Bias Aware Lexicon-Based Sentiment Analysis of Malay Dialect on Social Media Data: A Study on The Sabah Language

Mohd Hanafi Ahmad Hijazi, Lyndia Libin
and Rayner Alfred
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Sabah, Malaysia
hanafi@ums.edu.my, lyndia.88@gmail.com, ralfred@ums.edu.my

Frans Coenen
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
coenen@liverpool.ac.uk

***Abstract*—Sentiment Analysis (SA) has gained its popularity over the years for the benefit it brings to the development of economy, sociology and politic. SA enables observation, experiment, and quantification of emotions of the public toward a particular issue. However, there is not much SA done with respect to the Malay Language, especially in the context of the Malay dialects used in social media. The research presented in this paper aims to perform SA on one of the derivatives of the Malay language, namely Sabah Language. The Sabah Language, unlike many other languages, does not have a fixed spelling and, when used in an unstructured form as in the case of social media, poses particular difficulties for SA. This paper takes a lexicon-based approach to SA of the Sabah Language as used on social media. For the investigation, the corpuses selected were Facebook posts and tweets written in the Sabah language, 443 posts and tweets in total. Each was manually annotated as positive, negative or neutral by three annotators. As Sabah Language is a derivative of Malay language, the words used in Sabah Language contains most of Malay words. That is why, in Sentiment-Lexicon (SL) construction process, opinion-bearing Malay SL is retrieved, modified and expanded to build Sabah SL. Three different methods of assigning scores to the words in SL (opinion-bearing words) were employed during SL construction: (i) Simple PSA, (ii) Simple PSA with Switch Negation (PSA-SN) and (iii) Strength-based PSA. In this paper, pre-processing phase that includes spellchecker and shortform corrector is also implemented to reduce distinct word to be analyzed for SA. In classification phase, two classification methods, simple and bias aware classifications, were used to classify the posts. Experiments are conducted to show the effect of SL modification and expansion, the effect of pre-processing as well as the effect of bias-aware classification to the SA performed. Results show the highest accuracy of 85.10% was achieved using bias-aware classification with the modified and expanded SL, scores are assigned using Simple PSA and the pre-processed text.**

***Keywords*—bias-aware; lexicon-based; Sabah language; sentiment analysis; social media**

I. INTRODUCTION

Sentiment Analysis (SA) is one of the fields of study within the domain of Natural Language Processing (NLP) where peoples thoughts, feelings and opinions are analysed for its sentiment [1], [2]. The end-result of SA is the polarities of the analysed item, whether the sentence or document or word indicates positive, negative or neutral sentiment [1], [2]. This polarity serves as a feedback for the purpose of business/product/service improvement, tracking of political issues and socio studies, as well as serving as a benchmark for consumer decision-making. Automation of the sentiment analysis process maeans that large document collections can be easily processed. Researchers have proposed various algorithms and approached to analyse sentiment with respect to the text data found in social media, especially in Twitter and Facebook [2]. There are three fundamental approaches to SA. First is the lexicon-based approach, where two techniques can be employed; (i) Dictionary based and (ii) Corpus Based. The second approach utilises machine learning techniques, either supervised, semi-supervised or unsupervised [3]. The third approach is a hybrid approach that combines the lexicon-based and machine learning approaches. The approach presented in the paper is using lexicon-based approach where dictionary based approach is employed.

According to a survey [4], 7.9% of Malaysia Internet User are from the state of Sabah, which is ranked the third highest Internet User in Malaysia after Selangor and Johor [4]. This number of Internet users is proportional to Sabah as the state with the third highest population in Malaysia. One out of every 14 Malaysian Internet user, that use the internet for social networking, is from Sabah. Therefore it would be socially and commercially useful to apply SA to the Sabah Language as used on social media. However, there has been a very limited amount of research that has focused on SA in the Malay language, let alone the Sabah language (a derivative of Malay language) as most of the work found in the literature has focused on English. To the best of our knowledge, no work had been done directed at SA with respect to the Sabah language. The particular challenge here is that Sabah has no agreed spelling, a challenge compounded by the unstructured format of texts found on social media.

This paper proposed a lexicon based approach to SA to predict the sentiment of social media data written in the Sabah

978-1-5090-1721-8/16/$31.00 ©2016 IEEE

Language. The main contributions are:

- A framework for lexicon based SA of the Sabah language as used on social media. The lexicon is constructed by retrieving Malay SL, then modify and expand Malay SL by adding synonyms and antonyms found in the corpus; whether the words are in Sabah or Malay language.
- A lexicon based approach SA that employs dictionary based technique. During SL construction, three types of Polarity Score Assignations (PSA) are used to assign scores to the words in SL: Simple PSA, Simple PSA with Switch-Negation (PSA-SN) and Strength-based PSA, each coupled with bias-aware classification.

The organisation of this paper is as follows: In Section 2, related work concerning SA is presented. In Section 3 the overall framework for the proposed lexicon based social media SA for the Sabah Language is presented. Some experimental results are presented in Section 4. Section 5 concludes the paper with a summary of the main findings and some directions for future work.

II. Related Work

This section provides an overview of related work to the research presented in this paper. First machine learning and lexicon based SA are described and compared. Second, some previous work concerning bias-aware classification is presented.

A. *Machine Learning Approach*

Many researchers have used machine learning to do sentiment classification [5]. There are many machine Learning approaches that may be adopted, including: Naïve Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbour, Negative Selection and Maximum Entropy [1], [6]. Researchers favour machine learning algorithm such as NB and SVM to perform SA [5]. To do the classification, two different collections of documents are needed, a training collection and the test collection [1]. The training collection is used to train the classifier in order to learn the patterns contained in the training data; the test collection is used to test the resulting classification model. These machine learning approaches are usually trained using different feature sets including unigrams and bi-grams [1]. The advantages of using machine learning to perform text classification is that they perform well in specific and bounded domains [2]. However, they require a pre-labelled training set (preferably a large amount of data), so that patterns can be effectively learnt to distinguish positive, neutral and negative messages.

B. *Lexicon Based Approach*

Lexicon-based approach rely on detecting words that carry sentiment load [1], [2], [7]. The detection of these words requires a Sentiment Lexicon where words are annotated with their sentiment polarity [2]. The text grammar is then analysed and sentiment scores assigned to the sentences or words; according to the lexicon or dictionary [2]. Thus the quality of the lexicon plays an important role in determine the accuracy of the SA classification [6], [8]. An SL typically contains sentiment words, also known as opinion words, polar words and opinion-bearing words [9], along with their polarity scores. These words are usually divide into two class; positive words and negative words. Sentiment scores, extracted from an SL, can be used in a number of ways. One common approach is to calculate the overall polarity of the text from the sentiment score assigned to each word according to the SL. The simplest approach to the calculate of the overall polarity is a simple aggregate-and-average method [6] where the sum of all sentiment score of a sentence is calculated and the average is used as polarity label for that sentence.

To date, most available SLs are directed at the English language, very few are in other languages. Commonly used English language SL resources are WordNet [10] and SentiWordNet [11]. Most researchers when performing SA on languages other than English will resort to either build their lexicon manually [2] or by translating English lexicon to their preferred language [6]. A Malay SL has been manually created [12]. An alternative method of building a Malay Language SL might be to use WordNet or SentiWordNet to collect sentiment words, and then translate these words to their Malay equivalents [1].

SentiWordNet has been used to build a Malay lexicon [13]. However the SA accuracy was low due the polarity scores given by SentiWordNet do not reflect the true polarity of the equivalent Malay word. Another problem was that many Malay words had multiple English language equivalents with different sentiment scores.

C. *Comparison Between Lexicon-Based Approach and Machine Learning*

From the foregoing the differences between lexicon-based approach and machine learning approach are clear. In the context of languages other than English, lexicon-based approaches are preferred because of their flexibility. Another reason to employing the lexicon-based approach over the machine learnig approach is that in many cases it is the only feasible option when there is no training data available; this was the case with respect to a reported SA system based on the Spanish language as used in Facebook posts to give one example [2]. It has been shown that machine learning and lexicon-based approaches displays different accuracies when tested in specific domains and across domain. While machine learning shows better accuracy in specific and bounded domain, it shows poor accuracy when employed cross-domain [2], [6]. This is because machine learning has a tendency of over-fitting the training data set [2]. For specific domains, domain specific lexicons are required whose generation tends to be resource intensive.

D. *Bias-Aware Classification*

We wish to classify texts as featuring either positive or negative sentiment. The lexicon-based method can have a bias one way or the other [7], although recent research has revealed that lexicon-based approaches tend to have a bias towards positive sentiment [6], [7].This is due to the human tendency to favour positive statements. A method called Bias-Aware Thresholding (BAT) [7] has been proposed to minimize the sentiment bias rate, and consequently enhance the accuracy of lexicon-based SA. BAT uses cost-sensitive learning where the prediction threshold is adjusted to reduce sentiment error bias. This threshold is a controlling threshold where a positive value threshold will penalise positive prediction and a negative value threshold will penalise negative prediction using (1):

$$C_{BAT(d)} = \begin{cases} +, & \text{if}(S_A^+(d) - S_A^-(d)) > t \\ -, & \text{if}(S_A^+(d) - S_A^-(d)) < t \end{cases} \tag{1}$$

where $C_{BAT(d)}$ denotes classification of a document, d, using the BAT method, $S_A^+(d)$ denotes the sum of the positive scores for document d, $S_A^-(d)$ denotes the sum of the negative scores for document d and t is the controlling threshold. Polarity Bias Rate (PBR) is used to measure the bias rate within a document and is calculated using (2).

$$PBR = \frac{(FP - FN)}{N} \tag{2}$$

where FP denotes the number of false positives in document d and FN denotes the number of false negatives in a document d, while N is the total number of documents.

III. Bias-Aware Lexicon Based Sabah Language Social Media Data Sentiment Analysis

The framework for the proposed approach is shown in Fig. 1. It consists of four phases: (i) data collection, (ii) SL building (which include polarity score assignation), (iii) data pre-processing and (iv) classification. Each phases is described in further detail below.

1) *Data Collection*. The dataset was retrieved manually from social media; Facebook and Twitter. There were a total of 443 corpus; tweets and Facebook posts kept in textfile. The data was then annotated by three annotators.
2) *SL Building*. To build Sabah Language SL, a Malay SL is obtained from the Multilingual Sentiment in Data Science Labs2, serves as a foundation to Sabah SL that is built. As mentioned earlier in the introduction of the paper, as our lexicon approach is dictionary based, Sabah SL is built by adding corresponding Sabah Language words that is found in the corpus that are either synonyms or acronyms to the Malay-opinion words contained in the SL that is retrieved. The original SL, which has two textfile, positive Malay words and negative Malay words has duplicated entry whereby the same word occurs in both textfiles. There are also some English words included in both the textfiles as well. Modification of the SL is made by omitting the duplicate entries and by translating the English word first to Malay, before removing the word from the textfile. The SL is also further expanded with Malay acronyms and synonyms found in the corpus to enhance its coverage.

- *Polarity Score Assignation (PSA).* As already noted the assignation of polarity scores utilises three different methods as follows:
  a) **Simple PSA**. In this approach, feature that carries positive sentiment load was assigned a value of +1 and feature that carries negative sentiment load was assigned a value of -1.
  b) **PSA-SN**. In this approach, feature polarity score was determined according to the following rules:
  R1: if $WN_{AND}W^+$, then $P_S = -1 * P_S^+(W^+)$
  R2: if $WN_{AND}W^-$, then $P_S = -1 * P_S^-(W^-)$
  R3: $WN_{AND}W^0$, then $P_S = P_S^-(WN)$

  where: (i) P_S is the polarity score, (ii) WN is Negation Word, (iii) W^+ is a positive word, (iv) W^- is a negative word, (v) W^0 is a word that does not have a polarity score if it occurs after a Negation Word, (vi) $P_s^+(W^+)$ is the polarity score for a positive word W^+, and (vii) $P_s^-(W^-)$ is the polarity score for a negative word W^-. For example, two adjacent words, with its polarity value indicated in a bracket, “tidak (-1) suka (+1)” will have a PSA value of -1 according to R1.
  c) **Strength-based PSA**. In this approach, all features have scores determined by SentiStrength [14], ranging from -5 (most negative) to +5 (most positive). However, during the setup, it is found that many Sabah language-opinion bearing words have no polarity scores (valued zero) when computed using SentiStrength. This problem is called as cross-language limitation [15], where the source language word has no corresponding word, and hence no sentiment value in target language. As a solution, SentiStrength modification is performed where zero-valued opinion-bearing words is assigned a default value of -1 (for negative polarity words) and +1 (for positive polarity words) to preserve their polarity.
- *Data Pre-processing*. Pre-processing is important to reduce noise from the dataset and to make the dataset more manageable. Spellcorrector is employed to correct only misspelled Malay words as Sabah Language has no agreed spelling and short-form corrector is used to reduce number of distinct words in the corpus to be considered. The details of pre-processing phase is as follows:
  a) **Tokenization and Case Folding**. Every word in the corpus was identified from its associated whitespace. The word charcters were then transformed to lower case and the word tokenized. In this manner a Bag-Of-Words (BOW) representation was produced.
  b) **Symbol Elimination**. What were considered to

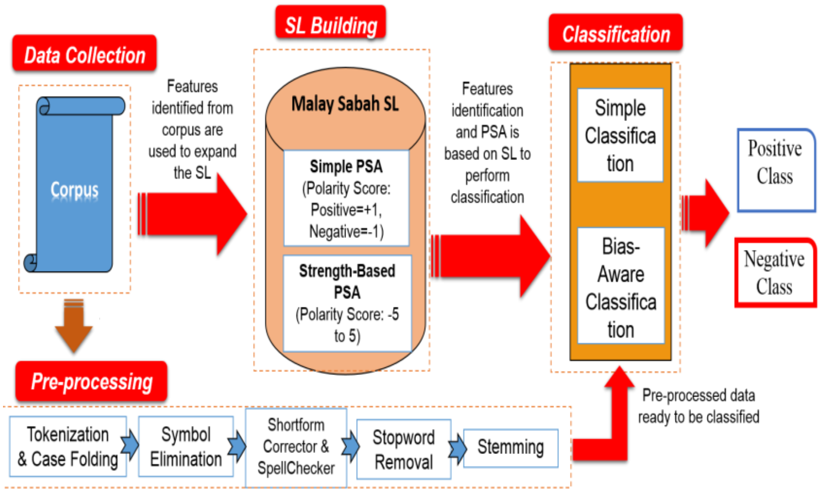
Data Collection
Corpus
Features identified from corpus are used to expand the SL
SL Building
Malay Sabah SL
Simple PSA (Polarity Score: Positive=+1, Negative=-1)
Strength-Based PSA (Polarity Score: -5 to 5)
Features identification and PSA is based on SL to perform classification
Classification
Simple Classifica tion
Bias-Aware Classifica tion
Positive Class
Negative Class
Pre-processing
Tokenization & Case Folding
Symbol Elimination
Shortform Corrector & SpellChecker
Stopword Removal
Stemming
Pre-processed data ready to be classified

Fig. 1: The proposed framework

be illegal characters, such as ‘!‘, ‘@‘, ‘#‘, ‘$‘, ‘%‘, ‘ˆ‘, ‘&‘, ‘‘, ‘(‘ and ‘)‘, and numbers were eliminated. Elimination is important to ensure the data is cleaned from unnecessary symbols and meaningless numbers that do not contribute to the classification at the later stage.

c) **Lemmatisation and Spellchecker**. Lemmatisation was performed that include transformation, repetition letter removal and short forms correction.

d) **Stopword Removal**. Stopword removal involves removal of all common Sabah Malay words that carry no significant meaning with respect to SA. However, a few stopwords words such as ‘aduh‘, ‘syabas‘, ‘adoh‘, ‘aduhai‘, ‘tidak‘ and ‘tetapi‘ were excluded from the list in this work as these words carry sentiment load.

e) **Stemming**. The Malay Stemming algorithm with Background Knowledge [16] was used for this purpose.

- *Classification*. After data has been pre-processed, the next step is to perform SA classification where the posts or tweets will be classified as either positive, negative or neutral. In this paper, bias-aware classification is used. For comparison purpose, simple classification was also performed.

a) **Simple Classification**. In this method the scores are summed together [9] and the classification is based on the following rules:
R4: if $\sum P_s > 0$ then positive
R5: if $\sum P_s \leq 0$ then negative

where $\sum P_s$ is sum of all positive and negative score in a post.

b) **Bias-Aware Classification**. The length of the sentence may affects classification as the sentiment label fluctuates over the sentence length [14]. As a solution, in this paper, each post polarity score was normalised over its length. The normalised value represents the intensity of the positive or negative sentiment of the post. A threshold, t, as described in Section 2, was defined and used to classify a post as either positive or negative according to the following rules [1]:
R6: if $(\sum P_s^+ - \sum P_s^-) > t$ then positive
R7: if $(\sum P_s^+ - \sum P_s^-) \leq t$ then negative

where $\sum P_s^+$ is the normalised sum of the positive scores of a tweet or facebook post and $\sum P_s^-$ is the normalised sum of the negative scores of the same tweet or facebook post. The advantage of this technique is that it will reduces bias to zero while maintaining prediction accuracy. Here, $t \in R$ is a controlling threshold. A positive value of t will penalise positive polarity predictions (by increasing the cost of false positive errors) while a negative value of t will penalise negative polarity predictions (by increasing the cost of false negative errors) [1].

IV. Experiments and Results

Two experiments were conducted on the work presented in this paper, the first experiment was conducted to: (i) identify the effect of Malay SL modification and expansion to the accuracy of SA performed and (ii) investigate the effect of SentiStrength score modification for Strength-based PSA to SA accuracy. The second experiment was conducted with the following objectives: (i) to identify the best method to assign scores to words in the SL (the best PSA) and (ii) to compare the performance of bias-aware classification with simple classification to SA.

TABLE I: Experimental results of the effect of SL expansion and SentiStrength score modification to SA classification accuracy (%)

Method	Classification accuracy (%)	
	Before	After
SL modification and expansion	**68.40**	**83.97**
SentiStrength score modification	77.43	81.72

From Table I, the performance of SA classification has improved from 68.40% to 83.97% (before and after SL modification and expansion were conducted). That is an improvement of 15.57%. SentiStrength score modification also has attributed to 4.29% increased in the accuracy. Based on this result, the next experiment considered only the modified and expanded SL. With respect to Strength-based PSA, the modified SentiStrength score is used.

TABLE II: Experimental results of the performance of different PSA and Classification method to SA with $t = 0.05$

PSA method	Classification method	Accuracy (%)
Simple PSA	Simple	83.97
Simple PSA	Bias-aware	84.88
PSA-SN	Simple	83.75
PSA-SN	**Bias-aware**	**85.10**
Strength-based PSA	Simple	81.72
Strength-based PSA	Bias-aware	82.62

In Table II, each PSA method was experimented with simple classification and bias-aware classification. From the table, it can be seen that each PSA technique performed better using bias-aware classification. The accuracy of SA increased by 1% to 2% when bias-aware classification was applied. The threshold value, t, used in this paper is 0.05 (the best t value produced in our early experiment which is not shown here). PSA-SN coupled with bias-aware classification have the highest accuracy of 85.10%. Strength-based PSA has the lowest accuracy.

Based on the results produced, it shows that PSA-SN coupled with bias-aware classification produced the best SA accuracy. This is because the switch-negation operation increased the accuracy of SA classification in each text as it takes account of the linguistic effect within the text that affects the sentiment scores assigned to words that is being negated. Without switch negation, when negation word is coupled with a positive word, it will produce a zero polarity score instead of a negative polarity score and classify the sentence as neutral instead of negative.

Strength-based PSA has lower accuracy compared to Simple PSA. We found that some of the Sabah words can be represented with multiple English words that have different polarity scores for each word. Choosing the best representation of English word for the Sabah words is therefore need a systematic approach. Without systematic approach, there is a possibility that scores assigned by SentiStrength to the chosen translation of the word failed to represent the word polarity strength accurately [13]. In this paper, the words are chosen based on the translation of Malay and Sabah opinion words to English, first by Google Translate. If the translation of the word is zero-valued in SentiStrength, Online Dewan Bahasa dan Pustaka translator is used. The English words that have non-zero polarity values in SentiStrength is selected. If none of the translated words have non-zero polarity, then it will be assigned a default value as described in Section 2.

With respect to SL building, how the SL is built also affects the accuracy of the SA. Based on the results presented in Table II, the Sabah SL expansion has increases SA accuracy significantly. The modification of SentiStrength scoring for Strength-based classification has also improved the SA classification accuracy.

V. Conclusion

This paper presents an approach to SA of Sabah language using lexical approach and bias-aware classification. An SL building that was based on Malay SL expanded with Sabah words found in the corpus was described. Experiments conducted shows the proposed approach produced promising SA classification accuracy. Like any other SA lexicon-based approach, the expansion of the Sabah SL plays an important role in the classification, as the detection of sentiment-bearing words in the corpus written in the language is dependent on the quality of the SL. However, as Malay and Sabah are often used together in Sabah sentences, some words such as “buli”, represent conflicting polarity in Sabah and Malay. Such cases need further contextual analysis of the sentence.

During pre-processing the challenge is the implementation of the short form corrector (lemmatisation) due to occurences of wrong-spelled words. Spellchecker algorithms, such as Lehvenstein distance, succeeded in correcting wrong-spelled words; however it is not efficient when correcting short form words. Contextual consideration may reduce the mistakes.

For Strength-based PSA, a more effective strategy is needed in order to increase the accuracy of the polarity scores assigned to individual words. The challenge is the translation of Malay and Sabah words into English words, and also in picking the best representation of the word in English as there is no SentiStrength database for Malay and Sabah language developed yet. More focus should therefore be directed at the translation and a method to assign the best score to individual Malay and Sabah words.

With respect to switch negation, the rules used can be improved in the future to heuristically detect patterns of negation in the sentence. This could be aided by Part-Of-Speech (POS) tagging. The adopted list of negation words can also be expanded. In short, improvements to switch negation can provide a sentence-level classification method rather than a limited to feature-level classification.

Acknowledgment

This work was supported in part by Artificial Intelligence Research Unit and Centre of Excellence of Semantic Agents, UMS.

References

[1] A. Alsaffar and N. Omar, “Integrating a lexicon based approach and k nearest neighbour for malay sentiment analysis,” *Journal of Computer Science*, vol. 11, no. 4, 2015.

[2] A. Ortigosa, J. M. Martn, and R. M. Carro, “Sentiment analysis in facebook and its application to e-learning,” *Computers in Human Behavior*, vol. 31, pp. 527–541, 2014.

[3] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[4] M. Communications and M. Commission, *Internet Users Survey 2014*. Malaysian Communications and Multimedia Commission, 2015.

[5] M. Puteh, N. Isa, S. Puteh, and N. A. Redzuan, “Sentiment mining of malay newspaper (samnews) using artificial immune system,” in *Proceedings of the World Congress on Engineering*, 2013.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[7] M. Iqbal, A. Karim, and F. Kamiran, “Bias-aware lexicon-based sentiment analysis,” in *30th Annual ACM Symposium on Applied Computing*. ACM, 2015, pp. 845–850.

[8] N. Azmina, M. Zamani, S. Z. Z. Abidin, N. Omar, and M. Z. Z.Abiden, “Sentiment analysis: Determining people’s emotions in facebook,” in *Applied Computational Science*, 2013, pp. 111–116.

[9] B. Liu, *Handbook of natural language processing*, 2010, ch. Sentiment Analysis and Subjectivity, pp. 627–666.

[10] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller, “Wordnet: An online lexical database,” *Int. J. Lexicograph*, vol. 3, no. 4, pp. 235–244, 1990.

[11] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *Proceedings of the 7th Conference on Language Resources and Evaluation*, 2010, pp. 2200–2204.

[12] B. Y. Liau and P. P. Tan, “Gaining customer knowledge in low cost airlines through text mining,” *Industrial Management and Data Systems*, vol. 114, no. 9, 2014.

[13] N. F. Shamsudin, H. Basiron, Z. Saaya, A. F. N. A. Rahman, M. H. Zakaria, and N. Hassim, “Sentiment classification of unstructured data using lexical based techniques,” *Jurnal Teknologi*, vol. 77, no. 18, 2015.

[14] M. T. K. B. G. P. D. C. A. Kappas, “Sentiment strength detection in short informal text,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[15] A. Das and S. Bandyopadhyay, “Towards the global sentiwordnet,” in *PACLIC*, 2010, pp. 799–808.

[16] L. C. Leong, S. Basri, and R. Alfred, “Enhancing malay stemming algorithm with background knowledge,” in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2012.