

Polarity classification tool for sentiment analysis in Malay language

Normi Sham Awang Abu Bakar, Ros Aziehan Rahmat, Umar Faruq Othman
Department of Computer Science, International Islamic University, Gombak, Selangor, Malaysia

Article Info

Article history:

Received May 30, 2019

Revised Aug 1, 2019

Accepted Aug 26, 2019

Keywords:

Classification
Malay sentiment analysis
Polarity
Sentiment lexicon
Twitter data

ABSTRACT

The popularity of the social media channels has increased the interest among researchers in the sentiment analysis (SA) area. One aspect of the SA research is the determination of the polarity of the comments in the social media, i.e. positive, negative, and neutral. However, there is a scarcity of Malay sentiment analysis tools because most of the work in the literature discuss the polarity classification tool in English. This paper presents the development of a polarity classification tool called Malay Polarity Classification Tool (MaCT). This tool is developed based on the AFINN sentiment lexicon for English language. We have attempted to translate each word in AFINN to its Malay equivalent and later, use the lexicon to collect the sentiment data from Twitter. The Twitter data are then classified into positive, negative, and neutral. For the validation purpose, we collect 400 positive tweets, 400 negative tweets, and 200 neutral tweets, and later, run the tweets through our sentiment lexicon and found 90% score for precision, recall and accuracy. Our main contribution in the research is the new AFINN translation for Malay language and also the classification of the sentiment data.

*Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Normi Sham Awang Abu Bakar,
Department of Computer Science, Faculty of ICT,
International Islamic University Malaysia,
Jalan Gombak, 53100 Kuala Lumpur, Malaysia.
Email: nsham@iium.edu.my

1. INTRODUCTION

Sentiment analysis (SA) is a process of computing, identifying and categorizing opinions expressed in words. It has been used to discover the writer's opinion towards a topic, product or issue. Sentiment analysis can be categorised as Opinion Mining (OM) process that uses Natural Language Processing (NLP) to extract and identify information within selected words [1] Developing the sentiment analysis for specific language will eventually benefit its native speakers in several fields such as business, politics and entertainment. In this research, sentiment analysis tool that focuses on analyzing Malay language texts will be developed. We aim to develop a Malay sentiment analysis tool that can identify the opinion of internet users, specifically on Twitter by classifying "tweets" according their polarity. Generally, polarity can be divided into three parts which are positive, negative and neutral [2]. The polarity of opinions can be divided into two groups; prior polarity and contextual polarity. The prior polarity is the polarity that represent the words in the lexicon, while contextual polarity is the polarity of the expression in a word [2]. Generally, there are various possible sources to collect the required data to conduct the sentiment analysis study. Among the most popular sources of SA data are the Social Network Sites (SNS) such as Twitter, Facebook, blogs and online forums. The similarities of these SNS are, it contains comments and opinions posted by people where emotions and opinions can be expressed through texts and emoticons. Besides determining the polarity of words, another dimension that is interesting to explore is the strength and

intensity of the word polarity. As such, one should be able to identify the intensity level of certain words, where different numerical scores will be given to the respective words to indicate the intensity or the strength of an opinion dimension expressed in the sentences [3].

The majority of work in investigating the polarity classification and strength are conducted for English language [2-8] and the work to perform such classification in Malay language is still scarce. To address this issue, we propose a Malay Polarity Classification Tool (MaCT), which has the capacity to determine the polarity and intensity of the opinions from our data source, Twitter. This paper presents the development of a polarity classification tool called Malay Polarity Classification Tool (MaCT). The tool is developed using AFINN lexicon which is widely used in English language where the polarity of the words are identified and the accuracy is analyzed using our classification method. We have successfully grouped 88.06% of the sentiment polarity of the tweets. For the classified tweets, we have managed to achieve 90% for the precision, recall and accuracy.

Background and Related Work

In general, there are three approaches for sentiment classification techniques: machine learning, lexicon based and hybrid approach. The Machine Learning Approach (ML) depends on the ML algorithm and applies the linguistic features, while the Lexicon-based Approach relies on sentiment lexicon, a collection of known and pre-compiled sentiment terms, and the Hybrid Approach is the combination of both approaches [1]. There exist a number of work in Sentiment Analysis, however, the majority of them are for English language. As such, the work of Bravo-Marquez et. al. suggest an approach to boost Twitter sentiment classification using different sentient dimensions, for example, opinion strength, emotion and polarity indicators, as meta-level features [3]. They found that the combination of sentiment dimensions provides significant improvement in Twitter sentiment classification tasks like polarity and subjectivity.

Another work by Agrawal et al built a new method for sentiment classification. Firstly, SentiWordNet is used to assign scores to a sentence and later, uses heuristics to handle context dependent sentiment expressions. They purport that their method shows significant improvement on movie review dataset over the baseline data [4]. In a similar line, Nasukawa et al. present a sentiment analysis approach to extract sentiments associated with polarities of positive (favourable) or negative (unfavourable) opinions toward a specific subject. They asserted that, in order to improve the sentiment analysis accuracy, it is important to properly identify the semantic relationships between the sentiment expressions and the subject [9]. From the perspective of sentiment analysis (SA) in Malay language, Handayani et al conducted a Systematic Literature Review (SLR) on the publications on SA, and the main focus is on the work done for Malay language [10]. They reported that, four out of ten papers on Malay SA utilize the Lexicon-based approach, while three papers use Machine Learning approach, two papers apply the Rule-based approach and one uses the Hybrid approach [10]. Furthermore, [11] tried to improve the classification performances by proposing a Malay SA classification model based on the semantic orientation and machine learning approaches. They collected 2,478 Malay sentiment-lexicon phrases and words assign synonyms to each word, and later, and the polarity is manually assigned with a score. A research conducted by Al-Moslmi et al highlights the effects of the common-used feature selection methods (Information Gain, Gini Index, and Chi-squared), and three machine learning classifiers (SVM, Naive Bayes, and K-nearest neighbor) for Malay sentiment classification. They conducted series of experiments using a Malay Opinion Corpus and found that the feature selection techniques improve the performance of the Malay sentiment-based classification [12]. The remainder of this paper is organized as follows: The next section highlights the methodology used in this research, while Section 3 discusses the results of the experiment. Finally, Section 4 concludes this paper and gives an overview of future work.

2. RESEARCH METHOD

Sentiment analysis research has been accelerated with the development of several lexical resources. Many sentiment classification tasks utilizes opinion words. Generally, positive opinion words are used to show desired states, while negative opinion words usually express undesired states. There are three main approaches to collect opinion word list. The first one is manual approach which is very time consuming, and usually combined with two automated methods, dictionary-based approach and corpus-based approach. Essentially, there are several classification approaches commonly used for sentiment analysis, i.e. OpinionFinder Lexicon [13], AFINN Lexicon [14], SentiWordNet Lexicon [15], SentiStrength Method [16], Sentiment140 Method [17], NRC Lexicon [18].

2.1. AFINN lexicon

Bradley and Lang coined a lexicon called Affective Norms for English Words lexicon (ANEW), which become the basis for AFINN lexicon [14]. The ANEW lexicon gives emotional ratings for vast number of English words. This lexicon include “valence”, which rates the psychological reaction of a person to a specific word. ANEW mostly covers formal language and does not include slang words commonly used in microblogging. Therefore, to address this gap, [19] created the AFINN lexicon, which covers the language in microblogging platforms. In this lexicon, slang, acronyms, and web jargon are also included. Positive words are scored from 1 to 5 and negative words from -1 to -5, and it is most suitable for strength estimation, and contains about 2477 words.

After deliberating on the possible lexicons to be used in our research, we decided to use AFINN lexicon because it is the most suitable to determine the polarity in microblogging and social media [20]. As such, we applied the words in AFINN lexicon to our data source, Twitter. However, the words contained in AFINN database are in English, whereas, our study is on the polarity of Malay words in Twitter. In order to enable the usage of AFINN for Malay language, we have translated each word in AFINN from English to Malay, including the polarity of the words. The words were compared to several Malay corpora hosted in the Malay Online Virtual Integrated Corpus (MOVIC) [21]. The translated Malay lexicon contains the exact translation of the English words in AFINN, and known as Malay Polarity Classification Tool (MaCT). The results of the translation is shown in Figure 1, where the English words in AFINN are translated to Malay. From Figure 1, we can see that, each word is given its polarity value, whether positive, negative and neutral. For example, the word “meninggalkan” is given a negative value of 2, which means the intensity of word is of Level 2, compared to the word “tidak hadir” which is given negative value of 1, which relatively means, is less intense than “meninggalkan”.

The sentiment analysis classification is done by collecting the Malay tweets on Twitter. For the purpose of collecting the tweets, Twitter4j [22], a Twitter API for JAVA, was used. This API provides many functionalities such as filtering tweets by time or by a hashtag. The processes of the polarity classification of the tweets in Twitter are depicted in Figure 2. From the figure, we found that when a word search is chosen, the system will extract the data from Twitter which later will be analysed to determine its polarity. Generally, the collected tweets contained noises that need to be removed, and several text pre-processing methods are applied to delete the noises [23]:

1. Tokenization
2. Removing numbers that do not express any emotions or attitudes.
3. Removing Punctuations

The next step is to extract the relevant tweets based on the search keywords, and later, the polarity is determined based on AFINN lexicon. Lastly, the polarity values are calculated using the polarity calculation algorithm.

```

1 {
2   "meninggalkan": -2,
3   "terbengkalai": -2,
4   "mengabaikan": -2,
5   "diculik": -2,
6   "penculikan": -2,
7   "penculikan": -2,
8   "jjjik": -3,
9   "membenci": -3,
10  "hina": -3,
11  "membenci": -3,
12  "kebolehan": 2,
13  "keupayaan": 2,
14  "menaiki": 1,
15  "tidak hadir": -1,
16  "membebaskan": 2,
17  "dibebaskan": 2,
18  "membebaskannya": 2,
19  "diserap": 1,
20  "penyalahgunaan": -3,
21  "didera": -3,
22  "penyalahgunaan": -3,
23  "kesat": -3,
24  "menerima": 1,
25  "diterima": 1,
26  "kemalangan": -2,
27  "sengaja": -2,
28  "mencapai": 2,
29  "dicapai": 2,
30  "tuduhan": -2,
31  "menuduh": -2,
32  "tertuduh": -2,
33  "menuduh": -2,

```

Figure 1. AFINN’s Malay translation

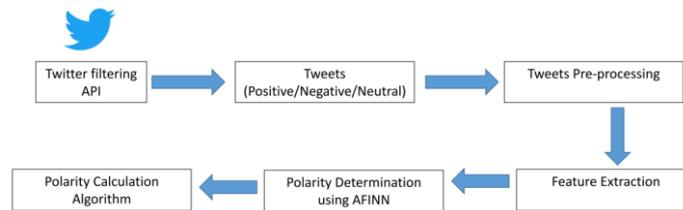


Figure 2. Polarity Classification process

2.2. Negation scoring strategy

Negation scoring strategy is very important feature for this sentiment analysis tool [24]. It helps the sentiment analysis tool to assign correct polarity when there are negation words in the text. Without negation scoring strategy words such as “tidak bagus” will be detected as positive as “bagus” is a token rated with positive value. The negation words in Malay language are shown in Figure 3. Generally, the orientation of the tweet polarity depends on the negation of words, where the meaning will totally change when negation is involved [25]. For example, the Tweet 1 is a positive tweet, in which the polarity is scope by the expression "bagus":

1. Kerja anda bagus. (Positive Tweet).

The tweet 2 use the form of the negation " tidak" is the negation of the Tweet 1:

2. Kerja anda tidak bagus. (Negative Tweet)

The processes of filtering the negation words are illustrated in Figure 4. First, the process starts with the identification of the word before the word token, if the negation word is found, a negative score will be given to the word token.

```

negators.json
1 {
2   "tidak": 1,
3   "tak": 1,
4   "bukan": 1,
5   "enggan": 1,
6   "taknak": 1,
7   "walaupun": 1
8 }
9 }
  
```

Figure 3. Malay negation words

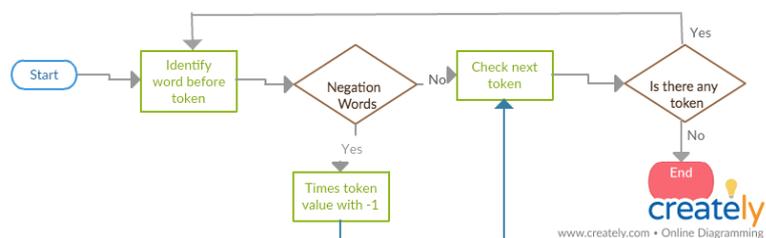


Figure 4. Processes to filter the negation words

The collection of words collected from Twitter was later tested using two steps: run half of the data as the training data and another half is treated as test data. The equations for the overall accuracy, precision and Recall are given below.

$$\text{Overall accuracy} = (TP+TN)/(TP+FP+TN+FN) \quad (1)$$

$$\text{Precision} = (TP)/(TP+FP) \quad (2)$$

$$\text{Recall} = (TP)/(TP+FN) \quad (3)$$

Where TP, TN, FP, and FN are number of true positives, true negatives, false positives, and false negatives, respectively.

3. RESULTS AND DISCUSSION

In general, our contribution in this paper are twofold: 1) The new AFINN version of Malay words, together with their polarity value (positive/negative/neutral) and 2) The polarity classification technique to identify the positive, negative or neutral tweets using several feature extraction techniques. Essentially, many researchers have used several statistical analysis techniques to analyze the collected data. The work by Iyer et al discusses the usage of deep neural network to analyse the data from a system that can detect,

classify and count vehicle that pass through the dedicated camera [26]. In a similar line, Al-Hagery attempted to extract the patterns from the dates product dataset using the machine learning technique, which is based on the association rules generation [27]. In addition, Sathyavikasini and Vijaya highlight the technique they use to distinguish the type of disease and they trained the classifiers using the supervised pattern learning technique [28]. Referring to Figure 1 in Section 2.1, the development of Malay version of AFINN has given the opportunity for Malay sentiment analysis researchers to study the polarity and also the strength of the words to further add more values of meanings in their work. Based on the literature, we first experimented using three feature extraction methods, based on the machine learning classifiers, SVM, Naive Bayes (NB) and K-nearest neighbour (KNN). The validation of our polarity classification method is done by randomly collecting 400 positive tweets, 400 negative tweets and 200 neutral tweets. Later, we gathered 100 respondents to label the words into positive and negative list and they are compared to our classification method. The next step is to compute the classification accuracy, precision and recall for our classification method using the formulas 1, 2 and 3 in Section 2.2. The results of the evaluation are shown in Table 1, where our classification method has successfully grouped 88.06% of the sentiment polarity of the tweets. For the classified tweets, we have managed to achieve 90% for the precision, recall and accuracy.

Table 1. Evaluation for Precision, Recall and Accuracy for our classification method

Percentage of words classified	Method Evaluation		
	Precision	Recall	Accuracy
88.06%	90%	90%	90%

From the results shown in Table 1, we can derive the polarity values from the list of words translated from AFINN database and this has solved the problem of having difficulty of finding the tool to work for the research in Malay language.

4. CONCLUSION

The development of sentiment lexicons in English language has been very useful for researchers over the years. A number of sentiment lexicons have been introduced, and one of them is AFINN. However, none of the lexicons are developed for Malay language. Thus, we have chosen to translate the words in AFINN and as a result, we got a comprehensive sentiment lexicon for Malay language. The main contribution of our research is the development of the Malay Polarity Classification Tool (MaCT) to classify sentiment lexicon for Malay language. The reason we choose AFINN is, it is the most suitable lexicon for social media data. The data for our experiment come from Twitter, and we use our method to classify the tweets to positive, negative and neutral. The validation results are excellent where our method scored 90% for precision, recall and accuracy. At the end of our research, we manage to meet the objectives of our study, which are, to find the polarity of the words in Twitter and perform the sentiment analysis based on the polarity. For future work, we plan to develop a more complete lexicon for Malay language so that we can obtain higher precision and accuracy score. The work in sentiment analysis in Malay language is still developing and we plan to continue the work to create a better sentiment lexicon tool.

REFERENCES

- [1] W. Medhat, A. Hassan and H. Korasky, "Sentiment analysis algorithms and applications: A survey". *Ain Shams Engineering Journal*, Vol. 5, pp 1093-1113, May 2014.
- [2] T. Wilson, J. Wiebe, and P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. Vancouver, British Columbia, Canada, 2005.
- [3] F. Bravo-Marquez, M. Mendoza, and B. Poblete. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13)*. ACM, New York, NY, USA, Article 2, 9 pages. 2013 DOI=<http://dx.doi.org/10.1145/2502069.2502071>.
- [4] S. Agrawal and T. J. Siddiqui. Using syntactic and contextual information for sentiment polarity analysis. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human (ICIS '09)*. ACM, New York, NY, USA, 620-623. 2009. DOI: <https://doi.org/10.1145/1655925.1656037>.
- [5] F. Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on Making Sense of Microposts*, May 2011.
- [6] E. Kouloumpis, T. Wilson, and J. Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

- [7] E., Cambria, R. Speer, C. Havasi, and A. Hussain SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *FLAIRS Conference*, pages 202–207, 2012.
- [8] P. Zhang, H. Zhongshi Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science* 41:4, 531-549. 2015.
- [9] T. Nasukawa and J.Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture (K-CAP '03)*. ACM, New York, NY, USA, 70-77. 2003 DOI=<http://dx.doi.org/10.1145/945645.945658>.
- [10] D. Handayani, N. S. Awang Abu Bakar, H. Yaacob and M. A. Abuzaraida, "Sentiment Analysis for Malay Language: Systematic Literature Review," *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, Kuala Lumpur, 2018, pp. 305-310. doi: 10.1109/ICT4M.2018.00063.
- [11] A. Al-Saffar, S., Awang, H., Tao, N., Omar, W. Al-Saiagh, and M. Al-bared. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS ONE*, 13(4), 1–18. (2018) <https://doi.org/10.1371/journal.pone.0194852>.
- [12] T. Al-Moslimi, S. Gaber, A. Al-Shabi, M. Albared, N. Omar. Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis. 1st ICRIL-International Conference on Innovation in Science and Technology (IICIST 2015). pp. 444-447.
- [13] T., Wilson, J., Wiebe, and P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. Vancouver, British Columbia, Canada, 2005.
- [14] M. M., Bradley, and P. J. Lang, Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings. *Technical Report C-1, the Center for Research in Psychophysiology* University of Florida, 2009.
- [15] S. Baccianella, A.Esuli, and F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010. Valletta, Malta.
- [16] M. Thelwall, K.Buckley, and G. Paltoglou, Sentiment strength detection for the social web. *JASIST* 63(1):163–173. 2012.
- [17] A. Go, R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. *Technical report Stanford University*, 2010.
- [18] S. M., Mohammad, and P. D. Turney, Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 2012.
- [19] F. Å. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on Making Sense of Microposts*, May 2011.
- [20] B., R. Becker, Kohavi, and D. Sommerfield. 2001. Visualizing the simple Bayesian classifier. In: *Information Visualization in Data Mining and Knowledge Discovery*, U. Fayyad, G. Grinstein, and A. Wierse, eds. San Francisco: Morgan Kaufmann Publishers.
- [21] N. S. Awang Abu Bakar, H. Yaacob, D. Handayani and M. A. Abuzaraida, "Malay Online Virtual Integrated Corpus (MOVIC): A Systematic Review," *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, Kuala Lumpur, 2018, pp. 243-248. doi: 10.1109/ICT4M.2018.00052.
- [22] S. Singh, T. N. Manjunanh and N. Aswini, A Study on Twitter4j Libraries for Data Acquisition from Tweets. *International Journal of Computer Applications*. pp. 29-32. 2016.
- [23] M.Youness, E. Mohammed, and B.Jamaa. 2018. Twitter Data Classification Using Big Data Technologies. In *Proceedings of the 2018 International Conference on Internet and e-Business (ICIEB '18)*. ACM, New York, NY, USA, 124-129. DOI: <https://doi.org/10.1145/3230348.3230368>.
- [24] A., Esuli, and F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation, 2006*.
- [25] G. A., Miller, R. Beckwith, C., Fellbaum, D., Gross, and K. Miller, Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.
- [26] P. R. Iyer, S. R. Iyer, R. Ramesh, MR. Anala, and K.N. Subramanya, Adaptive Real Time Traffic Prediction using Deep Neural Networks, *IAES International Journal of Artificial Intelligence (IJ-AI)* 8(2), pp. 107-119, 2019.
- [27] M. A. Al-Hagery, Extracting Hidden Pattern from Dates' product Data using a Machine Learning Technique, *IAES International Journal of Artificial Intelligence (IJ-AI)* 8(3), pp. 9-18, 2019.
- [28] K. Sathyavikasini and M. S. Vijaya, Identification of Rare Genetic Disorder from Single Nucleotide Variants using Supervised Learning Technique, *IAES International Journal of Artificial Intelligence (IJ-AI)*. 6(4), pp. 174-184, 2017.