An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis

Muhammad Fakhrur Razi Abu Bakar, Norisma Idris, Liyana Shuib
Faculty of Computer Science and IT
University of Malaya
Kuala Lumpur, Malaysia
fakhrurrazi0202@gmail.com, norisma@um.edu.my, liyanashuib@um.edu.my

***Abstract*—Nowadays, most Malaysians use social media such as Twitter to express their opinions toward any latest issues publicly. However, user individuality and creativity of language create huge volumes of noisy words which become unsuitable as dataset for any Natural Language Processing applications such as sentiment analysis due to the irregularity of the language featured. Thus, it is important to convert these noisy words into their standard forms. Currently, there are limited studies to normalize the noisy words for Malay language. Hence, the aim of this study is to propose an enhancement of Malay social media text normalization for lexicon-based sentiment analysis. This normalizer comprises six main modules: (1) advanced tokenization, (2) Malay/English token detection, (3) lexical rules, (4) noisy token replacement, (5) n-gram, and (6) detokenization. The evaluation has been conducted and the findings show that 83.55% achieved in Precision and 84.61% in Recall.**

Keywords—Malay social media text, Twitter, Noisy text, Lexicon-Based, Text Normalization

I. INTRODUCTION

Social Networking Service (SNS) like Twitter is increasingly popular as a medium for Malaysians to communicate or express their opinions overtly. The growing popularity of SNS attracts attention of researchers, especially in Sentiment Analysis (SA). Hybrid, lexicon-based, and machine learning were the three cardinal approaches for SA [1]. Lexicon-based approach is preferred when the English language is not being used because of its flexibility [1]. Based on our analysis, the noisy Malay text is not being handled completely at the pre-processing phase before being processed at the SA phase.

Twitter is considered as one of the top SNS in Malaysia [2]. According to [3], Malaysians write their respond on the SNS without following any abbreviations rules. Based on our analysis on 20k of Twitter messages from Malaysians, most of them use local dialects which give the same meaning but the spelling are totally different from the standard words (e.g. *awat* → *kenapa*), abbreviations where the patterns are not fixed or user generated text (e.g. *dgn, dgan* → *dengan*), local trend language where two similar words have totally different meanings compare to the meaning found in the dictionary (e.g. *payung(umbrella)* → *payung(belanja*), and mix languages which mostly consist of English and Malay (e.g. *i am so lapar*). According to [4], most of the Natural Language Processing (NLP) tools were predominantly trained on formal text. Hence, the existence of the noisy Malay texts become one of the major challenges in applying SA applications to public responds on SNS [5].

This paper proposed an enhancement of Malay text normalizer for lexicon-based SA. The remainder of this paper is organized as follows: The problem is identified from literature analysis and preliminary studies which is discussed in Section II. In Section III, the analysis on the 20k of Twitter data by Malaysian users is explained. Then the proposed architecture of the Malay Text Normalizer for lexicon-based SA in Section IV is presented. The evaluation of the normalizer is discussed in Section V. Finally, in Section VI we conclude the article and highlight the future works.

II. RELATED WORKS

Based on our analysis, there are 16 pre-processing techniques existed namely case folding, stop word removal, spelling correction, tokenization, stemming, lemmatization, intrinsic words removal, spam removal, characters removal, punctuation marks removal, non-words removal, social media tags removal, repeated characters removal, diacritics removal, symbols removal, and others. Case folding was used by several of the previous works such as [6], [7], [8], [1], and [9]. Case folding technique converts all the characters in a document into the same case, either all upper case or lower case. The second technique is stop words removal where it has been used by [6], [10], [11], [12], [13], [14], [1], [3], and [15] in their works. This technique removes words that carry unimportant meaning with respect to SA. Spelling correction is a technique to correct spelling errors due to abbreviations or typo. This technique was used by [10], [16], [14], and [1] in their works where it can be done by implementing spelling correction algorithm or creating a dictionary for converting the noisy word into its meaningful word. Another technique is tokenization which it helps to simplify the SA process. This technique was used by [10], [11], [13], [14], [1], [17], and [15] in their works. Stemming and lemmatization techniques both have been used by [1] in their study. For the basic removal techniques, [10] has removed intrinsic words in their work. Spam which largely be composed of insignificant words and emoticons were removed by [7] in their work. Other than that, [18] removed characters in their work. Another technique is punctuation mark removal which has been used by [10], [14], and [15] in their works. Non-words have been removed by [10] and [14] in their works. Besides than that, social media tags were removed by [12] and [14] in their works. Reference [12] has also removed repeated characters and diacritics in their works. The last basic removal technique which is symbols removal has been used by [7], [18], [1], and [3] in their works. Besides the above process, there are other pre-processing techniques used by the previous studies. Reference [19] cleaned their dataset which consisted of unwanted tags to obtain the words including abbreviations. For [9], their dataset was filtered by choosing tweets that were written in English, Malay or Indonesian language only. Other than that, they only chose the tweets which contained only one of the subjects being monitored in their project. To avoid huge numbers of neutral tweets, they filtered out any tweets that did not include any sentiment words which included in their sentiment lexicon. To avoid unreadable symbols and characters, the selected tweets must able to be encoded and decoded in UTF-8. In [6], they merged the word "tidak" with the next word to cater for negative words in the Malay language. Next, for [15], the duplicates that may modify the sentiment analysis's result were removed. In a study by [8], any words in a tweet that contained "www" or "https://", "#hashtag", and "@username" were converted to "URL", "hashtag", and "AT_USER" respectively. They also used another technique called trim.

To the best of authors' knowledge, there are only a few studies on normalizer for noisy Malay text. A study by [20] developed a dictionary-based system which also known as NoisyTerm to normalize Malaysian micro-texts. NoisyTerm has an ambiguity problem as highlighted by [5] in their study. Other than that, [20] removed directly the correct spelling in English or Malay language at the beginning of the process which will affect the latest Malay trend or noisy language actual meaning (e.g. *kite[saya/kita] → kite[layang-layang], payung[belanja] → payung [umbrella]*). In another study by [21] they proposed an approach for correcting the noisy Malay words without any interaction from the user. In addition, there is an ambiguity problem occur in their misspelled word dictionary which is the same as [20]. Finally, they removed symbols at the beginning of their approach which several of them will have a value in the latest informal patterns of the noisy Malay text. A corpus-driven analysis approach for normalizing Malay Twitter messages has been proposed by [5]. Reference [5] tagged any character excluding digit 2 and alphabetic characters as a proper noun which cannot be changed anymore at the upcoming steps. This will affect some of the latest informal patterns of the noisy text actual meaning and spelling (e.g. *$ → duit/money*). Other than that, the in-vocabulary words detection caused a loss of the latest Malay trend & noisy language actual meaning.

III. Analysis of Twitter Data

The Malay text normalizer is designed based on the Twitter analysis results. 20k Twitter messages by Malaysian users were analysed with guidance from three linguists from Faculty of Academic Studies of Malay, University of Malaya, Malaysia. The Twitter messages were collected based on location set to Malaysia using web scripting. The analysis process was done by reading all the Twitter messages one by one and classified the words into categories manually. Table I shows the summary of the Twitter analysis results where these datasets can be grouped into six categories. Based on the results, we found that most Malaysians used mix languages which consist of English and Malay in their tweets. Malay category refers to the whole tweet using a slang, standard and noisy Malay text. English category refers to the whole tweet using a slang, standard and noisy English text. Mix category refers to the mix of Malay and English languages, and some tweets consist of dialect in between. Dialect category refers to a minimum use of any Malay dialect in a standard Malay language tweet. Others refer to an identification name (@malaysiatravel) or any word starts with '#' symbol. The final category, decline, refers to other than the above mentioned.

The general analysis for the mix category is shown in table II where most of the messages used more Malay terms than English. Table III shows the analysis of formal and noisy words for the tweets that used mix languages. Table IV shows the analysis of formal and noisy words for the English category. The analysis of the formal and noisy words for the Malay category is presented in table V. Based on this analysis, we conclude that there is a high volume of noisy text occurred in Malay Twitter messages. The existence of the noisy text has become the biggest obstacle to applying text-mining. Thus, we propose an architecture of Malay text normalizer to clean the noisy text, before SA can be applied onto it.

TABLE I. Overall Analysis results

Category	Frequency	Percentage
Malay	5479	27.40%
English	4246	21.23%
Mix	6068	30.34%
Dialect	479	2.40%
Others	33	0.17%
Decline	3695	18.48%
Total	20000	100%

TABLE II. Frequency and Percentage of Dominant Language

Category	Frequency	Percentage
Malay (Dominant)	5110	84.21%
English (Dominant)	958	15.79%
Total	6068	100%

TABLE III. Mix Analysis Results

Category	Frequency	Percentage
Formal	463	7.63%
Noisy	5605	92.37%
Total	6068	100%

TABLE IV. ENGLISH ANALYSIS RESULTS

Category	Frequency	Percentage
Formal	3925	92.44%
Noisy	321	7.56%
Total	4246	100%

TABLE V. MALAY ANALYSIS RESULTS

Category	Frequency	Percentage
Formal	1292	23.58%
Noisy	4187	76.42%
Total	5479	100%

IV. PROPOSED TEXT NORMALIZATION

The overall process flow for SA is presented in Fig. 1 where in this work, we only focus on the Normalizer module which will be used to normalize the noisy Malay text found in the Twitter messages before proceeding to Polar Word Identification module. The overall process flow is from [3] where the data tokenization and data pre-processing module is replaced with our Normalizer module. Fig. 2 shows the architecture of the Normalizer module. In this overall process flow, tweets by Malaysian users will be used as the dataset. After going through the normalizer module, the polar words exist from the dataset will be identified by mapping them against the Malay sentiment lexicon. Lastly, the valence shifter will be handled, and the sentiment value will be identified.

Several modules and rules from this architecture were edited from [5], [20], and [21]. During the tokenization module process, all capital letters will be converted to small letters since capital letter is not having any orthographic value in noisy text [5]. Next, “\n”, “#”, and “http or link” will be removed after every extra blank have been converted to single blank. Any special character and mix special characters which do not have any meaning will also be removed. Any word which start with “@” or ended with “’” or “’s” will be tagged as Proper Noun Token (PNT). After that, any word that exist inside a Trend dictionary and Dialect dictionary will be tagged as Trend Dictionary Token (TDT) and Dialect Dictionary Token (DDT) respectively. There are another 19 Punctuation List (PL) tokens exist in this study for reducing the complexity and fasten the process of the proposed architecture. Finally, all the white spaces will be converted to new lines and every word will be tokenized.

After the tokenization process is completed, any token that exist in Global Malay/English dictionary will be tagged with In-Vocabulary Token (IVT). To fasten the process, only token that has been tagged with PL tokens or has not been tagged yet are allowed to go through this module. There is a condition where some of the noisy Malay tokens which have the correct spelling will have different meaning compare to the meaning found in the dictionary (e.g. *payung[noisy]* → *belanja, payung[dictionary]* → *umbrella*). Thus, this issue has been solved by tagging them with TDT before entering this module. There is also another condition for a tweet that used mix languages (English and noisy Malay text) where the noisy token “kite” can be referred to “saya/kita” or “layang-layang”. Based on the our Twitter analysis results, 84.21% of Malaysians used more noisy or standard Malay language in a single tweet compare to noisy or standard English language when they tweeted using mix languages. Hence, the token “kite” only will be tagged with IVT if the dominant language inside the whole tweet is English.

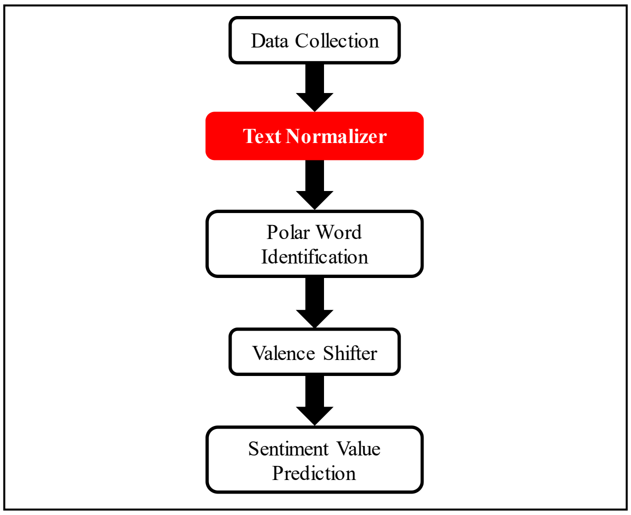

Figure 1. Overall Process Flow

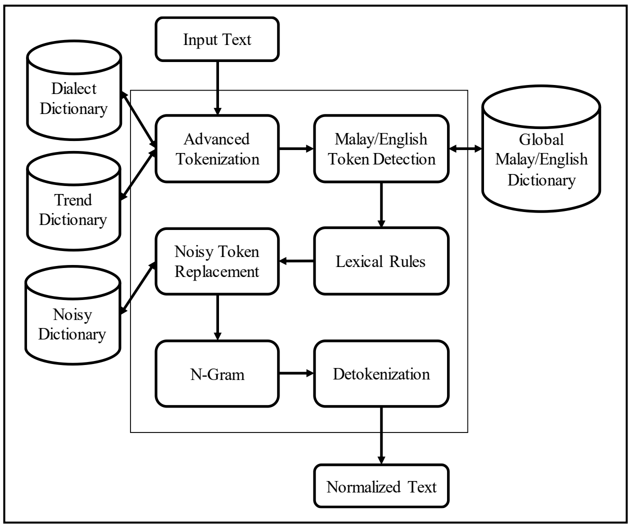

Figure 2. Enhancement of Malay Text Normalization

Lexical rules are the most complex and important module in this normalizer where it handles automatically most of the latest informal patterns of noisy Malay text as shown in Fig. 3. This module consists of 12 sub-modules namely Repeated Letter Elimination, Repetitive Words, Rules of X, Vowel Rules, Consonant Rules, Prefix Rules, CCV Rules, CC (1) Rules, CC (2) Rules, DPG Rules, RUYN Rules, and White List Rules. The lexical rules have two main functions namely tagging and process. During the tagging function, every token will go through the sub-modules one by one by following the order where any token that matches with one of the sub-modules will be tagged with related token. During the process function, the tagged token from the tagging function will be solved using related rules. During this process in most of the sub-

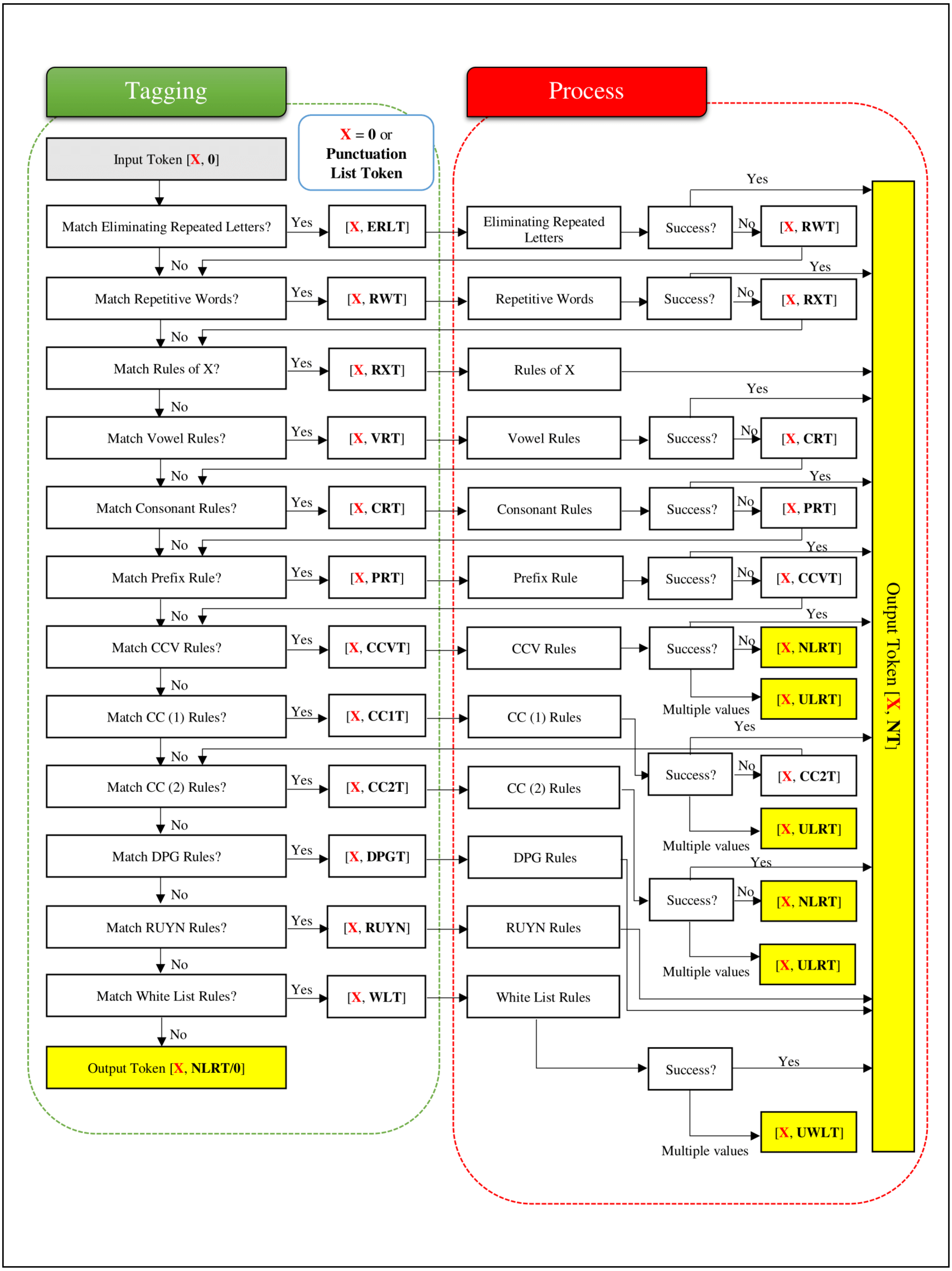

Figure 3. Lexical Rules Flow

modules, the normalized token will be checked automatically by using a dictionary which contains Malay words namely Global Malay dictionary. If the normalized token does not exist inside the Global Malay dictionary, it will be tagged with the next sub-module's token, otherwise the normalization of that token is considered as success. Other than that, stemming function also has been used to check the results in both Repeated Letter Elimination and Prefix Rules.

After the lexical rules process is completed, the token will next go to the noisy token replacement module. This module only allowed token which has been tag with Noisy Lexical Rules Token (NLRT) or has not been tag yet to fasten the process. If the token exists inside the noisy dictionary with more than one meaning, the token will be tagged with Noisy Dictionary Token (NDT), otherwise the token will be tagged with Normalized Token (NT). The N-gram module is provided to solve any of the ambiguity issues. Finally, all the tokens will be detokenized, and all the punctuation marks will be placed back by following the tagged PL tokens.

V. Evaluation

In evaluation stage, 100 Malaysians tweets have been used as a preliminary result in this study. Since 20k Twitter messages have been used to design the proposed architecture, another randomly 100 Malaysians Twitter messages have been collected for evaluation. The evaluation is conducted by comparing the normalized word produced by the proposed architecture with the normalized word produced by three linguists from Faculty of Academic Studies of Malay, University of Malaya, Malaysia. The result shows the proposed architecture produced a promising result with 83.55% Precision and 84.61% Recall.

VI. conclusion

This research paper explained briefly the conceptual theory of the proposed architecture since it is still an ongoing research. The aim of this research work is to produce a Malay text normalizer for lexicon-based SA. To develop the architecture of the normalizer, the dataset needs to be analysed thoroughly. Thus, we collected 20K Twitter data from Malaysian users and analysed the data to classify the words into different categories. From the results of the analysis, we proposed an architecture of the Malay Text Normalizer which comprises 4 dictionaries which are dialect dictionary, trend dictionary, global Malay/English dictionary and noisy dictionary. Besides than the 4 dictionaries, there are also 6 main processes proposed in the architecture which are advanced tokenization, Malay/English token detection, lexical rules, noisy token replacement, n-gram, and detokenization. To gauge the performance of the Malay Text Normalizer, the proposed architecture has been developed and evaluated using new Twitter dataset. The findings show that the proposed architecture achieved 83.55% in Precision and 84.61% in recall.

Acknowledgment

This research work is supported by University of Malaya (Faculty Research Grant: GPF007D-2018) and Cloud Connect Sdn. Bhd.

References

[1] M. H. A. Hijazi, L. Libin, R. Alfred, and F. Coenen, "Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language," *Proceeding - 2016 2nd Int. Conf. Sci. Inf. Technol. ICSITech 2016 Inf. Sci. Green Soc. Environ.*, pp. 356–361, 2017.

[2] N. A. Muhamad, M. A. Saloot, and N. Idris, "Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization," in *Journal of Physics: Conference Series*, 2017, vol. 806, no. 1.

[3] K. Chekima and R. Alfred, "Sentiment Analysis of Malay Social Media Text", *Computational Science and Technology*, vol. 488, pp. 205–219, 2018.

[4] T. Baldwin and L. Yunyao, "An In-depth Analysis of the Effect of Text Normalization in Social Media," *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 420–429, 2015.

[5] M. A. Saloot, N. Idris, and R. Mahmud, "An architecture for Malay Tweet normalization," *Inf. Process. Manag.*, vol. 50, no. 5, pp. 621–633, 2014.

[6] N. Samsudin, M. Puteh, A. Razak, and M. Zakree, "Immune Based Feature Selection for Opinion Mining," *Proc. World Congr. Eng.*, vol. III, pp. 1520–1525, 2013.

[7] N. F. Shamsudin, H. Basiron, Z. Saaya, A. F. N. Abdul Rahman, M. H. Zakaria, and N. Hassim, "Sentiment classification of unstructured data using lexical based techniques," *J. Teknol.*, vol. 77, no. 18, pp. 113–120, 2015.

[8] M. Naim, M. Ibrahim, M. Zaliman, and M. Yusoff, "Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception," *IEEE Conf. e-Learning, e-Management e-Services Twitter*, pp. 187–189, 2015.

[9] Y. F. Tan, H. S. Lam, A. Azlan, and W. K. Soo, "Sentiment analysis for telco popularity on twitter big data using a novel Malaysian dictionary," *Front. Artif. Intell. Appl.*, vol. 282, pp. 112–125, 2016.

[10] A. Alsaffar and N. Omar, "Study on feature selection and machine learning algorithms for Malay sentiment classification," *Conf. Proc. - 6th Int. Conf. Inf. Technol. Multimed. UNITEN Cultiv. Creat. Enabling Technol. Through Internet Things, ICIMU 2014*, pp. 270–275, 2015.

[11] A. Alsaffar and N. Omar, "Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis," *J. Comput. Sci.*, vol. 11, no. 4, pp. 639–644, 2015.

[12] T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar, "Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis," no. October, pp. 2–5, 2015.

[13] A. A. Sadanandan *et al.*, "Improving Accuracy in Sentiment Analysis for Malay Language," *Proceeding 4th Int. Conf. Artif. Intell. Comput. Sci.*, no. November, pp. 28–29, 2016.

[14] M. I. Eshak, R. Ahmad, and A. Sarlan, "A preliminary study on hybrid sentiment model for customer purchase intention analysis in socialcommerce," *2017 IEEE Conf. Big Data Anal. ICBDA 2017*, vol. 2018-Janua, pp. 61–66, 2018.

[15] 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*, vol. 13, no. 4, pp. 1–18, 2018.

[16] N. F. Shamsudin, H. Basiron, and Z. Sa'aya, "Lexical based sentiment analysis - Verb, adverb & negation," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 2, pp. 161–166, 2016.

[17] T. Al-Moslmi, N. Omar, M. Albared, and A. Alshabi, "Enhanced Malay sentiment analysis with an ensemble classification machine learning approach," *Journal of Engineering and Applied Sciences*, vol. 12, no. 20. pp. 5226–5232, 2017.

[18] S. Tiun, "Experiments on Malay short text classification," *Proc. 2017 6th Int. Conf. Electr. Eng. Informatics Sustain. Soc. Through Digit. Innov. ICEEI 2017*, vol. 2017-Novem, no. i, pp. 1–4, 2018.

[19] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment Analysis : Determining People's Emotions in Facebook," *Appl. Comput. Sci.*, vol. ISBN: 978-, pp. 111–116, 2014.

[20] N. Samsudin, M. Puteh, A. Razak, and M. Zakree, "Normalization of Common NoisyTerms in Malaysian Online Media," *Proc. Knowl. Manag. Int. Conf.*, no. July, pp. 515–520, 2012.

[21] S. B. Basri, R. Alfred, and C. K. On, "Automatic spell checker for Malay blog," *Proc. - 2012 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2012*, pp. 506–510, 2013.