English and Malay Cross-lingual Sentiment Lexicon Acquisition and Analysis

Nurul Amelina Nasharuddin[1], Muhamad Taufik Abdullah[1(✉)], Azreen Azman[1], and Rabiah Abdul Kadir[2]

[1] Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
{nurulamelina,mta,azreenazman}@upm.edu.my
[2] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
rabiahivi@ukm.edu.my

Abstract. Sentiment analysis finds opinions, sentiments or emotions in user-generated contents. Most efforts are focusing on the English language, for which a large amount of sources and tools for sentiment analysis are available. The objective of this paper is to introduce a cross-lingual sentiment lexicon acquisition method for the Malay and English languages and further being test on a set of news test collections. Several part of speech tags are being experimented using the Word Score Summation technique in order to classify the sentiment of the news articles. This method records up to 50% as experimental accuracy result and works better for verbs and negations in both the English and Malay news articles.

1 Introduction

Many existing research on textual information processing has focused on mining and retrieval of factual information. But recently with the growth of the Web, there is a lot of user-generated information available such as users' views and opinions on products on merchant sites, forums, blogs and discussion groups. Text processing focusing on opinions and sentiments has become increasingly important, not only to individuals but also for organisations. Sentiment analysis or opinion mining may be defined as a general method to find opinions, sentiments and emotions in text. Sentences in texts can be classified as objective or subjective where a subjective sentence may contains a positive, negative or neutral opinion. Most efforts are focusing on the English language, where a large amount of sources and tools are available especially for sentiment analysis. There are two major problems exist in the under-resourced sentiment analysis work: (1) limited number of formally standardised sentiment lexicon and (2) few number of sentiment classifier that are publicly available.

Hence, a better solution is to have a sentiment orientation analysis based on existing linguistics resources in highly-resourced languages, such as the English language. This work proposed an automatic cross-lingual sentiment lexicon acquisition for the English and Malay languages where Malay is being considered as an under-resourced language.

K. Kim and N. Joukov (eds.), *Information Science and Applications 2017*,
Lecture Notes in Electrical Engineering 424, DOI 10.1007/978-981-10-4154-9_54

Then an analysis on the sentiment for both English and Malay languages was experimented using the developed sentiment lexicon.

2 Related Work

2.1 Sentiment Analysis

There are two main approaches in extracting sentiment from texts namely the lexicon-based (an unsupervised approach) and the statistical or machine-learning (a supervised approach). Lexicon-based unsupervised learning approach uses sentiment dictionaries. Dictionaries or lexicons which are being used for this approach can either be constructed manually, using seed words, utilising existing dictionaries or making use of other lexical resources like WordNet [1, 2]. The pre-built dictionaries such as SentiWordNet [3] contain words along with their associated sentiment polarity (positive or negative) and strength. The SentiWordNet was developed following the structure of the English WordNet built by Princeton University [4]. The word class (nouns, verbs, adjectives and adverbs) are grouped into synonyms sets (or synsets) linked by semantic relations. Lexicon-based approach does not require storing a large data corpus and training, so the whole process is much faster. Classifiers built using statistical or supervised approaches usually perform very well in detecting the polarity of a text as they are generally trained on a large annotated corpus [5]. However, the performance is usually not good when the same classifier is applied on different domain that they have been trained on.

2.2 Malay Sentiment Analysis

Research in sentiment analysis for the Malay languages is scarce compared to the English language. Two studies focused on finding sentiments in Malay news document using Artificial Immune System technique inspired by the biological immune system responding toward foreign antigen [6, 7]. This supervised learning algorithm is widely adapted by other research such as in computer and network security. Zamani and his colleagues studied on how to analyse sentiments in English and Malay words in Facebook [8]. Their work focused on quantifying Facebook sentiments using a lexicon-based approach for both the English and Malay texts. Liau and Tan [9] also used a lexicon-based approach to study the customer's satisfaction level towards low-cost airlines in Malaysia, in order to understand the consumer' needs. These two researches created a small-scale lexicon and employed a very simple classifier to calculate the polarity of the sentiments.

One of a Malay lexical databases widely used in Malay language research study is the Wordnet Bahasa. It is a combination of lexical semantics from three different resources which are the Malay Wordnet, the Indonesian Wordnet and the Wordnet Bahasa [10]. It contains over 45,000 Malay words and 58,000 Indonesian words. The Wordnet Bahasa was also developed based on the English WordNet. Thus, a cross-lingual sentiment lexicon for English and Malay language can be developed by mapping both the Wordnet Bahasa and the English SentiWordNet with the English WordNet as they were built using the same structures. This mapping method is independent of

translation using a bilingual dictionary as both lexical databases have the same unique synsets value for each word entry.

3 Methodology

Figure 1 shows the methodology performed in this work where it comprised of four main phases namely, (i) pre-processing of test documents, (ii) cross-lingual sentiment lexicon acquisition, (iii) sentiment processor and finally (iv) document sentiment categorisation, in order to classify the documents into positive, negative or objective classes.

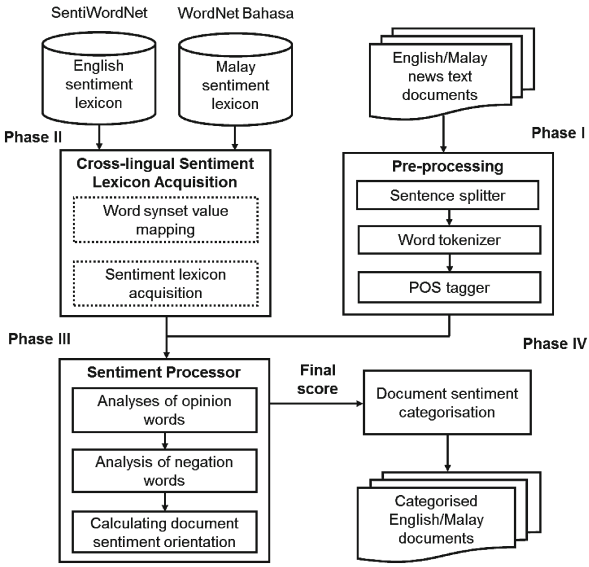

Fig. 1. The proposed methodology for the English and Malay sentiment analysis.

3.1 Test Documents Collection

In this work, a total of 883 of an unstructured English and Malay news articles were selected as the test documents. These articles which published in the year of 2005 were from a national news agency in Malaysia, the Bernama. The documents contain various contents which include politics, business, economy, executive reports and sports. This will avoid the domain-specific limitation during the experiments.

3.2 Test Documents Pre-processing

The English and Malay test documents were first annotated with the XML tags describing the author, publication date, headline as well as the contents. Then the documents were being pre-processed. Three main tasks of the pre-processing for both English and Malay documents were:

Lexical Analysis. The operations being done in this step were treating digits, hyphens and punctuation marks. All the digits in the documents were removed and the hyphens were replaced by space. All the capital letters were left unchanged as they were important indicators during the part-of-speech (POS) tagging.

Stopwords Removal. Next was to remove stopwords - the most frequent words that often do not carry much meaning. This work used the SMART stopword list of 571 words for the English documents and a list of 321 stopwords for the Malay documents.

Stemming and POS Tagging. The English test documents were stemmed using the Stanford CoreNLP Java library, a natural language analysis tool in order to remove the inflectional endings and return the base of the words in the documents. For extracting the POS of the English documents, the work employed the Stanford POS Tagger [11]. As for the Malay documents, this work implemented the stemmer from Abdullah, Ahmad, Mahmod and Tengku Sembok [12] and a rule-based POS tagger by Alfred and his colleagues [13] to tag the words into their POS.

3.3 English-Malay Cross-lingual Sentiment Lexicon Acquisition

A cross-lingual sentiment lexicon is needed to enable sentiment analysis to be performed on both English and Malay test documents. Thus, an automatic acquisition method was proposed to create a cross-lingual sentiment lexicon for English and Malay languages using the similarity structures of the English SentiWordNet and Wordnet Bahasa. The suggested approach consists of two main processing steps, explained in details as below:

Automatic Mapping of the Cross-lingual Synonyms. The Wordnet Bahasa was mapped to the English SentiWordNet using the synset value and POS value as the key to build the cross-lingual sentiment lexicon. Then, the synonyms for both languages were matched with each other. The mapping includes all types of POS in the Wordnet Bahasa which were nouns, verbs, adjectives and adverbs. As the Word Sense Disambiguation (WSD) was not one of the objectives of this work, word's prior polarity was being used to address the problem on how to compute the sentiment orientation of one word. Though the technique was less precise, it was guaranteed that the same score was given to the same word in different contexts. To find the prior polarity of a word, the weighted mean formula as in Eq. (1) was adopted

$$posScore = \frac{\sum_{i=1}^{n} \left(\frac{1}{i} \times posScore_i \right)}{\sum_{i=1}^{n} (\frac{1}{i})} \tag{1}$$

where *posScore* represents the positive value of a word, n represents the total number of senses of the word and $posScore_i$ represents the absolute value of the positive value of the i-th sense of that word [14]. The *objScore* (the objective or neutral value of a word) and *negScore* (the negative value of a word) were calculated using the same steps. In this technique, each sense weight was chosen according to a harmonic series according

to their frequency of being used. The choice was based on the assumption that more frequent senses should bear more "affective weight" than very rare senses.

Cross-lingual Sentiment Lexicon Acquisition. An English-Malay sentiment lexicon was generated from the previous step. The advantage of this method is it does not require translation unlike other cross-lingual lexicon-based approaches. The construction of this lexicon was based on the following assumptions; (1) the senses of the words in both Malay and English language were the same and (2) the sentiment score for each English word was similar to the matched Malay word.

3.4 English and Malay Sentiment Processor

In this step, the relevant words which used to find the sentiment of the document were extracted from each document. The relevant words include the adjectives, adverbs, verbs and negation words. The positivity, negativity and neutrality (or objectivity) values from the previous step were combined to determine the documents' sentiment orientations. For each document in both English and Malay languages, the method first finds the *posScore*, *negScore* and *objScore* values of the related word. The method then checks if there were negations used before the related word and if yes, the three sentiment values of the words will be updated. After all the related words have been checked, each of the sentiment values was aggregated to get the overall *posScore*, *negScore* and *objScore* values for the documents. Four different combinations of POS were being experimented in this work were $Adj + Neg$, $Verb + Neg$, $Adv + Adj + Neg$ and $Adv + Verb + Neg$.

For the aggregation method of the documents' sentiment scores, this work enhanced the Word Score Summation method by Hamouda and Rohaim by including the objective scores of the documents where these scores were important in a general domain text collection [15]. In this method, the sentiment scores for each positive, negative and objective word in a document were added up and negated if a negation word appeared. The overall sentiment orientation of the document was chosen based on which from these three overall scores had the highest value. It was assumed that the document's author sentiment was correlated to the choice and number of relevant words presented in the document.

3.5 Document Sentiment Categorisation

The results from the previous step were used to automatically categorise the documents into the three orientation categories, which are positive, negative and neutral. Then for each of the document, the overall document's orientation will be classified as positive if the score bigger than zero. If the overall score is less than zero, the document's orientation will be classified as negative and if equal to zero, the orientation is neutral.

4 Experiments

4.1 Experimental Settings

The automatic categorisation of the documents sentiments were then assessed manually by three language experts in both Malay and English languages. A contingency table was being used to compare the performances of the method and the manual assessments by the experts. Two different tables were created for the English and Malay test documents. The three-class contingency table representing the positive, negative and objective categories as in Table 1.

Table 1. Contingency table for three sentiment categories to compare between experts' assessments and proposed method

		Proposed method		
		Positive	Objective	Negative
Expert	Positive	T_{Pos}	E_{PosObj}	E_{PosNeg}
	Objective	E_{ObjPos}	T_{Obj}	E_{ObjNeg}
	Negative	E_{NegPos}	E_{NegObj}	T_{Neg}

For example, T_{Pos} is the total number of documents that were correctly categorised as positive and E_{ObjNeg} is when an objective document is wrongly categorised to negative category by the proposed method. From this table, the performance of the proposed method was measured using four commonly-used measurements which are the accuracy, precision, recall and F1. For example in calculating the performance of the positive category, the equations are as follows:

$$accuracy = \frac{T_{Pos} + T_{Obj} + T_{Neg}}{all\ instances\ in\ the\ table} \tag{2}$$

$$precision_{Pos} = \frac{T_{Pos}}{T_{Pos} + E_{ObjPos} + E_{NegPos}} \tag{3}$$

$$recall_{Pos} = \frac{T_{Pos}}{T_{Pos} + E_{PosObj} + E_{PosNeg}} \tag{4}$$

$$F1 = \frac{2 * precision_{Pos} * recall_{Pos}}{precision_{Pos} + recall_{Pos}} \tag{5}$$

4.2 Experimental Results

Table 2 shows the accuracy results for the proposed sentiment method for the English and Malay test documents. It can be observed that the POS combination of *Verb* + *Neg* shows the highest accuracy for both English and Malay documents at 50.2% and 47.9%, respectively. However, the worst performance is obtained by the *Adv* + *Adj* + *Neg*

combination with only 24.3% for English documents and 21.6% for Malay documents. The macro-averaged precision, recall and F1 results for English and Malay documents that are obtained are shown in Table 3. Macro-averaged measurements are the average values of the three categories of sentiment, being experimented.

Table 2. Accuracy comparison between different POS combinations for the English and Malay test documents.

	English	Malay
Adj + Neg	44.1	44.2
Verb + Neg	50.2	47.9
Adv + Adj + Neg	24.3	21.6
Adv + Verb + Neg	33.1	39.2

Table 3. Macro-averaged precision, recall and F1 comparisons between different POS combinations for the English and Malay test documents.

POS	Precision		Recall		F1	
	English	Malay	English	Malay	English	Malay
Adj + Neg	0.30	0.31	0.34	0.32	0.32	0.31
Verb + Neg	0.31	0.32	0.34	0.36	0.32	0.34
Adv + Adj + Neg	0.18	0.34	0.34	0.34	0.24	0.34
Adv + Verb + Neg	0.34	0.35	0.36	0.33	0.35	0.34

None of the macro-averaging F1 results of the proposed method are statistically significant. The best performance (0.35 and 0.34) of the proposed method is achieved when the POS combination of *Adv + Verb + Neg* are used for both English and Malay documents and also the combinations of *Verb + Neg* and *Adv + Adj + Neg* for Malay documents. In general, both accuracy and F1 scores for the English and Malay test documents are somewhat low. This finding was unexpected and suggests that rigorous steps need to be done in improving the scores. A possible explanation for this might be that the adoption of prior polarity which totally removes the senses of each words. Another possible explanation for this is that the lack of established and reliable pre-processing tools on the Malay language might poses problems to the overall experiments.

5 Conclusion and Future Work

This paper presents an extensive work on the development of a cross-lingual English and Malay sentiment lexicon and further used the lexicon in sentiment analysis task. The main contribution of this work is to effectively develop an automatic method for the cross-lingual sentiment lexicon construction which in turn will help the Malay language sentiment analysis in future. This paper attempts to determine which POS combinations perform best to detect sentiments on English and Malay documents. The results indicate that the best POS combinations are the *Adv + Verb + Neg* and *Adv + Adj + Neg* for both English and Malay test documents. Future work will be

focusing on improving the pre-processing phase of the Malay documents especially on the stemming task. The word sense disambiguation should also be included in future research to address the word sense problem for the Malay sentiment classification. Different kind of test collection could also be experimented in order to confirm the method robustness. Despite these limitations, it is believed that this study has contributed to this subject especially in the Malay sentiment analysis research.

Acknowledgments. This work was supported by the Fundamental Research Grant Scheme (FRGS), MOHE Malaysia. We thank Bernama for their assistance and advise during the data collection process.

References

1. Taboada, M., Brooke, J., Tofilofski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguis. **32**(2), 267–307 (2011)
2. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1075–1083 (2007)
3. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation. European Language Resources Association, pp. 2200–2204 (2010)
4. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
5. Boiy, E., Hens, P., Deschacht, K. Moens, M.-F.: Automatic sentiment analysis in on-line text. In: Leslie, C., Martens, B. (eds.) Proceedings of the 11th International Conference on Electronic Publishing, pp. 349–360 (2007)
6. Isa, N., Puteh, M., Raja Mohamad Hafiz, R.K.: Sentiment classification of Malay newspaper using immune network (SCIN). In: Proceedings of the World Congress on Engineering (2013)
7. Puteh, M., Isa, N., Puteh, S., Redzuan, N.A.: Sentiment mining of Malay newspaper (SAMNews) using artificial immune system. In: Proceedings of the World Congress on Engineering (2013)
8. Zamani, N.A.M., Abidin, S.Z.Z., Omar, N., Abiden, M.Z.Z.: Sentiment analysis: determining people's emotions in Facebook. In: Zaharim, A., Sopian, K., Psarris, K., Margenstern, M. (eds.) Proceedings of the 13th International Conference on Applied Computer and Applied Computational Science, pp. 111–116 (2014)
9. Liau, B.Y., Tan, P.P.: Gaining customer knowledge in low cost airlines through text mining. Ind. Manag. Data Syst. **114**(9), 1344–1359 (2014)
10. Bond, F., Lim, L.T., Tang, E.K., Riza, H.: the combined Wordnet Bahasa. In: Chung, S.-F., Nomoto, H. (eds.) Current Trends in Malay Linguistics, vol. 57, pp. 83–100 (2014)
11. Toutanova, K., Dan, K., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 173–180 (2003)
12. Abdullah, M., Ahmad, F., Mahmod, R., Tengku Sembok, T.: Rules frequency order stemmer for Malay language. Int. J. Comput. Sci. Netw. Secur. **9**(2), 433–438 (2009)

13. Alfred, R., Mujat, A., Obit, J.H.: A ruled-based part of speech (RPOS) tagger for malay text articles. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013. LNCS (LNAI), vol. 7803, pp. 50–59. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36543-0_6
14. Gatti, L., Marco, G.: Assessing sentiment strength in words prior polarities. In: Proceedings of COLING 2012: Posters, COLING 2012, pp. 361–370 (2012)
15. Hamouda, A., Rohaim, M.: Reviews classification using SentiWordNet lexicon. Online J. Comput. Sci. Inf. Technol. **2**(1), 120–123 (2012)