Lexical Based Sentiment Analysis – Verb, Adverb & Negation

Nurul Fathiyah Shamsudin, Halizah Basiron, Zurina Sa'aya
Faculty of Information & Communication Technology,
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia.
fathiyah.shamsudin@gmail.com

***Abstract*— Sentiment analysis is a method to determine whether the feedback given by the user is positive or negative. The comments posted by users consists noisy text which includes abbreviations, misspelling and short forms. Sentiment analysis becomes challenging when dealing with noisy data. The objective of this paper is to introduce a lexical based method in classifying sentiment of Facebook comments in Malay language. Two types of lexical based techniques namely Term Counting and Term Counting Average are implemented in order to classify the sentiment of Facebook comments. Several parts of speech tags are being taken into account. Pre-processing process is involved in dealing noisy texts in data. Term Counting works better for adjectives and adverbs while Term Counting Average performs better for verbs and negation words.**

***Index Terms*— Sentiment analysis; Lexical based approach; Term counting, Term counting average.**

I. INTRODUCTION

Facebook is the most popular social media site compared to the other popular social media platforms [1]. Up to 31 March 2015, there are 936 million Facebook daily active users with 270 million daily active users in the Asia-Pacific region. Social media users use comments to express their feelings and perceptions. According to [3], it shows that people prefers social media platforms to communicate with real personality. These comments are valuable to administrators to react strategically. In order to know whether the polarity of comments is positive or negative, sentiment analysis can be implemented.

The objective of this paper is to present a lexical based method in classifying sentiment of Facebook comments based on the adjectives, verbs, adverbs and negation. This paper is organized as follows. In Section II we describe the existing methods or techniques used in sentiment analysis. This section also includes the discussion on existing score dictionary or opinion lexicon. Next, the proposed lexical based sentiment analysis system is described in Section III. Section IV discusses the results of the proposed methods. Lastly, Section V is the conclusion of this paper and future work.

II. RESEARCH BACKGROUND

Sentiment analysis is the computational study of people's opinion or feedback, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [4]. There are many research in sentiment analysis conducted for different languages such as English, Spanish, French, and German [5]. However, there seems to be lack of research focusing on sentiment analysis for Malay language [6, 7]. There are two main approaches in sentiment analysis – machine learning based and lexical based. Many researches have implemented the former technique and only currently the latter technique is selected. In this paper, we use the term "sentiment classification".

A. Methods Used in Sentiment Classification

One of the methods in performing sentiment analysis is b using any machine learning methods. This method is also known as the supervised method. In machine learning, training and testing data are crucial in performing this approach. When applying the machine learning approach in sentiment analysis, these data are based on the feedback sentences. From these feedback sentences, the classifiers are identified [9]. Machine learning techniques such as Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME) and k-Nearest Neighbour (kNN) are commonly used in previous works. Another important matter in machine learning approach is the feature selection. It is the process of selecting a set of attributes or features that is relevant to the mining processes [10].

Pang et al. [9] applied SVM, NB and ME in document-level sentiment classification for movie reviews. Puteh et al. [6] developed the sentiment analysis system for Malay newspaper using the Negative Selection Algorithm (NSA) of Artificial Immune System (AIS) technique. Samsudin et al. [7] studied opinion mining from online Malaysian movie reviews uses machine learning classifiers: SVM, NB and kNN. Samsudin et al. [10] introduced a feature of selection algorithm based on artificial immune network system (FS-INS) on pre-processed online Malay messages and used three different machine learning techniques to analyze the FS-NIS: NB, kNN and Sequential Minimal Optimization (SMO).

The latter approach is the unsupervised method. It is a method which is based on words or phrases from the feedback [4]. The method that we are using for this paper is the natural language processing (NLP) focuses on the syntax and semantic of the feedback. According to [9], this approach involves calculating an orientation of a text document from the semantic orientation of words or phrases in the document.

Taboada et al. [8] developed a system called Semantic-

Orientation CALculator (SO-CAL), which uses lexical based approach in extracting sentiment from text. Their score dictionary is a collection of English words with its orientation or polarity of adjectives, nouns, verbs, adverbs, intensifiers and negation. Ohana et al. [13] classified film reviews using a method called Term Counting (TC). TC classifies the reviews by counting the positive and negative term score based on their SentiWordNet opinion lexicon. Hamouda et al. [14] studied the TC method by [13] and improvised the efficiency of using the SentiWordNet based opinion lexicon by considering the magnitude of the positivity and negativity of the words. They introduced two methods called Term Score Summation (TSS) and Average on Sentence and Average on Review (ASAR). TSS method is done by summing up the positive scores and negative scores for each word while the ASAR method is done by calculating the average of the positive scores and negative scores for each word in every sentence. Both of these methods determine the classification of the review based on the highest class score.

B. Score Dictionary

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 [12]. In this research, we use the term "score dictionary" for our version of sentiment lexicons. The idea of creating a score dictionary based on existing wordnet is inspired by a work done by [15].

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. We choose Wordnet Bahasa, an open source of Malay and Indonesia lexicon dictionary as our score dictionary resource [16]. In this paper, we select adjectives, adverbs and verbs of Malay words from the Wordnet Bahasa only.

III. SYSTEM OVERVIEW

Our proposed sentiment analysis system consists of three major phases as shown in Figure 1. In this section, we explain our progress during Data Processing phase and Sentiment Analysis phase in Section 3.1 and 3.2 respectively.

A. Data Processing

Every noise words found in the comments are converted into its full form or meaningful word. For example, the word "makan/eat" can be represented as noisy words of "mkn" and "mkan". The task is being done by comparing the noisy word in the comments with a collection of data of noisy words which contains noisy word (short form) and its corresponding meaningful word (longform), and then the noisy texts in the comments are replaced by its corresponding full form. We called this data the Fullform dictionary (Figure 2). The Fullform dictionary is created by manually extracting the noisy texts from the comments and manually hand-tagged its corresponding full form word.

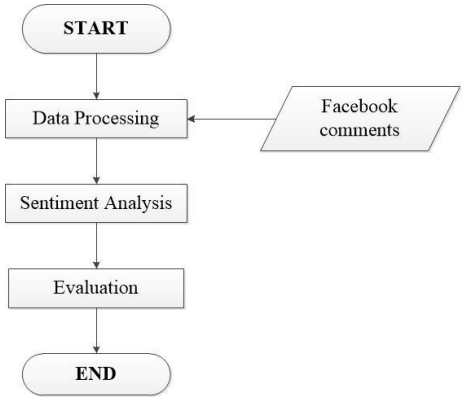

Figure 1: The Overall Process Flow

The next task is the data preparation. In this paper, we create three additional Malay score dictionaries which will be used as the source of calculating the polarity score of the comments. We have created the Verb score dictionary (Figure 4), the Adverb score dictionary (Figure 5) and the Negation dictionary (Figure 6). We also update the entries of our existing Adjective dictionary. To achieve this, we use Wordnet Bahasa as our score dictionary resource [16]. For the adjectives and verbs word, the words are manually categorized as negative (-1) or positive (+1). As for the adverbs, the words are act as modifier to the word paired with the adverb. These adverbs are manually categorized to a word that doubles (2) or halves (0.5) the polarity score of the paired word. As for the negation words, these words act as the polarity inverter of the word which paired with the negation word. The negation words do not need to be assigned a score.

Shortform	Longform
ad	ada
ade	ada
ader	ada
de	ada
adk	adik
adik2	adik-adik
adill	adil

Figure 2: A sample of Fullform dictionary

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

Figure 3: A Sample of Malay Adjective Score Dictionary

ID	WORD	TAG
300016	berair	neu
300017	terlanjur	neg
300019	main	neu
300020	tetap	pos
300021	bersuara	neu

Figure 4: A Sample of Malay Verb Score dictionary

ID	WORD	VALUE
400006	mungkin	0.5
400007	hampir	0.5
400008	belum	0.5
400009	kemudian	0.5
400010	agaknya	0.5

Figure 5: A Sample of Malay Adverb Score Dictionary

Kata Nafi
tidak
jangan
bukan
tiada
enggan
entah

Figure 6: A Sample of Malay Negation dictionary

B. Sentiment Analysis

The Sentiment Analysis phase is where the sentiment score calculation is done. Based on its sentiment score, the Facebook comments are categorized into two classes (positive and negative) by implementing the lexical based approach. Figure 7 outlines the tasks performed in the sentiment analysis process. In this paper, we are focusing on the Sentiment Scoring process. More details of the other processes are explained in [17].

i) Adjectives, Verbs, Adverbs and Negation

In our previous work ([17]), we choose Adjective as our source of our scoring methods. In this paper, we include verbs, adverbs and negation for our scoring methods. The idea of using the adjectives, verbs, adverbs and negation is inspired by a system called Semantic Orientation CALculator (SO-CAL) developed by [8] but we decide to implement a simpler method.

(A) "Saya sedang berusaha untuk belajar."

"I am trying to learn."

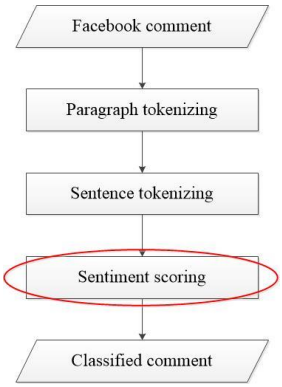

Figure 7: Sentiment Analysis Process

Verbs are calculated as same as the adjectives. For example, in Sentence (A), "berusaha/*trying*" and "belajar/*study*" are verbs. The word "berusaha/*trying*" has (+1) polarity, same goes to the word "belajar/*study*" which also has (+1) polarity. If we apply Term Counting method, the calculation of the scoring is:

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

So, this sentence has a score of 2 and it indicates this sentence is positive.

(B) a) "Saya tidak belajar." (*"I do not study."*)

b) "Saya tidak hodoh." (*"I am not ugly."*)

Negation negates the current polarity of paired words. When a negation word is paired with a positive word, it will become negative. When a negation word is paired with a negative word, it becomes positive. We apply negation calculation when negation is paired with adjective or verb. For example, in sentence (B), we choose the word "tidak/*not*" as a negation. In (a), there is one negation word and one verb word ("belajar/*study*") is found. The negation word "tidak/*not*" is paired with the verb "belajar/*study*" which has (+1) polarity. So, it negates the positive polarity of the word "belajar/*study*", from "belajar/*study*" (+1) to "tidak belajar/*not studying*" (-1). Thus, the sentence (a) is categorized as negative. In (b), there is one negation word and one adjective word ("hodoh/*ugly*") is found. "hodoh/*ugly*" is an adjective which has (-1) polarity. So, the negation word "tidak/*not*" negates the negative polarity of the word "hodoh/*ugly*", changing into "tidak hodoh/*not ugly*" (+1). Thus, the sentence (b) is categorized is positive.

(C) a) "Saya selalu belajar." (*"I always study."*)

b) "Saya jarang belajar." (*"I rarely study."*)

Adverbs alter the magnitude of current polarity of the paired words. The adverb is either doubles or halves the polarity value of the paired words. For example in C, the word "selalu/*always*" and "jarang/*rarely*" are the examples of adverbs in Malay. In sentence (a), the adverb "selalu/*always*" which has a modifier value of 2, is paired with a verb "belajar/*study*" which has a positive polarity (+1). The calculation is:

$$(+1) \times 2 = 2$$

So, the sentence (a) has a positive polarity with a score of (+2). In sentence (b), the adverb "jarang/*rarely*" which has a modifier value of 0.5 is paired with a verb "belajar/*study*" which has a polarity value of 1. The calculation is:

$$(+1) \times 0.5 = 0.5$$

So, the sentence (b) has a positive polarity with a score of (+0.5). The word "selalu belajar/*always study*" in sentence (a) shows stronger positive polarity and the word "jarang belajar/*rarely study*" in sentences (b) shows weaker positive polarity compared to the word "belajar/*study*" itself.

ii) Scoring Methods

We are using two types of scoring methods known as Term Counting and Term Counting Average. These methods will calculate the sentiment score of the comments and the comments will be categorized based on the sentiment score

(refer Table 1). A comment with a positive sentiment score value is categorized as positive while comment with a negative sentiment score value is categorized as negative. The comment is categorized as neutral if it has a sentiment score value of 0.

Term Counting (TC): This method is introduced by [13]. 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).

Table 1
Polarity Criteria

Condition	Polarity
Sentiment Value > 0	Positive
Sentiment Value = 0	Neutral
Sentiment Value < 0	Negative

For example, we apply TC method on two different sentences. According to the Malay adjective score dictionary, there are two adjective words in the Sentence (D) which are "banyak/*a lot*" and "naik/*increase*". Both "banyak/*a lot*" and "naik/*increase*" are categorized as positive (+1). It is clear that this sentence is a positive sentence since there are more positive word counts compared to negative word count. Sentences (E) is a negative sentence because it contains more negative words ("malas/*lazy*" and "tidak/*not*") compared to positive words.

(D) "Berat badan saya naik banyak."

"My weight increases a lot."

(E) "Saya malas dan tidak rajin."

"I am lazy and not hardworking."

Term Counting Average (TCAvg): This method is a modified TC method, inspired by Average on Sentence and Average on Review method introduced by [14]. In our previous work ([17]), we applied the exactly the same method by [14] which combines Malay and English score dictionary. In this paper, we use only Malay dictionary and rename the method as Term Counting Average. Our TCAvg is done by calculating the average of polarity score in each sentence. Then, the total of average scores of the comment will determine the polarity of the comment.

For example, we are going to determine the polarity of a comment (F) which consists of two sentences. In this example, the word "berusaha/*trying*" with a score of (+1), "belajar/*study*" with a score of (+1) in the first sentence and "bodoh/*stupid*" with a score of (-1) in second sentence are selected for our scoring calculation. Thus, the calculation for the first sentence is:

$$((+1) + (+1)) / 2 = 1$$

and the calculation for second sentence is:

$$(-1)/1 = -1$$

Then, we total up the average score for both sentences to get the score for the comment. The calculation is:

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

Since the total score is (0), the comment is categorized as neutral.

(F) "Saya sedang berusaha untuk belajar. Saya orang bodoh."
"I am trying to learn. I am stupid."

Inspired by the work from [18] and [19], we decide to perform the scoring methods on several different combination of words based on their part of speech (POS) tags. The combinations of POS tags are listed below:

Adj + Neg: This is combination of adjective and negation word. If a negation word is found before adjective word, the negation word will negate the polarity of the adjective word. If the adjective word has a sentiment score of (+1), it will change to (-1) and vice versa. If no adjective or verb presents after the negation word, the negation word is assigned a score of (-1).

Adj + Adv: This is a combination of adjective and adverb word. If an adverb word is found before an adjective word, the adverb word will alter the magnitude of the adjective word's sentiment score. If the adverb of value (2) is found before an adjective with value of (+1), it doubles the score from (+1) to (+2). If the adverb of value (0.5) is found before an adjective word with a value of (+1), it halves the score from (+1) to (+0.5).

Verb + Neg: This is combination of verb and negation word. This combination works the same as Adj + Neg where if the negation word is found before the verb word, the negation word negates the polarity of the verb word. If no adjective or verb presents after the negation word, the negation word is assigned a score of (-1).

Verb + Adv: This is a combination of verb and adverb word. This combination works the same as Adj + Adv where if the adverb word is found before the verb word, the adverb word alters the magnitude of the verb word's sentiment score.

C. Evaluation

We conduct our experiment by testing the TC and TCAvg methods with a data of 450 Facebook comments, extracted from a public page posted by Malaysian users. These comments are manually hand-tagged to its corresponding polarity (refer Table 2). There are 49 comments are excluded in our experiment as these comments are consists of spams and unreliable data.

The evaluation of the proposed methods is measured by comparing the results from proposed methods with the human-tagged categorization data. The data are compared with classification results that produced by the lexical-based techniques. The comparison results are referred in calculating accuracy - the ratio of all correctly classified instances against all predicted instances.

Table 2
Polarity Criteria

Class	Total data
Positives	80
Negatives	249
Neutral	72
Out	49
Total	450

IV. Results And Discussion

Figure 5 illustrate the accuracy of different POS combinations used in TC and TCAvg method. Firstly, let us observe the accuracy of TC and TCAvg methods, which uses Adjectives only as their scoring source, on the original data Adj(with noise) and pre-processed data Adj. The TC and TCAvg methods applied on pre-processed data produces better result compared to the original data with an improvement of 10.03% and 6.95% respectively. This shows that data pre-processing is crucial in producing a quality result.

Secondly, let us compare the results of Adjective POS combinations on pre-processed data. The Adj + Neg combination shows the highest accuracy for both of the methods while Adj shows the lowest accuracy. Next, we compare the results of Verb POS combinations. We can see that the Verb + Neg combination shows the highest accuracy while the Verb + Adv combination shows the lowest accuracy. Overall, the Verb + Neg combination of TCAvg has the highest accuracy (52.12%) while the Verb + Adv combination of TC method has the least accuracy (6.36%).

V. Conclusion

In this paper, we implement a few parts of speech (POS) in lexical based method for Malay language sentiment analysis. We take verbs, adverbs and negations into account and create a list of POS combination to be implemented in Term Counting (TC) and Term Counting Average (TCAvg) scoring methods. Overall, the TC method outperforms TCAvg when adjectives are present while TCAvg performs better than TC when verbs present in the comment. It also shows that TC method works better for Adverb words while TCAvg works better for Negation words. As for future work, we plan to refine our existing scoring method to utilize the true definition of each word. We could also make use of different sets of data to measure the reliability of our methods.

Acknowledgment

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.

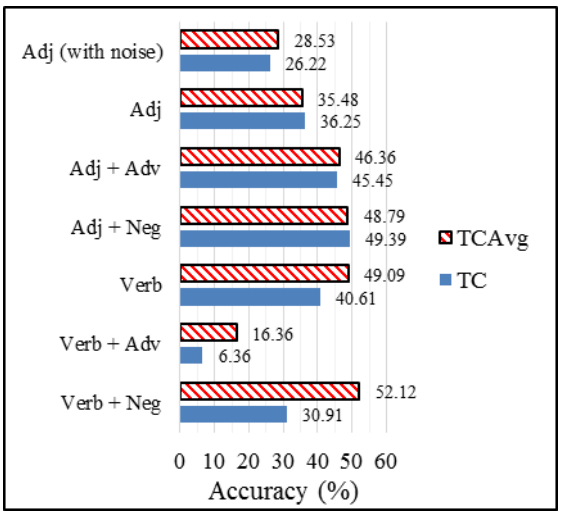

Figure 8: Accuracy comparison between POS combinations

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. 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] 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.

[4] Liu, B. (2012). Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

[5] Korayem, M., Crandall, D., & Abdul-Mageed, M. (2012). Subjectivity and Sentiment Analysis of Arabic: A Survey, Advanced Machine Learning Technologies and Applications.

[6] 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).

[7] 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 (pp. 38-43). IEEE.

[8] 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.

[9] 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(pp. 79-86). Association for Computational Linguistics.

[10] Samsudin, N., Puteh, M., Hamdan, A. R. and Ahmad, M. Z. (2013). Mining Opinion in Online Messages. In International Journal of Advanced Computer Science and Applications, 4(8), 19- 24.

[11] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. In Foundations and Trends in Information Retrieval 2(1-2).

[12] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

[13] Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. In 9th. IT & T Conference (p. 13).

[14] Hamouda, A., & Rohaim, M. (2011). Reviews classification using sentiwordnet lexicon. In World Congress on Computer Science and Information Technology.

[15] Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[16] Noor, N. H. B. M., Sapuan, S., & Bond, F. (2011). Creating the Open Wordnet Bahasa. In PACLIC (pp. 255-264).

[17] Shamsudin, N. F., Basiron, H., Saaya, Z., Rahman, A. F. N. A., Zakaria, M. H., & Hassim, N. (2015). Sentiment Classification of Unstructured Data Using Lexical Based Techniques. Jurnal Teknologi, 77(18).

[18] Subrahmanian, V. S., & Reforgiato, D. (2008). AVA: Adjective-verb-adverb combinations for sentiment analysis. Intelligent Systems, IEEE, 23(4), 43-50.

[19] Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., & Subrahmanian, V. S. (2007, March). Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. In ICWSM.