is the process of marking the word in the text (corpus) in accordance with a particular part of the sentence. POS Tagging aims to determine the exact meaning of a word in a sentence by considering its relationship with the closest words and linked in phrases, sentences, or paragraphs. This process is carried out after the separation of each sentence carried out. There are eight parts of speech roomates are nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections.

F. Sentiment Identification

In the previous pre-processing, a sentence will be identified whether the sentence contains emoticons. The results of that process will then proceed to the identification in accordance with certain criteria. Here is the explanation in the form of a diagram. Here is a brief explanation on a Fig. 2.

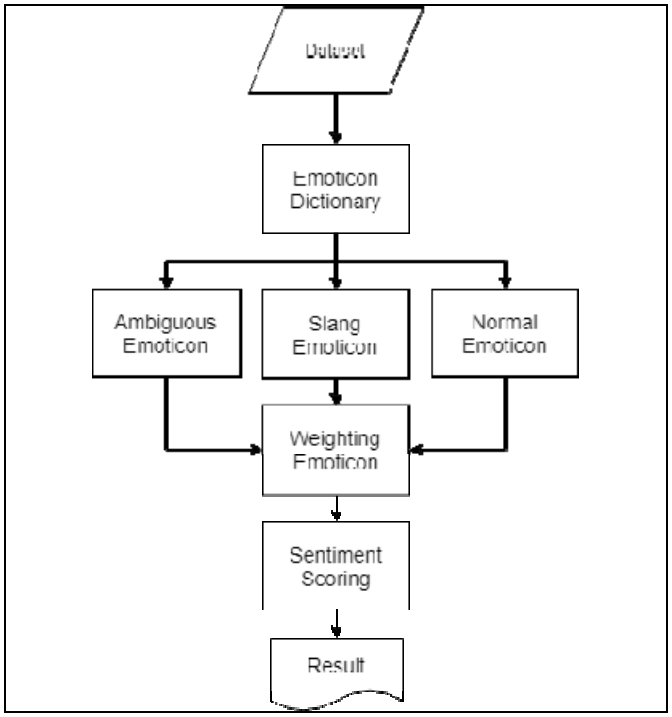

Fig. 2. Sentiment Identification Emoticon-based

1.) Text-based

For text that identified as ordinary word, the process will continue with the search sentiment score of each word in the Sentriwordnet. The end result of this process is the number of positive and negative scores in one sentence. This value is obtained from the sum of positive and negative scores of each word.

2.) Emoticon-based

For text that identified as emoticon, the process is continued in this function. Identified emoticon will be checked. There are three conditions that make a text indicated as emoticon. This condition refers to the emoticon dictionary that was created in the previous method.

a. Ambiguous Emoticon

Ambiguous emoticon is definition of mixed emotions or condition. Condition where the emotion was in the middle between two opposite emotions.

For example emoticon :"), which is a combination of emoticons happy and crying. In this context these emoticons is defined as a condition where the person is experiencing conditions of emotion but happy. This can be found in a tweet that described the touched condition and happy situation at once such as graduation, wedding moments , etc.

TABLE II. AMBIGUOUS EMOTICON SCORE

Ambiguous Emoticon	Emoticons	Positive Score	Negative Score
:')	:)	0.875	0.0
	:(	0.25	0.0
Result		0.56	0.0

Scores for emoticon of this type is the result of the merger of negative or positive score of both original emoticons. See Table II.

b. Slang emoticon

These emoticons commonly used the statuses on social media. In this case, emoticons written in a format that is not appropriate. As example is emoticon :)))), it is written with ‘)’ excess characters

The addition of ‘)’ could mean conditions 'very'. So that could mean significant emoticons very happy situation. as well as emoticons :(((. That means the condition where a person is in a very sad feeling.

TABLE III. SLANG EMOTICON SCORE

No.	Slang Emoticons	Positive Score	Negative Score
1	:)	0.875	0.0
	:))	0.975	0.0

	:)))...)	1	0.0
2	:(	0.125	0.75
	:((	0.125	0.65
	:(((... (	0.125	0.0

Scores for this emoticon type is addition and subtraction score for each character added. The value of each character added is 0.1. Addition or subtraction depends on the tendency of sentiment on the emoticon. See Table III above.

c. Normal emoticon

Normal emoticons are emoticons that don't have special characteristics and like emoticons in general. Scores for normal emoticon according to the emoticon dictionary that refers to a sentiment score on SentiWordNet.

3.) Weighting Emoticon

In this process, the text of which was identified as an emoticon will going to this process. In this process we perform weighting score in emoticons. We assume that emoticons have more affect on all the tweets and can describe an emotion than ordinary words. So we give more weight that value is double on each emoticon identified.

(1)

$$Positive\ Score = 2 * positive$$

(2)

$$Negative\ Score = 2 * negative$$

4.) Sentiment Scoring

After identifying the text into words and emoticons, the next step is to analyze sentiment score to each sentence m. Sentiment score is obtained from SentiWordNet containing positive and negative scores.

After the positive score of each word in the sentence summed. See Formula. 3. Similarly, the negative score. See Formula. 4 In addition, every word and emoticons that have a value of sentiment also summed. See Formula. 5. These results then that would be a divider in the equation Formula. 6.

(3)

$$S_{positive} = \sum_{i=1}^{n} Positive\ score_i$$

(4)

$$S_{negative} = \sum_{i=1}^{n} Negative\ score_i$$

(5)

$$S_{count} = \sum_{i=1}^{n} Word_i + Emoticon_i$$

The end result of the identification of this sentiment is the number of positive score of equation (3) added of negative scores from equation (4) divided by the number of words in equation (5) in a sentence. For details see the formula in equation. 6

(6)

$$S_{Result} = \frac{S_{positive} + S_{negative}}{S_{count}}$$

G. Sentiment Classification

After going through the process of identification sentiment, we get a score that is the final result to determine the classification of sentiment.

A sentence will have a positive sentiment if the final score is more than or equal to 0.66. The sentence which has the final score less than or equal to 0:34 will be categorized as negative sentiment. And the last, if final score is more than 0:34 and less than 0.66 will be categorized as neutral sentence.

(5)

$$S_{Sentences} \begin{cases} Positive\ ,if\ S_{Result} \geq 0.66 \\ Negative\ ,if\ S_{Result} \leq 0.34 \\ Neutral\ ,if\ S_{Result} > 0.34\ and\ S_{Result} < 0.66 \end{cases}$$

IV. RESULT

TABLE IV. SENTIMENT CLASSIFICATION RESULT

	Positive	**Neutral**	**Negative**
Accuracy	0.77	0.60	0.78
Average		0.74	

Table IV shows the results of the experiment for each type of dataset are positive, negative, and neutral. This result shows the accuracy for each type of sentiment. The average of the three types of 0.74.

The table below shows comparison results of accuracy, specify, precision and recall of algorithms that use emoticons-based and without using emoticons-based.

TABLE V. COMPARISON RESULT

	SentiWordNet	Emoticon SentiWordNet-based
Accuracy	0.59	0.74
Specify	0.58	0.73
Precision	0.59	0.72
Recall	0.59	0.74

In Table V it can be seen that the accuracy of the algorithm that uses emoticons SentiWordNet-based reach 0.74. This value is much larger than than the usual accuracy of the algorithm SentiWordNet which was only 0.59.

As well as on the value Specify, precision and recall. Third value derived from the results of SentiWordNet algorithm is much greater than the value of the ordinary SentiWordNet algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper, we try to identify the sentiment of a sentence by considering emoticons. We assume that an emoticon expresses more obvious emotions than words. So we give more weight to the emoticons. From the experiments that have been done, we get an accuracy of 0.74. Overall these results are much better than using SentiWordnet algorithm withouut emoticons-based, which only has an accuracy of 0.59.

For the study in future work is the use of WSD (Word Sense Disambiguation) [13] [14] in a word in order to know the exact meaning in a word so the sentiment score will give appropriate results.

REFERENCES

[1] M. Taboada, J. Brooke and M. Tofiloski, "Lexicon-based methods for sentiment analysis," Computational linguistics, pp. 267-307, 2011.

[2] Ohana B. and B. Tierney, "Sentiment classifiation of reviews using SentiWordNet," in 9th. IT & T Conference 2009.

[3] E.W. Pamungkas, R. Sarno, A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models, Telecommunication, Computing, Electronics and Control (TELKOMNIKA)", Vol 15, No 1, 2017. DOI: http://dx.doi.org/10.12928/telkomnika.v15i1.3176

[4] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC, vol. 10, pp. 2200-2204, 2010.

[5] W. Devid Haryalesmana, N. Azhari S, "peringkasan Esktraktif sentiment on Twitter Using TF-IDFdan Hybrid Cosine Similarity", Indonesian Journal of Computing and Cybernetics Systems, Vol.10, No.2, July 2016, pp. 207 ~ 218

[6] M. Edison and A. Aloysius "based Lexicon Acronyms and Emoticons of Sentiment Classification Analysis (SA) on Big Data", International Journal of Database Theory and Application, Vol.10, No.7 (2017), pp.41-54 (2017).

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University of Science, 1989.

[9] WordNet, A Lexical Database for English, "https://wordnet.princeton.edu/" [Accessed 28 April 2018]

[10] List of emoticons, "https://en.wikipedia.org/wiki/List_of_emoticons/" [Accessed 28 April 2018]

[11] F. H. Rachman, R. Sarno, C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus-Based Emotion", International Journal of Electrical and Computer Engineering (IJECE), Vol 8, No 3, pp. 1720-1730, Juni 2018. DOI: 10.11591/ijece.v8i3.ppi-720-1730.

[12] F. H. Rachman, R. Sarno, C. Fatichah, "CBE : Corpus-Based off Emotion for Emotion Detection in Text Document". The 3rd International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Pages: 331 – 335, 2016. DOI: 10.1109/ICITACEE.2016.7892466.

[13] B. S. Rintyarna., R. Sarno, "Adapted Weighted Graph for Word Sense disambiguation", The 4th International Conference on Information and Communication Technology (ICoICT), 2016. DOI: 10.1109 / ICoICT.2016.7571884.

[14] B. S. Rintyarna, R. Sarno, C. Fatichah, "Enhancing The Performance Of Sentiment Analysis Task On Product Reviews By Handling Both Local And Global Context". International Journal of Information and Decision Sciences. 2018 Vol. 11.