Jurnal
Teknologi

Full Paper

# SENTIMENT CLASSIFICATION OF UNSTRUCTURED DATA USING LEXICAL BASED TECHNIQUES

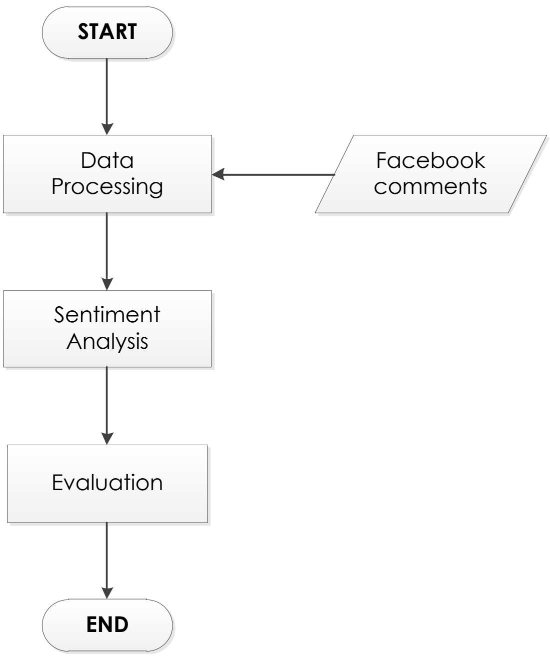
Nurul Fathiyah Shamsudin[a]*, Halizah Basiron[a], Zurina Saaya[a], Ahmad Fadzli Nizam Abdul Rahman[a], Mohd Hafiz Zakaria[a], Nurulhalim Hassim[b]

[a]Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[b]Faculty of Engineering Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

**Article history**
Received
*15 May 2015*
Received in revised form
*1 July 2015*
Accepted
*11 August 2015*

*Corresponding author
fathiyah.shamsudin@gmail.com

## Graphical abstract




## Abstract

Sentiment analysis is the computational study of people's opinion or feedback, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. There are many research conducted for other languages such as English, Spanish, French, and German. However, lack of research is conducted to harvest the information in Malay words and structure them into a meaningful data. The objective of this paper is to introduce a lexical based method in analysing sentiment of Facebook comments in Malay. Three types of lexical based techniques are implemented in order to identify the sentiment of Facebook comments. The techniques used are term counting, term score summation and average on comments. The comparison of accuracy, precision and recall for all techniques are computed. The result shows that the average on comments method outperforms the other two techniques.

*Keywords*: Sentiment analysis, lexical based approach, term counting, term score summation, average on sentence and average on comments

# 1.0 INTRODUCTION

The growth in popularity of online social media (SM) has been remarkable. A survey in [1] has shown that Facebook is the most popular social media site compared to the other popular platforms such as Twitter and Instagram. Facebook was created in 2004 and according to [2], up to 31 March 2015, there are 936 million Facebook daily active users with 270 million daily active users in the Asia-Pacific region alone. The quantity of qualitative data that is generated by these applications is immense. For example, SM offers a rich dataset which can be used by higher educational institutions to enhance their understanding of how students form their opinions about the institution's offerings [3]. Top management in higher education or any local government agencies can definitely benefit from the idea of sentiment analysis. They can simply know how they engage with the common public, how people react (either negatively or positively) to any recent moves that they make and determine which one triggers people's attitude.

Comments posted on SM platforms are deemed honest and authentic as the platforms are unmoderated and can be used freely to express their feelings and perceptions. The results from [4] shows that people uses SM to communicate with real personality. The comments posted are useful to administrators to react strategically. In order to know whether the polarity of comments is positive or negative, sentiment analysis can be implemented. Sentiment analysis is the computational study of people's opinion or feedback, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [5]. There are many research in sentiment analysis conducted for different languages such as English, Spanish, French, and German [6]. However, there seems to be lack of research focusing on sentiment analysis for Malay language [7, 8].

The objective of this paper is to introduce a lexical based method in analysing sentiment of Facebook comments in Malay language. Term scoring, term score summation and average score method will be used in adjective scoring and the score adjective dictionary will be used to assign scores on words and thus determine the sentiments of each comments. Certain words which have more than one meaning or senses will be treated as one by referring which part of speech it belongs to. In other words, word sense disambiguation case is not considered in this research.

The rest of this paper is organized as follows. In Section 2.0 we describe the existing methods or techniques used in sentiment analysis. This section includes some reviews on the existing works for Malay language. Next, the proposed lexical based sentiment analysis system is described in Section 3.0. Section 4.0 discusses the results of the proposed methods. Last but not least, the conclusion and future works are outlined in Section 5.0.

# 2.0 RESEARCH BACKGROUND

As for techniques used in sentiment analysis, there are two main techniques: machine learning and lexical-based. Many researches have implemented the former technique and only currently the latter technique is selected. Existing research in sentiment analysis for Malay have been implemented machine learning technique. However machine learning technique does not perform well across domain and on unseen data. On the other hand lexical based approach performance is consistent across domain and on completely unseen data [9].

None has chosen the lexical based technique for the technique to analyse the polarity of feedback comment in Malay language. Hence, this is the proposed technique for this research. The reliability of sentiment analysis in social media content for Malay language can be improved using lexical based approach.

## 2.1 Machine Learning

The supervised method uses any existing machine learning methods to perform sentiment analysis. It involves building classifiers from feedback sentences in which the feedback sentences (such as movie reviews) are used as training and testing data [10]. The commonly used machine learning techniques are Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME) and k-Nearest Neighbour (kNN). Generally, this method starts with the collection of data. The data will be split into a training set and a test set. The training set is used for classifier learning process and the test set is used in testing the classifier's performance after the learning process is done. Feature selection is the process of selecting a set of attributes or features that is relevant to the mining processes [11]. In machine learning, feature selection is useful as the important features are shown and can be used in predictions.

[10] applies Support Vector Machine, Naïve Bayes and Maximum Entropy in document-level sentiment classification for movie reviews. The classifiers are trained on unigrams and bigrams features. Based from the results, the size of the training data affected the performance of the classifiers. NB classifiers perform better for smaller training data but when it comes to larger training data set, SVM perform best compared to NB and ME.

Machine learning techniques perform very well in a domain that their classifiers are trained on. The performance drops when the same classifier is used in a different domain [9, 12]. The main weakness of supervised method is that the performance of the classifier is dependent on the training data. A training data with larger size and higher quality information results in better classification. The lack of information and insufficient amount of training data may lead to misclassification.

### 2.2 Natural Language Processing

The unsupervised method is a method which is based on words or phrases from the feedback [5]. The natural language processing (NLP) is one of the example which focuses on the syntax and semantic of the feedback. This approach is also known as lexical based. This approach involves calculating an orientation of a text document from the semantic orientation of words or phrases in the document [9]. For example, a comment posted in Facebook is extracted to determine whether the comment is a positive or negative respond. Using a dictionary which consists of positive and negative words or phrases, the words in the comment are compared. If the comment consists of more positive words, then the comment is categorized as positive. Otherwise it is deemed as negative.

The semantic orientation of words or phrases can be referred as the lexicon of sentiment words or phrases. For most sentiment analysis algorithms, the sentiment lexicon is the most important resource [13]. In this research, we use the term "adjective score dictionary" and "score dictionary" for our version of sentiment lexicon. Our proposed score dictionary is inspired by [20]. The score dictionary mainly consists of Malay adjectives.

[9] presents a lexical based approach in extracting sentiment from text. They built the Semantic-Orientation CALculator (SO-CAL) based on dictionary of words with its orientation or polarity which includes adjectives, nouns, verbs, adverbs, intensifiers and negation. [14] uses SentiWordNet as the source of their opinion lexicon in film review classification. The reviews are classified by counting the positive and negative term score based on their opinion lexicon known as Term Counting. The review is classified based on the highest class (positive or negative) count. [15] proposes a technique to improve the efficiency of using SentiWordNet by [14]. The proposed technique is originally from Term Counting while considering the magnitude of the positivity and negativity of the words. Term Score Summation method is done by summing up the positive scores and negative scores for each word. Average on Sentence and Average on Review method is done by calculating the average of the positive scores and negative scores for each word in every sentence. Then, the average positive scores and negative scores for each sentence is calculated to get the positive and negative score for the review. Both of these methods determine the classification of the review based on the highest class score.

[5] confirms that even though this lexical approach does not invariably outperform machine learning method, but its overall track record is better. [9] also claims that lexicon based methods for sentiment analysis are robust, resulting in good cross-domain performance, and can be easily enhanced with multiple sources of knowledge.

### 2.3 Existing Works for Malay Language

Currently, most of the resources and systems built for sentiment analysis are tailored to English and other Indo-European languages [6]. There are few which focused on Malay language. [7] develops the sentiment analysis system for Malay newspaper. They focus on sentiment mining on textual data collected from newspapers and classifying the sentiment value using Negative Selection Algorithm (NSA) of Artificial Immune System (AIS) technique. [8] studies opinion mining from online movie reviews uses machine learning classifiers: Support Vector Machine (SVM), Naïve Baiyes and k-Nearest Neighbour. The data used are collected from several forums and blogs written by the Malaysian. [11] creates a feature of selection algorithm based on artificial immune network system (FS-INS) on pre-processed online Malay messages and uses three different machine learning techniques to analyse the FS-NIS: Naïve Bayes (NB), k-Nearest Neighbour (kNN) and Sequential Minimal Optimization (SMO). Table 1 shows the comparison of existing techniques used in sentiment classification on Malay language.

## 3.0 SYSTEM OVERVIEW

Our proposed sentiment analysis system consists of three major phases as shown in Figure 1. In this paper, we focus on the Sentiment Analysis tasks which will be explained in Section 3.2.

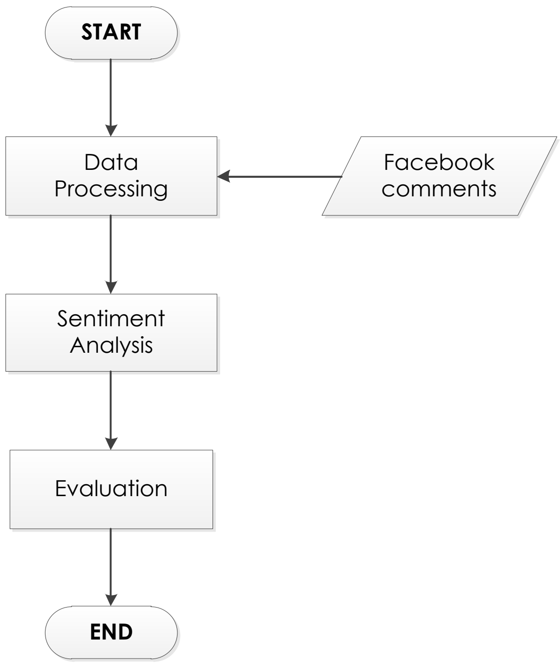



**Figure 1** The Proposed Sentiment Analysis on Facebook Malay Comments Overall Process Flow

**Table 1** Existing techniques of sentiment analysis on Malay language

| Paper | Method | Features | Dataset | Type of language |
|---|---|---|---|---|
| Puteh 2013 [7] | Artificial Immune System (AIS) | Negative Selection Algorithm (NSA) | Malay newspaper | Formal |
| Samsudin & Hamdan 2011 [8] | • SVM<br>• Naïve Bayes<br>• k-NN | Standard method | Online Malay movie review from forums and blogs | Informal |
| Samsudin *et al.* 2013 [11] | • NB<br>• kNN<br>• SMO | Feature Selection based on Immune Network System (FS-INS) | Online Malay messages | Informal |

### 3.1 Data Processing

In this phase, there are two tasks involved towards the data: data processing and data testing. In the data processing task, the data used for this research are Facebook comments written by Malaysian in public pages post. These Facebook comments are freshly extracted into CSV file. These raw data contains rows of spam messages, Facebook emoticons and other languages comments. These are also comments written in an improper manner. Thus, data cleaning needs to be done in order to provide a reliable source for the proposed methods. Below is the list of cleaning task done during our research:

a) Removing spams and irrelevant words from the comments.
b) Removing emoticons and excessively used symbols from the comments such as :),'…','!!!!'
c) Converting all texts for each comment into lowercases.

After data cleaning, the next step is data preparation. In this task, we prepare two types of data. The score dictionary data is a list of adjective words which will be used in lexical based sentiment analysis. We prepared two types of score dictionary for this research – Malay adjective score dictionary and Malay-English (BM-ENG) score dictionary. Malay adjective score dictionary is constructed by extracting the commonly used Malay adjective words from the data. Then, each of the adjective words is manually categorized as negative with a value of -1 or positive with a value of +1. This dictionary only contains adjective of Malay words.

We also consider adjective that have polarity value does not converts to 100% positive sentiment. The same goes to the negative words which may or may not possess a 100% negative sentiment. This is where we created English score dictionary based on the Malay adjective words. The English words in the English score dictionary are the translation of the Malay adjectives words. For example, the Malay word "cukup" is translated into "enough", "exactly", "sufficient" and "adequate". The creation of English score dictionary is inspired from [16] in which they use WordNet as their resource of English dictionary for sentiment identification. Alternatively, we use SentiWordNet as our English score dictionary. This dictionary includes other part of speech apart from adjectives for the English words.

The final step is the preparation of the testing data. The testing data is used in order to test our proposed methods. The testing data consist of rows of user comments along with its positive or negative category (refer Figure 2). The categorizing of the comments in the testing data is manually hand tagged. The performance of the proposed methods will be tested on this testing data. Further details in measuring the performance will be discussed in Section 3.3.

| | A | B |
|---|---|---|
| 1 | neg | Let s do it prof  jangan sembang kencang sambil duduk dalam aircond  Ketahian semu |
| 2 | neg | Orang Kita Bukan Malas Tapi Pendapatan Bulanan Tidak Padan Dengan Kos Sara Hidup |
| 3 | neu | Aku pun protes jugak dengan kenaikan harga mcm2 barang dan jugak kos hidup yang |
| 4 | neg | selagi wat masuk pendatang asing atau mahkluk asing selagi tu la ko ckp malas bangsa |
| 5 | neg | Ada yg ckp anak muda Malaysia malas x mcm pekerja asing sanggup keja apa sj walau |
| 6 | neg | Ada beza malas dengan buat keje yang x berbaloi  pekerja asinh dtg keje sini sbb curre |
| 7 | pos | Wahai bangsa melayu ku  fahimi lah maksud malas tu  ia bukan bermaksud malas bek |
| 8 | neu | Bkn sume mls ada juga yg kerja siang mlm tp ttp x ckup nk sara kos hdup di kl dgn fmli |
| 9 | neg | Bile ckp ayat xsepatUtnye kt kaum lain dikate rasis  tp pangkt pRof perli org melayu ap |
| 10 | pos | bagi aku xsalah pom prof tu cakap  laki aku keje kontraktor kecik2  mula2 pakai org m |
| 11 | neu | sape usaha lbh die dpt lbh tu je agama pun ckp camtu tak kire lah asing ke bukan asin |
| 12 | pos | mesti ramai yang akan marah bila aku ckap ni   orang kita memang malas   aku bekerj |
| 13 | pos | Ada betulnya Jgn cpt melatah Aq ni student je Percaya tak aq ckp gaji aq masa lepas s |
| 14 | neg | melayu malas kepala pungkok dia setakat sembang duduk dalam aircond boleh la ceri |
| 15 | neu | Orang kita bukan malas tapi miskin hidup dinegeri sendiri |
| 16 | pos | Orang kita memang malas Bukan takat malas malah memilih Aku tengok dengan mata |
| 17 | neg | Ak setuju dgn kenyataan prof tu ak idop dkelilingi realiti ap y belaku skrg tgk je kt baza |
| 18 | neg | Ayoo prof ptt buat kajian  draf taraf kehidupan rakyat kebanyakkan selaras dgn keada |
| 19 | pos | btul la ckp prof tu keje2 mcm bangla tu buat bukn kita nak buat ckp jijik ah x de kelas |
| 20 | neu | Sorry to say  hakikatnya warga asing dpt meniaga buka kedai2 atas lesen rakyat tempa |
| 21 | neg | Skrg blaja je dh hutang Blaja tinggi2 belum tentu dapat kerja Ramai ada kelayakan dipl |
| 22 | neu | aku kerja gaji 900 sebulan mmpu tmpung adk aku sekolah fom 3  cuma nk ingt kn seti |

**Figure 2** Sample of Testing Data

### 3.2 Sentiment Analysis

This is the most crucial part of sentiment analysis system process. In this phase, the Facebook comments are categorized into two classes namely positive and negative by implementing the lexical based approach. Figure 3 outlines the tasks performed in the sentiment analysis process.

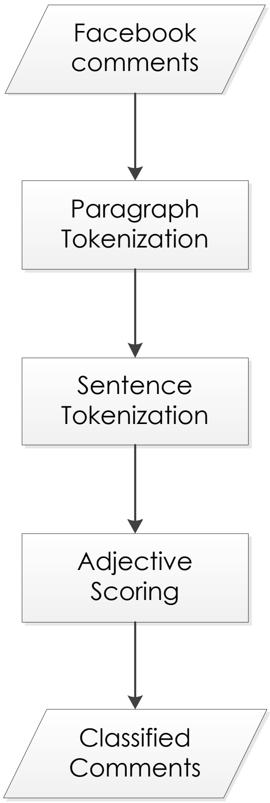



**Figure 3** Proposed Sentiment Analysis Process

### 3.2.1 Paragraph and Sentence Tokenization

We treat one comment from a user as one paragraph which means one comment consists of many sentences. The process of breaking comments (paragraph) into sentences is called Paragraph Tokenizing. Each sentence consists of many words. These words are then compared, one by one, with the adjective word dictionary. Words that match with the adjective dictionary will be used in determining the polarity of the sentences and comments.

### 3.2.2 Adjective Scoring

Adjective scoring is a task in which the comments are being classified into positive or negative. We use three different methods in sentiment classification: Term Counting, Term Score Summation, and Average on Sentence and Average on Comments. The total scores of positive and negative will determine the polarity of the comments. We also refer to our Malay score adjective dictionary and BM-ENG score adjective dictionary as the source of positive and negative scores for the words. Our main focus is the Malay adjective words. Since there are non-adjective English words exists in BM-ENG score dictionary, these words are considered as adjectives as they are derived from the Malay adjective word. By knowing the polarity scores of the Malay adjectives, we can now proceed to sentiment classification by implementing the lexical based methods. One of the methods are using Malay score dictionary while the other two uses BM-ENG score dictionary. The three methods used are described below:

***Term Counting (TC)***: This method is introduced by [14]. It is a simple method to classify positive and negative review by counting the positive or negative words found in a review. The sentiment polarity is based on which class received the highest score. This method is solely depending on the Malay adjective score dictionary (refer to Figure 4).

For example, we apply Term Counting method on two different sentences. According to the Malay adjective score dictionary, there are two adjective words in the Sentence (A) which are "*banyak*" and "*naik*". Both "*banyak*" and "*naik*" are categorized as positive (+1). It is clear that this sentence is a positive sentence since there are more positive word count compared to negative word count. Sentences (B) is a negative sentence because it contains more negative words ("*malas*" and "*tidak*") compared to positive words.

(A) "Berat badan saya naik banyak."
(B) "Saya malas dan tidak rajin."

| ID | Word | Category | Value |
|---|---|---|---|
| 100000 | cukup | pos | 1 |
| 100001 | boleh | pos | 1 |
| 100002 | naik | pos | 1 |
| 100003 | banyak | pos | 1 |
| 100004 | tidak | neg | -1 |
| 100005 | malas | neg | -1 |
| 100006 | bukan | neg | -1 |

**Figure 4** A Sample of Malay Adjective Score Dictionary

***Term Score Summation (TSS)***: This method is introduced by [15]. It is an enhanced version of Term Counting by [14]. The positivity, negativity and objectivity magnitude for each word is considered in calculating the review's score. The summation of positive and negative scores for each word that found in a sentence is calculated to get the positive and negative scores for a comment. Then, the sentiment of the comment is determined based on which score has the highest value [15]. This method uses both Malay adjective score (refer Figure 4) and BM-ENG score dictionary (refer Figure 5).

For example, we apply the Term Score Summation method on a comment consisting of 2 sentences as shown in Example (C). In the first sentence, it contains one Malay adjective word- "*cukup*" with the Malay ID 100000 in Malay adjective score. By using the Malay ID, according to the BM-ENG score dictionary (refer Figure 5), the word "cukup" can be translated into 4 English words ("enough", "exactly", "sufficient" and "adequate").The Therefore, the positive score of this sentence is:

0.125 + 0.375 + 0.125 + 0.125
= 0.75

and the negative score is:

0 + 0 + 0.75 + 0
= 0.75
Same procedure is used for the second sentence. The second sentence contains one Malay adjective – "*boleh*" (Malay ID 100001) with 6 English words ("can", "able", "permit", "ability", "capable" and "talented"). The second sentence positive score is:

0.625 + 0.125 + 0 + 0.5 + 0.125 + 0.5
= 2
and the negative score is:
0.105 + 0 + 0.25 + 0 + 0 + 0
= 0.355.
Then, we add up the positive scores and negative scores of all sentences. Thus, the comments positive score is:
0.75 + 2
= 2.75
and the negative score is:
0.75 + 0.355
=1.055.
The comment is categorized as positive since the positive score is higher that the negative score.

(C) "Bagi kami, makanan itu sudah cukup"
"Makanan itu boleh kami kongsi kepada yang lain"

| ID | MALAY | WORD | TAG | POS | NEG |
|---|---|---|---|---|---|
| 200000 | 100000 | enough | Adj | 0.125 | 0 |
| 200001 | 100000 | exactly | Adv | 0.375 | 0 |
| 200002 | 100000 | sufficient | Adj | 0.125 | 0.75 |
| 200003 | 100000 | adequate | Adj | 0.125 | 0 |
| 200004 | 100001 | can | Nou | 0.625 | 0.105 |
| 200005 | 100001 | able | Adj | 0.125 | 0 |
| 200006 | 100001 | permit | Ver | 0 | 0.25 |
| 200007 | 100001 | ability | Nou | 0.5 | 0 |
| 200008 | 100001 | capable | Adj | 0.125 | 0 |
| 200009 | 100001 | talented | Adj | 0.5 | 0 |

**Figure 5** A Sample of BM-ENG Score Dictionary

***Average on Comments (ASAC)***: This method is originally named as Average on Sentence and Average on Review which is introduced by [15] in classifying review. For each sentence in a comment, positive and negative scores are determined by calculating the average of positive and negative scores for each word found in it. Then, the averages of positive and negative scores for these sentences are calculated to get the positive and negative scores for the comment. Determining sentiment polarity is based on which score has the highest value [15].

By using the same example in the previous method, we apply the Average on Comments method. Instead of summing up the positive and negative scores, this method finds the average positive and negative scores for each sentence. Thus, in the first sentence, the positive score of this sentence is:
(0.125 + 0.375 + 0.125 + 0.125)/4 = 0.75/4
= 0.1875
and the negative score is:
0 + 0 + 0.75 + 0 = 0.75/4
= 0.1875.
The second sentence positive score is:
(0.625 + 0.125 + 0 + 0.5 + 0.125 + 0.5)/6 = 2/6 = 0.3333
and the negative score is:
(0.105 + 0 + 0.25 + 0 + 0 + 0)/6 = 0.355/6
= 0.05916.
Then, we find the average positive score and negative score for the comment by using all the scores of the sentences in the comment. Therefore, the positive score of the comments is:
(0.1875 + 0.3333)/2 = 0.5208/2
= 0.2604
and the negative score is
(0.1875 + 0.05916)/2 = 0.24666/2
= 0.12333.
Since the average of positive scores is higher than the average of negative scores, the comment is categorized as positive.

## 3.3 Evaluation

The evaluation of the proposed methods is measured by comparing the results from proposed methods with the human-tagged categorization. The human-tagged data are shown in Table 2. The data are compared with classification results that produced by the lexical-based techniques. The comparison results are referred in calculating accuracy, precision and recall.

The formula of accuracy, precision and recall is based on the confusion matrix shown in the Table 3.

**Table 2** Total of classification of human-tagged data

| Class | Total data |
|---|---|
| Positives | 81 |
| Negatives | 246 |
| Neutral | 46 |
| None | 77 |

**Table 3** Confusion Matrix

| | | Actual Class | |
|---|---|---|---|
| | | P | N |
| Predicted Class | P | ***tp*** (true positive) | ***fp*** (false positive) |
| | N | ***fn*** (false negative) | ***tn*** (true negative) |

Accuracy is the ratio of all correctly classified instances against all predicted instances stated in formula (1).

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \quad (1)$$

Precision is ratio correctly classified instances against its relevant predicted class. The formula for positive and negative precision is stated in formula (2) and (3) respectively.

$$Precision\ (positive) = \frac{tp}{tp + fp} \quad (2)$$

$$Precision\ (negative) = \frac{tn}{tn + fn} \quad (3)$$

Recall is the ratio of correctly classified instances to the total number of instances of its relevant actual classes. The formula for positive and negative recall is stated in formula (4) and (5) respectively.

$$Recall\ (positive) = \frac{tp}{tp + fn} \quad (4)$$

$$Recall\ (negative) = \frac{tn}{fp + tn} \quad (5)$$

## 4.0 RESULTS AND DISCUSSION

The algorithms of the lexical based methods are tested with a data consists of 450 Facebook comments extracted from a public page. As mentioned in the Section 3.0, the data is pre-processed before it is used for sentiment analysis task.

Table 4 illustrates the comparison of accuracy, precision and recall for Term Counting, Term Score Summation and Average on Comments. The accuracy and recall results show that Average on Comments performs best as compared to Term Counting and Term Score Summation. Even though Average on Comments shows the lowest in precision but it shows the highest in accuracy (63.41%) and recall (78.7%). We believe that the poor accuracy result is due to the adjective scoring method in BM-ENG adjectives score dictionary. In the current scoring method, Malay adjectives are first translated to English and a polarity score is determined based on the corresponding English words in SentiWordNet. The polarity value for certain English words are not equivalent to Malay words. For example, the English translation of Malay word "cukup" are "enough", "exactly", "sufficient", and "adequate". Referring to Figure 5, each English translation word for "cukup" has different polarity values. Therefore, there is a need to develop Malay SentiWordNet in order to determine the accurate polarity value for Malay adjectives.

**Table 4** Comparison of Results

| | TC | TSS | ASAC |
|---|---|---|---|
| **Accuracy (%)** | 62.89 | 60.97 | 63.41 |
| **Precision** | 0.762 | 0.760 | 0.751 |
| **Recall** | 0.756 | 0.724 | 0.787 |

## 5.0 CONCLUSIONS AND FUTURE RESEARCH

In this paper we have introduced lexical based method for sentiment analysis specifically for content in Malay language from social media. We have applied three adjectives scoring methods namely Term Counting, Term Score Summation and Average on Comments with our own version of BM-ENG Score Dictionary for sentiment classification. We have described the results of comprehensive evaluation based on different adjectives scoring methods. In particular, the Average on Comments technique outperforms Term Counting and Term Count Summation methods in accuracy and recall. These initial results are promising. They demonstrate the potential improvement in our adjective scoring methods and BM-ENG adjectives score dictionary.

As for future work, we plan to add more Malay words in the BM-ENG Score Dictionary and enhance the capability of this dictionary. Specifically, the scores in adjective score dictionary can be replaced with more reliable sources which is based on the Malay words rather than English words. We could also evaluate our adjectives scoring methods using different set of data. We will also consider the usage of jargons in comments as features for sentiment classification. This type of content might contain useful information that leads to the polarity of the comments.

## Acknowledgement

We are grateful the Ministry of Higher Education for the grant Research Acculturation Grant Scheme (RAGS/1/2014/ICT07/FTMK/B00089) that provided financial support of this research.

## References

[1] Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A. & Madden, M. 2014. Social Media Update 2014. *Pew Internet and American Life Project*. From: http://www.pewinternet.org/2015/01/09/social-media-update-2014/. [Accessed on 7 May 2015].

[2] Argaez, E. D. 2015. Internet World Stats–Usage and Population Statistics. *Facebook Users in the World*. [Online]. From: http://www.internetworldstats.com/facebook.htm. [Accessed on 7 May 2015].

[3] Roblyer, M. D., McDaniel, M., Webb, M., Herman, J., & Witty, J. V. 2010. Findings on Facebook in Higher Education: A Comparison of College Faculty and Student Uses and Perceptions of Social Networking Sites. *The Internet and Higher Education*. 13(3): 134-140.

[ 4 ] Back, M., Stopfer, J., Vazire, S., Gaddis, S., Schmukle, S., Egloff, B. and Gosling, S. 2010. Facebook Profiles Reflect Actual Personality, Not Self-Idealization. *Psychological Science*. 21(3): 372.
[ 5 ] Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
[ 6 ] Korayem, M., Crandall, D., & Abdul-Mageed, M. 2012. *Subjectivity and Sentiment Analysis of Arabic: A Survey*. Advanced Machine Learning Technologies and Applications.
[ 7 ] Puteh, M., Isa, N., Puteh, S., & Redzuan, N. A. 2013. Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. In *Proceedings of the World Congress on Engineering*. Vol. 3.
[ 8 ] Samsudin, N., Puteh, M., & Hamdan, A. R. 2011. Bess or xbest: Mining the Malaysian Online Reviews. In *Data Mining and Optimization (DMO), 2011 3rd Conference on IEEE*. 38-43.
[ 9 ] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. 2011. Lexicon-based Methods for Sentiment Analysis. *In Association for Computational Linguistics*. 37(2): 267-307.
[ 10 ] Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics. 79-86.
[ 11 ] Samsudin, N., Puteh, M., Hamdan, A. R. and Ahmad, M. Z. 2013. Mining Opinion in Online Messages. *International Journal of Advanced Computer Science and Applications*. 4(8): 19- 24.
[ 12 ] Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 2(1-2).
[ 13 ] Feldman, R. 2013. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*. 56(4): 82-89.
[ 14 ] Ohana, B., & Tierney, B. 2009. Sentiment Classification of Reviews Using SentiWordNet. In *9th. IT & T Conference*. 13.
[ 15 ] Hamouda, A., & Rohaim, M. 2011. Reviews Classification Using Sentiwordnet Lexicon. In *World Congress on Computer Science and Information Technology*.
[ 16 ] Kim, S. M., & Hovy, E. 2004. Determining the Sentiment Of Opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics. 1367.
[ 17 ] Eguchi, K and Lavrenko, V. 2006. Sentiment Retrieval Using Generative Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 345-354.
[ 18 ] Hajmohammadi, M. S., Ibrahim, R., & Othman, Z. A. 2012. Opinion Mining and Sentiment Analysis: A Survey. *International Journal of Computers & Technology*. 2(3): 171-178.
[ 19 ] Indurkhya, N. and F. Damerau. 2010. *Handbook of Natural Language Processing*. Chapman & Hall/Crc Machine Learning & Pattern Recognition.
[ 20 ] O'Neill, A. 2009. Sentiment Mining for Natural Language Documents, Book Sentiment Mining for Natural Language Documents', Australian National University.
[ 21 ] Thelwall, M., Buckley, K., and Paltoglou, G. 2012. Sentiment Strength Detection for the Social Web. *Journal of America Society Information Science Technology*. 63(1):163-173.
[ 22 ] Tromp, E. 2011. Multilingual Sentiment Analysis on Social Media. Master Thesis, Dept. Computer Science, Eindhoven University.