Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach

Nur Imanina Zabha, Zakiah Ayop, Syarulnaziah Anawar, Erman Hamid, Zaheera Zainal Abidin
Center for Advanced Computing Technology (C-ACT),
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

Abstract—Sentiment analysis is a process of detecting and classifying sentiments into positive, negative or neutral. Most sentiment analysis research focus on English lexicon vocabularies. However, Malay is still under-resourced. Research of sentiment analysis in Malaysia social media is challenging due to mixed language usage of English and Malay. The objective of this study was to develop a cross-lingual sentiment analysis using lexicon based approach. Two lexicons of languages are combined in the system, then, the Twitter data were collected and the results were determined using graph. The results showed that the classifier was able to determine the sentiments. This study is significant for companies and governments to understand people's opinion on social network especially in Malay speaking regions.

Keywords—Opinion Mining; Sentiment Analysis; Lexicon-based Approach; Cross-lingual

I. INTRODUCTION

The amiable contextual definition of Big Data is dataset characterized by the 3Vs; Variety, Velocity and Volume that require particular Analytical Methods and Technology to transform into Value [1]. According to a study [2], Big Data can come in multiple forms including structured and non-structured data such as text files, genetic mappings, multimedia files and financial data. Three main types of data structures are: (1) Structured data which contain a defined data type, format, and structure (that is simple spreadsheets, traditional RDBMS, online analytical processing [OLAP] data cubes, CSV files, and even transaction data), and any data that reside in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets; (2) Semi-structured data is a part of structured data that does not adjust with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless enforce hierarchies of fields and records in the data and contains markers or other tags to classify semantic elements. Examples of semi-structured data are XML and JSON; (3) Unstructured data: Data that has no inherent structure, which may include text documents, PDFs, images, and video all of which require different techniques and tools to process and analyze.

Big data is where timeliness, diversity, distribution and/or scale will demand the use of new technical structures, analytics, and different tools to enable perceptivity that may unlock new origins that will be a business value. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data sample reveals and leverages business change. However, the larger the set of data, the more difficult it is to manage [3]. Businesses can benefit from analyzing larger and more intricate data sets that will probably require near-real time or real time capabilities that leads to a necessity for new tools, data architectures, and analytical methods.

A survey [4] reported that Malaysia's Internet penetration has reached 76.9% where owners own one or several accounts in Facebook, Instagram, YouTube, Twitter and etc. According to [5], social media eases the creation and sharing of career interests, ideas, opinion, information and other sorts of expression through virtual communities in different networks. Business companies are increasingly using social media monitoring tools to monitor, track, and analyze online conversations on the Web about their brand or products or about related topics of interests. A key component of social media monitoring tool is sentiment analysis.

Most sentiment analysis research focuses on English language lexicon. However, Malay or Bahasa Melayu is the major language spoken in Malaysia, Brunei, Indonesia, Singapore and Thailand. With 19.05 million social media users [6] in Malaysia, there is a need to provide source and tools available in the Malay language for sentiment analysis. Two major problems exist in Malay sentiment analysis [7]; limited number of standardized sentiment lexicon and (2) scarcity of sentiment classifier that is publicly available. The use of mixed languages, *Bahasa Rojak*, emoticons or emoji by social media users to express their opinions has increased the difficulty of classifying sentiments.

This paper is organized as follows. Section II discusses research background in Malay sentiment analysis. Section III introduces proposed system overview. Section IV outlines the discussion of results followed by experimental results in Section V. Section VI conclude the overall of this study.

II. RESEARCH BACKGROUND

Sentiment analysis (SA) is a process of detecting, deriving and classifying sentiments, opinions and attitudes expressed in texts concerning current issues, services, products, organizations, individuals, events, topics and their attributes [8]. SA derives opinion from the social media and formulates a negative or positive sentiment and based on which, sentiment classification is performed. There are two approaches for SA in

classifying the sentiments [9]: machine learning approach which is also known as a supervised approach and lexicon-based approach is also known as unsupervised approach.

A. *Machine Learning Approach*

Many researchers utilize machine learning approach to classify sentiments such as Naïve Bayes [10][11], k-NN [12], [13][14], and Support Vector Machine (SVM) [11]. In general, this approach collected large data and split into two; training data set and test data set. Training data set is used for classifier learning process and test data set is used to evaluate the classifier performance after the learning process is completed.

Sentiment analysis on movie review was proposed using the combination of Naïve Bayes (NB), Maximum Entropy (ME) and SVM [11]. The results show that SVM produces the best accuracy using unigram features. The authors believed that discourse analysis, focus ascertainment and co-reference identification could improve the accuracies of sentiment classification.

In Malay SA, [15], Malay text classifier on Malay movie review using SVM, NB and k-NN was proposed. The authors claimed that data collected are very 'noisy' compared to English reviews. Without feature selection and feature reduction technique, the performance of sentiment classification is low. The authors continued their work to improve the Malay SA [16]. They proposed Feature Selection based on Immune Network System (FS-INS) inspired from Artificial Immune System (AIS) in sentiment training process with three common classifiers; NB, k-NN and SVM. FS-INS works better than other feature selection techniques in filter category.

B. *Lexicon-based Approach*

A study [17] was the first to approach phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. A distinct approach on analyzing sentiment is by preparing a lexicon of negative and positive phrases. It is also known as the process of computationally categorizing and identifying opinions conveyed in phrases or texts, particularly in order to find out whether the writer's tendency towards a particular topic, product, etc. is neutral, positive or negative. Some other advances to anticipate the sentiments of words, expressions or documents are Natural Language Processing (NLP) [17]–[20], [21] and pattern-based [21]. A study [22] proposed Opinion Observer to compare various product features using language pattern mining.

In Malay SA, a Malay text classifier is developed using Lexicon-based approach from selected Facebook posts [23]. Text classifier was developed to study consumer opinion on low-cost carriers collected from Twitter posts [24]. Both studies [25], [24] built their own lexicon consisted of small number of sentiment words. To date, only a few research has developed Malay vocabulary text classifier using lexicon-based languages such as [7] and [26].

Machine learning approach produced high accuracy due to its high-quality training data. However, the performance drops when the same classifier is implemented in a different domain [27]. In contrast, lexicon based approach can be implemented in various domains, but slightly less accurate to machine learning approach.

C. *Proposed Cross-language Lexicon-based Approach*

Based on the previous work in the past sentiment analysis, many researchers perform sentiment analysis on data from social media like Facebook, Twitter, Pinterest, etc. and mostly focus on English language because most natural language toolkit have excellent datasets of English language with emoticons and slangs. There is only few research of sentiment analysis that use other languages besides English. Malay SA of [25] and [24] used small-scale lexicon classifier. This work proposed cross-lingual sentiment analysis on the existing English and Malay lexicons on twitter data where Malay is considered as a limited resource language. This research is closely related to Malay SA by [25] and [14]. However, study [25] classify the data into emotions and study [14] used Malay Review Corpus. In this study, sentiment analysis on Twitter users in Malaysia using large lexicon classifier available was experimented and the accuracy is evaluated. Twitter data are chosen in this study due to its source of data. According to a study, [28] on Twitter an average amount of tweets sent are around 6000, which in minutes matches up to 350,000 tweets are being sent, which sums up to 500 million tweets in 24 hours and overall of 200 billion tweets a year. The number of tweets is used by many organizations, institutions and companies as their informative source.

The English [22] and Malay [26] lexicons were being tested in terms of their category whether the words are positive or negative. For English lexicon, they take the frequent pairs of nouns or noun phrases that appear at the beginning of a sentiment as sample of opinions whether negative or positive points of view and the most frequent sample in a positive opinion will be considered a "positive word" and vice versa. For Malay lexicon, Wordnet is used to rate the words through their meaning and synonym and using Naïve Bayes technique to recheck their accuracy points given by Wordnet. This study proposed creating a database of positive and negative lexicon with a mixture of English and Malay and using R to determine if the sentiment was positive or negative.

III. METHODOLOGY

The proposed sentiment analysis system consists of three major phases as shown in Fig. 1.

A. *Preprocessing*

A search of a certain subject or hashtag will provide all tweets regarding that particular subject or hashtag. For example from the coding below will return a 1500 tweets of the 'subject' that was searched

```
tweet <- searchTwitter('subject', n=1500)
```

The data extracted for this research are Tweets from public accounts written by Malaysians on twitter. Raw tweets extracted from twitter is not suit for extracting features. Most tweets consist of special characters, messages, stop words, usernames, empty spaces, emoticons, time stamps, URL's, hash tags, abbreviations, etc. Thus, by cleaning the twitter data we are able to pre-process this data using R functions.
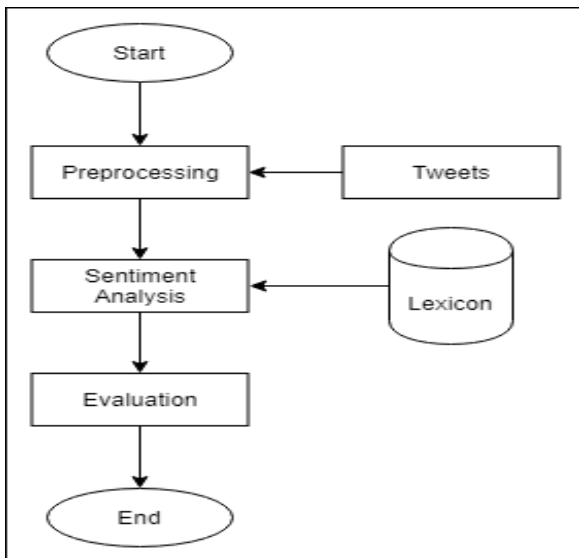
Fig. 1.   The Proposed Sentiment Analysis Process Flow.

B. Lexicon

There are two data that was used which combine the two most used language in Malaysia which is English and Malay. The English lexicon were generously provided by [22] for future researchers and data was regularly updated. The data was compiled by identifying the polarity of sentences. Frequent features are identified through association mining, and heuristic-guided pruning is applied. The technic of taking the frequent pairs of nouns or noun phrases that appear at the beginning of a sentiment will be a "representative sample" of opinions, both the negative and positive points of view will be covered in noun phrases or frequent nouns while others are placed in the infrequent features. Researchers [26] built a Malay sentiment lexicon based on Wordnet. All Malay words were already given their polarity where 1 represent a positive word while -1 represent a negative word whether the words are positive or negative (refer Fig. 2).

The English and Malay words are both grouped together for positive as well as the negative words. This combination will make analysis of twitter data easier as it can detect straight away where each word belongs to. Conjunction words like 'for', 'do', 'and', 'yet', 'or', 'ada','di', 'agar', 'ialah', 'kalau', etc, and Noun words like 'Doctor', 'food', and 'chicken', 'sempadan', 'wayang', 'buku', etc, are not in the Lexical Database and are considered as neutral (0).

C. Sentiment Analysis

This is where the sentiment score calculation is done. Based on its sentiment score, the tweets are categorized into two classes (positive and negative) by implementing the lexical based approach. Term Counting (TC) will be used for the calculation method as the positive or negative words are found on each document, and are used to determine the sentiment score [29]. This is a simple method of counting the positive and negative words found in the tweets. For example, if the sentence is

(A) "the movie was horrible and *mahal"*

Fig. 2.   A Sample of Malay Lexicon by [23].

The word horrible has (-1) polarity same goes to the word *mahal* which also has (-1) polarity. If the TC were to be applied the result would be

$$(0) + (0) + (0) + (-1) + (0) + (-1) = -2$$

(B) "*saya suka membaca,* it's my favourite hobby"

The word *suka* has (+1) polarity same goes to the word favourite which also has a (+1) polarity. If the TC were to be applied the result would be

$$(0) + (+1) + (0) + (0) + (0) + (+1) + (0) = +2$$

A classifier was built with packages from R to calculate the sentiment from each tweet to count the polarity of the tweet whether positive, negative or neutral. A bar plot will also be generated after the score is calculated.

D. Evaluation

The data from sentiment analysis will be analyzed manually by a native Malaysian speaker. Then, it will be compared to the proposed method results in terms of accuracy. Due to the standardize nature of the lexicon, the system are bound to make a few mistakes while analyzing the data because the actual mean of the shorten word or dialect that Malaysian used could not be recognize by the system.   For example:

(A) *"loning mu gi* interview *cari* experience*"*

The actual translation in English would be "Now you will attend the interview to gain experience".

(B) "I tk ske main game td"

Sentimen ▼	text	Human ▼
Positive	RT @rizkinvisa: Support bisnes laki-laki anak malaysia. p	Positive
Negative	RT @RTM_Malaysia: Program Lawatan Kerja YB Tuan @g	Neutral
Negative	RT @info_malay: Absence of clear, practical policies cast	Negative
Negative	RT @info_malay: Tersadai di tapak parkir #Malaysia http	Neutral
Positive	RT @info_malay: Jumlah kes rasuah yang dilaporkan di N	Positive
Positive	RT @SSK00123: Top 5 Overseas Markets for #Viswasam :1	Neutral
Negative	RT @info_malay: Guna aset kerajaan, jamu makan langg	Negative
Negative	info_malay: Visa problem for Indian transgender nothin	Negative
Positive	info_malay: Works ministry to get RM90 million extra fc	Neutral
Neutral	info_malay: ADUN bergelar Datuk direman #Malaysia ht	Negative
Neutral	info_malay: 'Sudah-sudahlah kami dibebankan...' #Mala	Negative
Negative	info_malay: SPR Beri Amaran Salah Guna Aset Kerajaan,	Negative
Neutral	urjoblessfan: [Malaysia GO](MAGAZINE) MAPS FEBRUAR	Neutral
Neutral	@rameshlaus: At the #Malaysia Box Office, for the 2nd v	Neutral
Positive	RT @anneroslan_: Senarai semak perbelanjaan untuk pe	Positive
Neutral	昨日はツインタワーにいってきたよ〜〜あの、こ	Neutral
Negative	mr_gadget: #penangbridge crash was a result of too muc	Negative
Neutral	RT @TERAGRE01964783: https://t.co/Uk0mCyHu3A#bloc	Neutral
Neutral	RT @newsfirst_ta: இலங்ச, ஊழல் ஆணைக்குழு	Neutral
Neutral	RT @AyaIsleemEn: ðµð‚#Palestine || ð²ð¾ #Malaysia ba	Negative
Neutral	Join our affiliate program to earn highest Recurring cash	Positive
Neutral	இலஞ்ச, ஊழல் ஆணைக்குழு அதிகாரிகளை	Neutral

Fig. 3. Example result for #Malaysia.

The full sentences that the system would be able to analyse are "I *tidak suka main* game *tadi*".

(C) "我喜欢鸡肉"

Malaysia is made up of various culture and ethnic but this research will focus on two most used languages in Malaysia which is Bahasa Melayu and English.

Fig. 3 below show the example result of search tweets on #Malaysia. Due to recent restrictions by Twitter only the latest tweets can be extracted. From 1000 tweets set in the settings, it returns 100 tweets. From the pre-processing, the cleaning data produced 57 clean data.

The categorization of the sentiments is assessed manually by a native Malaysian speaker. TABLE I. will be used to compare the performance of the method and the manual assessment by the speaker [7]. The three class contingency table represents the positive, negative and objective sentiments.

TABLE I. CONTINGENCY TABLE FOR SENTIMENT CATEGORIES TO COMPARE BETWEEN HUMAN ASSESSMENT AND PROPOSED METHODS

		Propose methods		
		Positive	**Objective**	**Negative**
Expert	**Positive**	Tp	Epo	Epn
	Objective	Eop	Too	Eon
	Negative	Enp	Eno	Tnn

T_p is the total data that were correctly categorized as true positive and Eon is when an objective data is wrongly categorized to negative category by the proposed method. From this table, the performance of the proposed method was calculated using four measurement, which are the accuracy, precision, recall and F1. For this study, the equations to calculate the performance in positive category are as follows:

$$accuracy = \frac{Tp + Too + Tnn}{total\ instances} \qquad (1)$$

$$precision = \frac{Tp}{Tp + Eop + Enp} \qquad (2)$$

$$recall = \frac{Tp}{Tp + Epo + Epn} \qquad (3)$$

$$F1 = \frac{2 * precision * recall}{precision * recall} \qquad (4)$$

IV. RESULT

TABLE II. shows the performance measurement results of the proposed method for the English and Malay tweets.

TABLE II. PERFORMANCE MEASUREMENT OF THE PROPOSED METHOD

Accuracy	0.554
Precision	0.478
Recall	0.733
F1	0.578

It can be observed that the F-Score from the result is quite adequate because usually the score are in a range of 0.0 - 1.0, where 1.0 would be a perfect system. The system has a high score on recall which shows that the system can accurately determine neutrality of the tweets. The accuracy and precision of this system is not very high, this is probably due to the usage of slangs, shorten words and dialects which the system were unable to process. In addition, the number of clean data collected is quite low which affect the confident level of the result.

V. EXAMPLE ANALYSIS

The proposed system can still be used to search for controversial topics just to understand the public perception or opinion toward that topic.
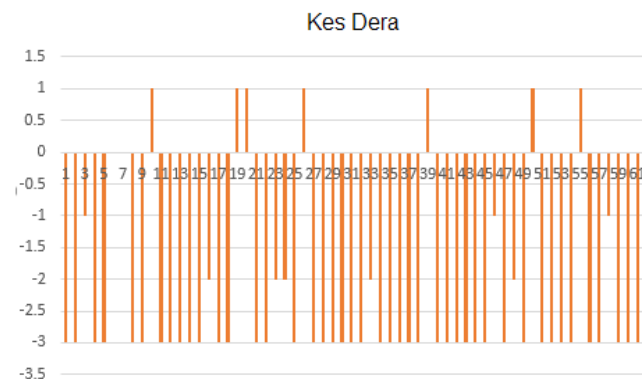

Fig. 4. Result of 'Kes Dera'.

A few controversial topics were being analyzed to view the accuracy of the system. Abuse has always been a serious issue. With this system, the term "Kes Dera" was searched. From Fig. 4, it can be seen that the public was angry and posted various negative remarks about the issue.

Another hot issue was service tax. A comparison between Goods and Services Tax (GST) and the new tax Sales and Services Tax (SST) was searched using this system. Fig. 5 and Fig. 6 show the latest public perception or opinion on both tax systems.

In Fig. 5, there were mixed feelings towards GST taxing system but if compared to Fig. 6, there were a lot of negative opinions on SST than GST.
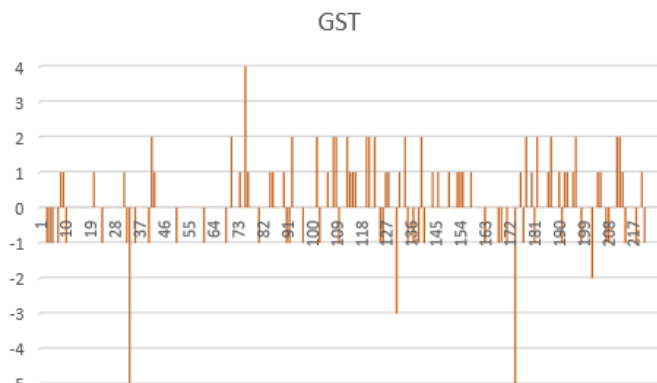

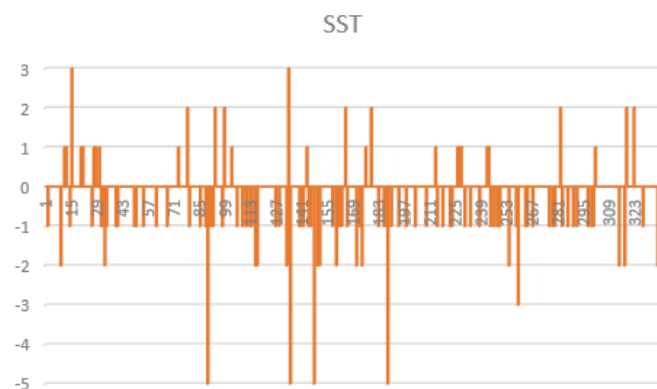Fig. 5.   Result of GST.

Fig. 6.   Result of SST.

Fig. 7.   Result of LGBT.

Recently, LGBT has been a debatable issue that impacted the Malaysians. Malaysia is a country where Islam is an official religion, but other religions are allowed to follow their own religious practices. LGBT is an ongoing movement to fight for its belief. The movement is not widely accepted by local communities globally but are more popular in the United States. Various public opinion on LGBT in Malaysia can be viewed (refer Fig. 7).

Fig. 7 shows a balanced amount of negative and positive opinion. Each person is entitled to their own opinion on any matter. This system is only created to mine positive, negative or neutral comments, opinion or messages from twitter users.

VI. CONCLUSION

In conclusion, a cross-lingual sentiment analysis has been developed to analyze tweeter data from Malaysian social media users. The system has successfully analyzed the two largely used languages in Malaysia; English and Malay. Four performance parameter has been measured. The results show that it has high recall which shows that the result can be trusted. However, the proposed method has average accuracy due to the usage of slangs, shorten words and dialects. This model can still analyze what the public feel about a certain topic that revolves in Malaysia. This could greatly help a certain brand or company to understand what a person or a customer comment about their product. However, this system cannot analyze dialect and detect shorten words that is widely used in the social media. For future work, the lexicon will include words in short form, slangs, different dialects and stop words.

REFERENCES

[1] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," Libr. Rev., vol. 65, no. 3, pp. 122–135, 2016.

[2] D. Dietrich, B. Heller, and B. Yang, "Data Science & Big Data Analytics: Discovering," Anal. Vis. Present. Data, 2015.

[3] P. Russom, "Big Data Analytics. TWDI best practices report," 2012.

[4] Malaysian Communications and Multimedia Commission, "Internet Users Survey 2017," 2017.

[5] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," Bus. Horiz., vol. 54, no. 3, pp. 241–251, 2011.

[6] "Number of social network users in Malaysia 2022 | Statistic." [Online]. Available: https://www.statista.com/statistics/489233/number-of-social-network-users-in-malaysia/. [Accessed: 26-Dec-2018].

[7] N. A. Nasharuddin, M. T. Abdullah, A. Azman, and R. A. Kadir, "English and Malay cross-lingual sentiment lexicon acquisition and analysis," in International Conference on Information Science and Applications, 2017, pp. 467–475.

[8] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Syst., vol. 89, pp. 14–46, 2015.

[9] E. Kasmuri and H. Basiron, "Subjectivity analysis in opinion mining - A systematic literature review," Int. J. Adv. Soft Comput. its Appl., vol. 9, no. 3, pp. 132–159, 2017.

[10] A. Lin, "Improved Twitter Sentiment Analysis Using Naive Bayes and Custom Language Model," arXiv Prepr. arXiv1711.11081, 2017.

[11] B. Pang, L. Lee, S. Vaithyanathan, and S. Jose, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[12] R. Alfred, W. W. Yee, Y. Lim, and J. H. Obit, "Factors affecting sentiment prediction of malay news headlines using machine learning approaches," Commun. Comput. Inf. Sci., vol. 652, pp. 289–299, 2016.

[13] 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.

[14] 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. 1–12, 2018.

[15] N. Samsudin, M. Puteh, and A. R. Hamdan, "Bess or xbest: Mining the Malaysian online reviews," in Data Mining and Optimization (DMO), 2011 3rd Conference on, 2011, no. May 2016, pp. 38–43.

[16] N. Samsudin, M. Puteh, A. R. Hamdan, M. Zakree, A. Nazri, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, 2013, vol. 3, pp. 3–5.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005, pp. 347–354.

[18] K. Hiroshi, N. Tetsuya, and W. Hideo, "Deeper sentiment analysis using machine translation technology," in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 494.

[19] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70–77.

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in Proceedings of the 19th international conference on World wide web, 2010, pp. 641–650.

[21] M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad, "Named Entity Recognition using Fuzzy C-Means Clustering Method for Malay Textual Data Analysis," J. Telecommun. Electron. Comput. Eng., vol. 10, no. 2–7, pp. 121–126, 2018.

[22] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in Proceedings of the 14th international conference on World Wide Web, 2005, pp. 342–351.

[23] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment analysis: determining people′s emotions in Facebook," Univ. Teknol. MARA, Malaysia, pp. 111–116, 2013.

[24] B. Y. Liau, P. P. Tan, B. Y. Liau, P. P. Tan, B. Yee Liau, and P. Pei Tan, "Gaining customer knowledge in low cost airlines through text mining," Ind. Manag. Data Syst., vol. 114, no. 9, pp. 1344–1359, 2014.

[25] N. A. M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z. Abiden, "Sentiment analysis: determining people's emotions in Facebook," Univ. Teknol. MARA, Malaysia, pp. 111–116, 2013.

[26] N. U. R. S. Alexander and N. Omar, "Generating a Malay Sentiment Lexicon Based on Wordnet," J. Teknol. Mklm. dan Multimed. Asia-Pasifik, vol. 6, no. 1, pp. 126–140, 2017.

[27] B. Lu and B. K. Tsou, "Combining a large sentiment lexicon and machine learning for subjectivity classification," 2010 Int. Conf. Mach. Learn. Cybern. ICMLC 2010, vol. 6, pp. 3311–3316, 2010.

[28] "Twitter Usage Statistics - Internet Live Stats." [Online]. Available: http://www.internetlivestats.com/twitter-statistics/. [Accessed: 22-Nov-2018].

[29] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," Sch. Comput. 9th. IT T Conf., p. 13, 2009.