Automatically Generating A Sentiment Lexicon For The Malay Language

Article · July 2016

DOI: 10.17576/apjitm-2016-0501-05

CITATIONS
16

READS
705

3 authors:

Mohammad Darwich
Universiti Kebangsaan Malaysia
14 PUBLICATIONS **81** CITATIONS
SEE PROFILE

Shahrul Azman Mohd Noah
Universiti Kebangsaan Malaysia
238 PUBLICATIONS **1,976** CITATIONS
SEE PROFILE

Nazlia Omar
Universiti Kebangsaan Malaysia
73 PUBLICATIONS **613** CITATIONS
SEE PROFILE

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 5 No. 1, June 2016: 49 - 59
e-ISSN: 2289-2192

AUTOMATICALLY GENERATING A SENTIMENT LEXICON FOR THE MALAY LANGUAGE

MOHAMMAD DARWICH
SHAHRUL AZMAN MOHD NOAH
NAZLIA OMAR

ABSTRACT

This paper aims to propose an automated sentiment lexicon generation model specifically designed for the Malay language. Lexicon-based Sentiment Analysis (SA) models make use of a sentiment lexicon for SA tasks, which is a linguistic resource that comprises a priori information about the sentiment properties of words. A sentiment lexicon is an indispensable resource for SA tasks. This is evident in the emergence of a large volume of research focused on the development of sentiment lexicon generation algorithms. This is not the case for low-resource languages such as Malay, for which there is a lack of research focused on this particular area. This has brought up the motivation to propose a sentiment lexicon generation algorithm for this language. WordNet Bahasa was first mapped onto the English WordNet to construct a multilingual word network. A seed set of prototypical positive and negative terms was then automatically expanded by recursively adding terms linked via WordNet's synonymy and antonymy semantic relations. The underlying intuition is that the sentiment properties of newly added terms via these relations are preserved. A supervised classifier was employed for the word-polarity tagging task, with textual representations of the expanded seed set as features. Evaluation of the model against the General Inquirer lexicon as a benchmark demonstrates that it performs with reasonable accuracy. This paper aims to provide a foundation for further research for the Malay language in this area.

Keywords: Malay language, sentiment lexicon, sentiment analysis, opinion mining, sentiment dictionary

INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining (OM) models that use the lexicon-based approach incorporate a sentiment lexicon, which is a linguistic resource that consists of sentiment words labelled with their corresponding polarity. Information about sentiment words and phrases contributes to, and is an indispensable resource for, SA tasks. This is evident in the emergence of a large volume of works that focus on the SA subtask of marking words with a sentiment polarity. Positive sentiment words express desired affective states, while negative sentiment words express their undesired counterparts. Sentiment lexicons have been compiled either manually by hand, or via automated algorithms. The latter method makes use of either a dictionary or a corpus as a resource for this task.

Manually tagging words to produce a sentiment lexicon is costly in terms of annotator time and effort, and at the same time, is often associated with subjective bias in annotation, since the judgment of annotators varies to a particular degree (Andreevskaia & Bergler, 2006; Dragut et al., 2012). The increasing popularity of sentiment analysis has resulted in the demand of automatic sentiment lexicon generation algorithms that involve minimal human intervention.

Most prior studies on automatic sentiment lexicon generation focus on English. However, applying readily available off-the-shelf algorithms constructed for English on foreign languages such as Malay is difficult, since the resources available in foreign languages are limited, while an abundance of resources exist for English. Mihalcea et al. (2007) develop a sentiment lexicon in a foreign language by translating an existing sentiment lexicon originally

in English. An evaluation of this translation technique using hand-labelled annotation indicates that a only small volume of words in the target lexicon preserve their original sentiment values after translation; this is because words are translated from a source language into senses other than the ones intended in the target language, increasing the issue of ambiguity. As a consequence, researchers have resorted to compiling and utilizing resources in the target language for automated sentiment lexicon generation in that particular language, as opposed to translating these resources from other languages such as English.

Prior work on sentiment lexicon generation for low-resource languages such as Malay is lacking. This has brought up the motivation to propose an algorithm that automatically labels words with a polarity, specifically for this particular language. A dictionary-based approach was used for this task. WordNet Bahasa (WNB) was mapped onto the English version of WordNet to construct a multilingual word network. This approach is minimally supervised, since a seed set comprising of only one strong positive word and one strong negative word were used to automatically propagate through WordNet's synonymy and antonymy relations for term-polarity labelling. The expanded seed set was used to train a binary classifier to label the remaining words that were not picked up through WordNet propagation. The proposed algorithm was evaluated against the intersecting words in the General Inquirer (GI; Stone et al. 1966) to demonstrate that it performs with reasonable accuracy, and is worthy to be further investigated.

The outline of this paper is structured as follows. Section 2 reviews the related work on sentiment lexicon generation for both English and foreign languages. Section 3 presents the methodology carried out to construct the proposed algorithm. The evaluation and results are discussed in Section 4. Section 5 concludes.

RELATED WORKS

Previous works have been devoted to the development of reliable sentiment lexicons by both manual and automatic means. Automatic approaches can be divided into two categories: (1) dictionary-based (e.g. using WordNet, Merriam Webster, etc.); and (2) corpus-based. Regardless of the approach, the end result is a list of words and their corresponding sentiment polarity that is used for sentiment classification on larger pieces of text (e.g., user reviews or social media posts).

The underlying intuition in using an online dictionary is that words are not only semantically related by meaning, but for the most part, are also related in terms of their semantic orientation. An online dictionary in this approach plays the role of a semantic knowledge base that includes extensive coverage of words defined in a natural language. Also, dictionaries tend to come with semantic relations between words such as synonymy and antonymy. Although these works are outdated, they reflect the foundation of sentiment lexicon generation. More recent works employ these algorithms on various languages, including Hindi, French, Arabic, and Japanese, among others.

Hassan et al. (2011) employ a Markov random walk algorithm on Arabic WordNet and Hindi WordNet, as well as a seed set of manually annotated terms to generate a sentiment lexicon for both Arabic and Hindi respectively. They yield an accuracy of 0.92 and 0.75 respectively. Rao and Ravichandran (2009) use a label-propagation algorithm on Hindi WordNet and on French OpenOffice Thesaurus to generate sentiment lexicons in Hindi and French respectively, suggesting their algorithm is applicable on languages besides English. Japanese sentiment lexicons were constructed for single terms (Takumara et al. 2005) using a sophisticated Ising spin algorithm.

Distance-based semantic similarity measures between subjective seed terms and target terms were proposed to mark target terms with a polarity (Kamps et al., 2004; Williams &

Anand, 2009). WordNet synonymy and antonymy relations were exploited in a bootstrapping algorithm (Hu & Liu, 2004). A word's synset occurrence frequency (Kim & Hovy, 2004) and gloss information (Esuli & Sebastiani, 2006) were used as features for supervised classification. Label propagation was applied on WordNet subgraphs (Blair-Goldensohn et al., 2008; Rao & Ravichandran, 2009). A random walk algorithm was proposed in which predefined words act as absorbing boundaries (Hassan & Radev, 2010). Morphological (affix) features of words were modified to generate new words, while preserving the sentiment features of the original (Mohammad et al., 2009; Neviarouskaya & Prendinger, 2009).

The corpus-based approach relies on co-occurrence statistics or syntactic patterns in a (generally large) text corpus and a set of seed or reference words for automatically marking words with a polarity. Both a dictionary and a corpus extracted from a social media platform were utilized to label words with a polarity and strength (Peng & Park, 2011). A massive corpus was used to measure the polarity of a word based on whether its co-occurrence with a set of positive reference words is greater than its co-occurrence with a set of negative reference words (Turney & Littman, 2003). An approach that extracts pairs of adjectives conjoined by *but* and *and* from a corpus was proposed, with the intuition that conjoined adjectives are subject to linguistic constraints. In general, *and* conjoins a pair of adjectives with equal polarity, while *but* conjoins a pair of adjectives with opposing polarity (Hatzivassiloglou & McKeown, 1997).

The above mentioned works, however, mostly focus on English. There is no previous work that has been devoted to automatically assign Malay words with a polarity, hence, the motivation for this work. Mihalcea et al. (2007) build a sentiment lexicon in a target language by translating a readily-available sentiment lexicon in English. An evaluation of this method using human annotation indicates that only a small portion of the entries in the target lexicon preserve their sentiment properties after the translation process, primarily due to the issue of ambiguity in both the target and source languages. Therefore, terms are first translated from English to Malay by mapping terms from the English version of WordNet to the Malay version via their synset offset values, and then their sentiment polarity was deduced, as opposed to directly translating the final sentiment lexicon using a bilingual dictionary.

CONSTRUCTING THE PROPOSED ALGORITHM

The proposed algorithm to automatically label words in the Malay language with a sentiment polarity is presented. The algorithm was constructed in several steps. The first involves mapping WordNet Bahasa (WNB) onto the English version of WordNet 3.0 to construct a Malay word network, while at the same time preserving the semantic relations between words. This is followed by using a seed set to propagate through the constructed network and automatically label words. The updated seed set was then used to train a binary (positive-negative) supervised classifier to label the remaining words that have not been picked up by the propagation process in the previous step.

MAPPING WORDNET BAHASA ONTO THE ENGLISH WORDNET

WordNet (Fellbaum, 1998; Miller et al., 1990) plays the role of a lexical database with the characteristics of both an online dictionary and a thesaurus, and exhibits an ontological, semantic structure between words. WordNet 3.0 contains about 155,287 words organized into 117,659 synsets. The most important semantic relation is synonymy, which groups words that are synonymous to each other into synonym sets (or synsets). Each synset refers to a unique meaning or concept, and contains a gloss (or a definition), and one or more usage examples. Different senses of a polysemous word are found in different synsets, based on the meaning each sense is associated with. The four major word classes (nouns, verbs, adjectives and

adverbs) for English make up four large subgraphs. Words in each subgraph form dense relations between each other, while relations between words of different word classes are sparse. Nouns and verbs have a taxonomic structure that is defined by hyponym-hyperym relations, while adjectives are grouped into adjective clusters based on direct antonymy. Adverbs form the smallest part of WordNet, and are extracted from adjectives via morphological affixation.

Since the algorithm utilizes the dictionary-based approach, and deals with the Malay language, there is a need for a lexical database in this language. MalayWordNet (Tze & Hussein 2006) contains Malay versions of synsets for only nouns and verbs. The interest is solely in adjectives. Therefore, another Malay version of WordNet was used. The Global WordNet Association aims to provide a platform to connect and standardize WordNet versions for multiple languages. WordNet Bahasa (WNB; Bond et al., 2014; Noor, Sapuan & Bond, 2011) is the current standardized Malay and Indonesian version of WordNet. It contains a total of 49,668 synsets, 145,696 senses and 64,431 unique terms. Table 1 shows a sample of the WNB database. The first column shows a sense's offset-pos value, which is an eight-digit number representing the sense's offset, followed by its part-of-speech (nouns, verbs, adjectives and adverbs are denoted as n, v, a and r respectively). WNB contains senses in Malay, Indonesian, and general Bahasa (denoted as M, I and B respectively); these can be observed in the 'Lang' column. The 'Target language translation' column contains the translations of English senses into the target languages. Note that there may exist one or multiple translations for a particular English sense, as can be observed for the offset '15297472', which contains three different translations.

TABLE 1. Sample of WNB database.

Offset-pos value	Lang	Goodness	Target language translation
15297303-n	B	M	*tempoh percubaan*
15297472-n	B	L	*Percubaan*
15297472-n	B	L	*waktu percubaan*
15297472-n	I	O	*masa percobaan*
15298011-n	B	L	*perbelanjaan*
15298507-n	M	X	*waktu pertanyaan*

Since a portion of WNB was automatically checked, a 'goodness' value is specified for each sense, which represents the degree of the reliability of the translation from English to the target languages. Y, O, M, L and X refer to 'hand checked and good', 'automatically checked and good', 'automatically checked and medium', 'automatically checked and probably bad', and 'hand-checked and bad' respectively. The WNB database was cleaned by discarding all senses with "lang = I" and "goodness = X". This is to remove any terms associated with Indonesian, as well as entries that have been manually checked and confirmed to be of bad quality.

The Malay and Bahasa version senses were mapped to the English WordNet senses via their offset values. For example, the first sense of the adjective "early#1" in the English WordNet corresponds to several translations in Malay, namely, *dini, muda, awal*, *mula, siang-siang*, and *terdahulu*, via the offset value '812952'. The constructed multilingual network therefore contains a node that refers to three different pieces of information, namely, an English sense, the Malay translation, and the offset value, as shown in Table 2.

TABLE 2. Table of sample nodes, each with an offset value, an English sense and the corresponding Malay translation.

Offset	English	Malay
1740	able	*boleh*
2312	abaxial	*abaksial*
2098	unable	*tidak boleh*
4296	last	*akhir*
4615	cut	*memotong*

The focal concern is on adjectives, since they are modifier terms that express judgment on nouns, and are often used to express sentiment and affective states. According to Esuli and Sebastiani (2006), about 35.7% of adjectives, and 39.6% of adverbs, are at least partially subjective and express sentiment, followed by only 9.98% of nouns, and 11.04% of verbs. Consequently, only the adjectives are extracted from WNB, and then mapped onto their corresponding English versions. This generates an adjectives subgraph, keeping all of the original lexical relations and paths of the English WordNet intact. The final adjectives subgraph comprises a total of 10,716 senses and 7,371 unique terms.

Adjectives and adverbs have a different structure than that of nouns and verbs in WordNet. They are organized in adjective clusters rather than in a taxonomy, and the main relations between them is direct antonymy. In an adjectives cluster, head adjectives come in symmetrical bipolar pairs, and have contrasting relations (i.e. they are direct antonyms of each other). In the case of Figure 1, *wet* and *dry* make up the head adjectives in the cluster. Each one of the bipolar adjectives is a head adjective in its half-cluster, and is surrounded by satellite adjectives that are synonymous to it (e.g. arid (*gersang*) is a synonym of dry (*kering*)). Since a head adjective is a direct antonym to the adjective in the opposing half-cluster, it is also an indirect antonym to each of the satellite adjectives in the opposing half cluster (e.g. arid (*gersang*) is an indirect antonym of wet (*basah*)).

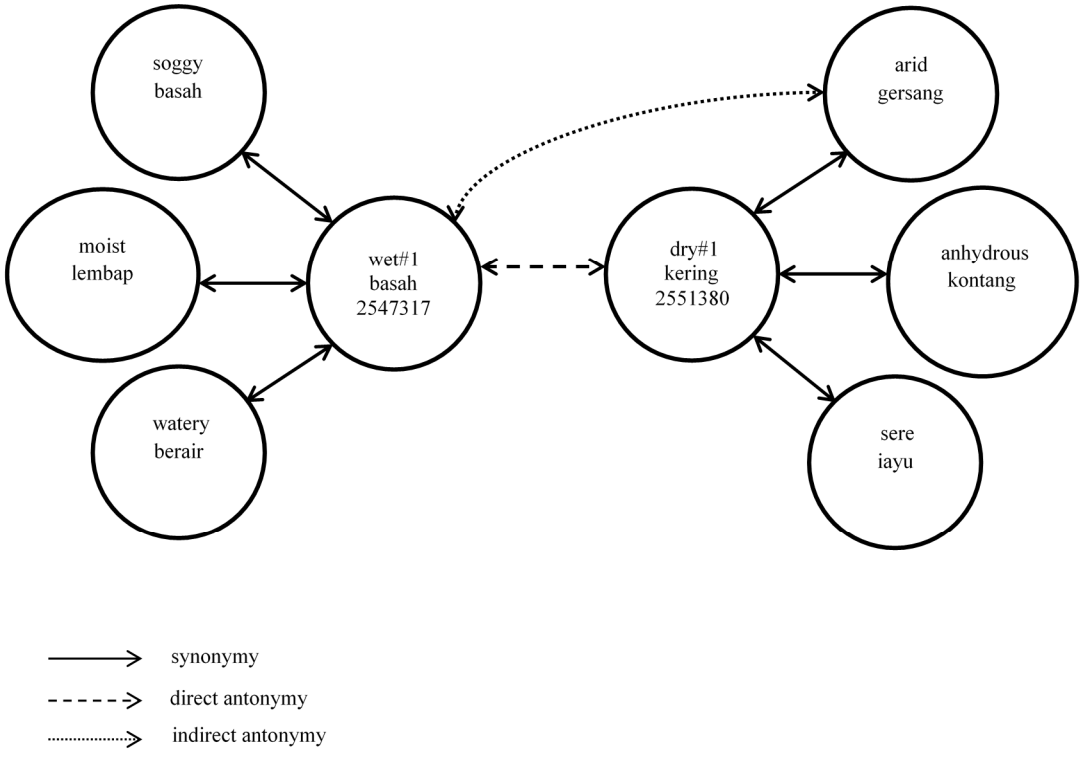

FIGURE 1. Example adjectives cluster for wet and dry

SEED SET

The point of an automatic means of labelling words with a polarity is to minimize human involvement. The seed sets $S^+ = \{baik\}$ and $S^- = \{buruk\}$ were used to represent the positive and negative classes respectively ($S_i = S_i^+ \cup S_i^-$), where i represents the number of iterations of WordNet propagation. The adjectives '*baik*' and '*buruk*' in Malay translate to an English equivalent of 'good' and 'bad' respectively.

This algorithm is unsupervised in that it uses an initial seed set of only two words to *define* the positive and negative classes, rather than to train the classification model. The training data is generated automatically by means of applying the seed set to propagate through WordNet relations, and there is no human involvement associated with this step.

WORDNET SYNONYMY AND ANTONYMY PROPAGATION ALGORITHM

The seed set is used to propagate through WordNet synonymy and antonymy relations. The underlying intuition is that synonymous words do not only have similar meanings, but also generally have similar semantic orientations. Antonyms have opposing meanings, hence, opposing semantic orientations. Starting with the initial seed set (S_0) of words with a labelled semantic orientation, and propagating through WordNet's synonymy relations, the sentiment polarity of words is preserved. In contrast, propagating through antonymy relations flips a word's polarity. This allows for automatically assigning unseen adjectives with a polarity.

For a seed word in the positive set, after one iteration, all of its synonyms are also added to the positive set (S_i^+), while all of its antonyms are added to the negative set (S_i^-). For a seed word in the negative set (S_i^-), all of its synonyms are also added to the negative set, while all of its antonyms are added to the positive set. Figure 2 depicts the algorithm used for WordNet Propagation.

```
WordNet Synonym and Antonym Propagation Algorithm
Input:  Si+, Si-
Output: Si+ expanded,  Si- expanded
Begin
For entry ei in Si+:
     Si+ <- Si+ + synonyms(ei)
     Si- <- Si- + antonyms(ei)
End For
For entry ei in Si-:
     Si- <- Si- + synonyms(ei)
     Si+ <- Si+ + antonyms(ei)
End For
Si+ expanded <- Si+
Si- expanded <- Si-
End
```

FIGURE. 2. WordNet Synonym and Antonym Propagation Algorithm

After four iterations for the positive set, a total of 1000 words were generated ($S_4^+ = 1000$), with 876 via the synonymy relation and 124 via the antonymy relation. For the negative set, a total of 1156 words were generated ($S_4^- = 1156$), with 1020 words via the synonymy relation and 136 via the antonymy relation.

After five iterations for the positive set, a total of 2139 words were generated ($S_5^+ = 2139$), with 1858 words via the synonymy relation and 281 via the antonymy relation. For the

negative set, a total of 2281 words were generated (S_5^- = 2281), with 1974 words via the synonymy relation and 307 via the antonymy relation.

After propagation using five iterations, the tagged words were used as a means to train a classification model in the next step. It is worthy to note that the training data has been automatically generated, with no human involvement. This data is used to train a classifier to label unseen adjectives with a positive or a negative polarity.

BINARY WORD-POLARITY CLASSIFICATION

The words labelled by the propagation algorithm were used to train a classifier to label unseen words with a polarity. Since a synset contains words that all refer to the same meaning or concept, they generally have the same semantic orientation. For simplicity, for an unseen target term to be classified in the positive or the negative category, if it has synonyms that appear more often in the positive class, and antonyms that appear more often in the negative class, it is classified as positive. Otherwise, it is classified as negative. Pre-processing or cleaning of data is not carried out, since the training data contains single isolated terms only, as opposed to larger pieces of text, for which additional pre-processing steps such stemming may increase performance.

A binary naïve Bayes (NB) classifier was employed for the classification task. This type of classifier abides Bayes' theorem, which states that:

$$P(C \mid f_1,..., f_n) = \frac{P(C)P(f_1,..., f_n \mid C)}{P(f_1,..., f_n)} \quad (1)$$

where P represents the probability of class C, given a features vector composed of features $f_1,..., f_n$ (McCallum & Nigam 1998).

This is simplified to become:

$$P(C \mid f_1,..., f_n) = \frac{P(C)\prod_{i=1}^{n} P(f_i \mid C)}{P(f1,..., f_n)} \quad (2)$$

Since $P(f1,..., f_n)$ remains constant given any input, it is discarded, and thus becomes:

$$P(C \mid f_1,...f_n) \propto P(C)\prod_{i=1}^{n} P(f_i \mid C) \quad (3)$$

This can be reformulated as:

$$C_{\text{MAP}} = \underset{C \in C_{\text{all}}}{\operatorname{argmax}}\, P(C)\prod_{i=1}^{n} P(f_i \mid C) \quad (4)$$

where C_{MAP} is 'maximum a posteriori', or the most probable class for a given input. The prior probability of a class $P(C)$ is computed by dividing the count of words in that current class by the count of words in all classes.

$$\hat{P}(f_i \mid C) = \frac{(occ(f_i), C)}{count(C)} \quad (5)$$

where $(occ(f_i), C)$ is a binary value denoting the occurrence of feature f_i in C, and $count(C)$ is the total number of occurrences of f_i in class C.

The NB classifier mentioned above is applied in this model as follows:

$$\underset{C_{\text{Polarity}}}{\operatorname{argmax}}\, P(C_{Polarity} \mid w) = \underset{C_{\text{Polarity}}}{\operatorname{argmax}}\, P(C_{Polarity})P(w \mid C_{Polarity}) \quad (6)$$

where $C_{Polarity}$ is representative of the positive class or the negative class.

Furthermore, every likelihood of an unseen word $P(w \mid C_{Polarity})$ is computed by:

$$P(w \mid C_{Polarity}) = P\left(\frac{\sum_{i=1}^{n} occ(synmember_i, C)}{total(C)} \right) \quad (7)$$

where $occ(synmember_i, C)$ is the occurrence of each synonym member of the word *w* in class *C*, and $total(C)$ is the total number of words in *C*. Each unseen word is represented as a binary features vector, where each binary value denotes the occurrence (or absence) of each of the synonyms of the unseen word.

Using this model, all of the remaining words in the extracted adjectives subgraph that were not tagged by the previous propagation step were classified to generate the resultant sentiment lexicon. Table 3 shows a sample of the final lexicon.

TABLE 3. Sample of the final lexicon.

Positive Words	Negative Words
kekuasaan	*ceroboh*
berhasil	*hilang tenaga*
khusus	*teruk sekali*
ketenangan	*tidak senang hati*
puas	*membahayakan*
bersusila	*menjadikannya tidak tulen*
perbuatan yang baik	*dengan sedih*

The lexicon was manually cleaned to filter out objective (i.e. neutral) adjectives, or adjectives that do not carry a polarity (e.g. triangular). This final step of manual cleaning is necessary, since there is no means to filter out objective adjectives by the algorithm, an issue investigated in future work.

RESULTS AND DISCUSSION

The GI was manually constructed by categorizing terms based on their semantic properties. It comprises a total of 1,915 words marked as 'Positiv' (sic) and 2,291 words marked as 'Negativ' to represent entries labelled as having positive and negative polarities respectively. A term may be included in multiple categories, each one denoting a particular trait associated with it. Due to its reliability, the GI has been used as a gold standard for benchmarking purposes in the majority of works in this area. Both the accuracy of WordNet propagation, and the accuracy of the word-polarity classification model, are measured separately by comparing the generated set of words of each against the intersecting words in the GI. Even though the GI is in English, it could be used for evaluation, since the Malay word synsets from WordNet Bahasa in this experiment are aligned to their English translations in English WordNet via synset offset values.

WORDNET PROPAGATION RESULTS

Table 4 shows the yielded accuracy for the synonymy and antonymy propagation algorithm after four and five iterations. The number of intersecting words between the GI and the words generated after four iterations (S_4) was 524. The synonymy and antonymy propagation algorithm after four iterations yielded an overall accuracy of 0.643. It yielded an accuracy of 0.546 to label positive words only, and an accuracy of 0.727 to label negative words only. After

five iterations (S_5), the number of intersecting words against the GI was 942. The algorithm yielded an overall accuracy of 0.567. It yielded an accuracy of 0.482 to label positive words only, and an accuracy of 0.644 to label negative words only.

TABLE 4. WordNet propagation algorithm results

Iterations	# of words generated (pos/neg)	Intersecting words against GI	Overall Accuracy	Accuracy (pos only)	Accuracy (neg only)
S_4	1000/1156	524	0.643	0.546	0.727
S_5	2139/2281	942	0.567	0.482	0.644

This demonstrates that, as the number of iterations increases to automatically label words with a semantic orientation through synonymy and antonymy relations, the number of words generated increases as well. However, coverage increases only at the cost of accuracy. According to the table, although there is an increase in coverage, the overall accuracy drops by 7.5% (from 0.643 to 0.567) when using five expansions (S_5) instead of four (S_4). The reason for the decrease in accuracy is that the semantic association of words becomes weaker as the distance between them increases.

WORD-POLARITY CLASSIFICATION RESULTS

This model was trained on a training set formulated using the five level synonymy and antonymy relations propagation. The test set was formulated using the intersecting words between the remaining words in the adjectives subgraph (i.e. words not picked up after WordNet propagation) and the words in the GI. This is to ensure that the test set is independent and does not overlap with the training set. The remaining words that were not part of the training data were about 7,600. The amount of intersecting GI words with the remaining words was 1,051 (540 with the positive words in the GI, and 511 with the negative).

After evaluation, the word-polarity classification model yielded an overall accuracy of 0.720, indicating that it is able to label words with 7.7% more accuracy than the four level WordNet propagation (0.643), and with 15.3% more accuracy than five level propagation (0.567). This demonstrates that its ability to accurately label words relies on the quality the training data used. Since the four level propagation generated higher quality data compared to five level propagation, this has helped to better train the classifier. The five level propagation seems to have generated an increased volume of noise (i.e. mislabelled adjectives), resulting in poor training of the classifier. There is no previous work on Malay sentiment lexicon generation algorithms, hence the absence of a baseline to allow for a feasible comparison.

CONCLUSION

This paper presented an automatic sentiment generation algorithm specifically for the Malay language. WordNet Bahasa was first mapped onto the English version of WordNet to construct a multilingual word network, and a dictionary and a supervised classification model were employed for tagging words with a positive or negative polarity. Evaluation of the algorithm against the manually annotated General Inquirer lexicon demonstrates that it performs with reasonable accuracy. The contribution of this paper toward the development of SA models that are specifically constructed for the Malay language is twofold. First, it provides a foundation for further progress on sentiment lexicon generation algorithms in this target language. Second, it defines a baseline that can be used as a benchmark in future work.

For future work, the proposed algorithm may be improved by considering other word classes such as nouns and verbs, which also carry sentiment. Another task is to incorporate a

filtering mechanism for objective adjectives that do not carry any sentiment (i.e. a multi-class classifier with an objective class). Along with synonyms of an input word as features for the classifier, augmenting the gloss of the input words may further enrich the training data with beneficial cues for the classification model, since a subjective word may also contain subjective words within its gloss.

A different direction may be to use a combination of a dictionary and a corpus in the target language for polarity classification, which may exploit the benefits of both approaches, and in turn improve performance. The main aim in this work is not to develop a state-of-the-art sentiment lexicon generation algorithm, but rather to provide a baseline for future work in this area.

REFERENCES

Andreevskaia, A. & Bergler, S. 2006. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *European Chapter of the ACL*, vol. 6: 209-216.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A. & Reynar, J. 2008. Building a sentiment summarizer for local service reviews. *WWW Workshop on NLP in the Information Explosion Era*, vol. 14.

Bond, F., Fellbaum, C., Hsieh, S. K., Huang, C. R., Pease, A. & Vossen, P. 2014. A multilingual lexico-semantic database and ontology. *Towards the Multilingual Semantic Web*, 243-258. Berlin Heidelberg: Springer.

Dragut, E., Wang, H., Yu, C., Sistla, P. & Meng, W. 2012. Polarity consistency checking for sentiment dictionaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol. 1: 997-1005. Stroudsburg PA: Association for Computational Linguistics (ACM).

Esuli, A. & Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of Language Resources and Evaluation Conference*, vol. 6: 417-422. Genoa, Italy:[s.n.].

Fellbaum, C. 1998. *WordNet*. Wiley Online Library. Blackwell Publishing Ltd.

Hassan, A. & Radev, D. 2010. Identifying text polarity using random walks. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 395-403. Stroudsburg, PA, USA: ACM.

Hatzivassiloglou, V. & McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics*, 174-181. Stroudsburg, PA, USA: ACM.

Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data mining*, 168-177. Seattle, WA, USA: ACM.

Kamps, J., Marx, M., Mokken, R. J. & De Rijke, M. 2004. Using WordNet to Measure Semantic Orientations of Adjectives. *Language Resources and Evaluation Conference*, vol. 4: 1115-1118. Lisbon, Portugal: UvA-DARE.

Kim, S. M. & Hovy, E. 2004. Determining the sentiment of opinions. *Proceedings of the 20th international conference on Computational Linguistics*, 1367. Stroudsburg, PA: Association for Computational Linguistics (ACM).

McCallum, A., & Nigam, K. 1998. A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, vol. 752: 41-48..

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, vol. 3(4): 235-244.

Mohammad, S., Dunne, C. & Dorr, B. 2009. Generating high-coverage semantic orientation lexicons from ovetly marked words and a thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 2: 599-608. Singapore: ACM.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. 2009. Sentiful: Generating a reliable lexicon for sentiment analysis. *Affective Computing and Intelligent Interaction and Workshops ACII 2009.*

3rd International Conference on. Institute of Electrical and Electronics Engineers. Amsterdam, Netherlands, 1-6. Sept.
Noor, N. H. B. M., Sapuan, S. & Bond, F. 2011. Creating the Open Wordnet Bahasa. In *PACLIC*, 255-264. Singapore:[s.n.].
Peng, W. & Park, D. H. 2011. Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization. *Fifth International AAAI Conference on Weblogs and Social Media*. Urbana, Illinois.
Rao, D. & Ravichandran, D. 2009. Semi-supervised polarity lexicon induction. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 675-682. Stroudsburg, PA: ACM.
Stone, P. J., Dunphy, D. C. & Smith, M. S. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Oxford: M.I.T. Press.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, vol. 37(2): 267-307.
Takamura, H., Inui, T. & Okumura, M. 2005. Extracting semantic orientations of words using spin model. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 133-140. Stroudsburg, PA, USA: ACM.
Turney, P. D. & Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, vol. 21(4): 315-346.
Tze, L. L. & Hussein, N. 2006. Fast prototyping of a malay wordnet system. *Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing Applications (LAICS-NLP) Summer School Workshop*, 13-16. Bangkok, Thailand.
Williams, G. K.,& Anand, S. S. 2009. Predicting the polarity strength of adjectives using WordNet. *Proceedings of the Third International ICWSM Conference*. Menlo Park, California: Association for the Advancement of Artificial Intelligence (AAAI) Press.

Mohammad Darwich
Shahrul Azman Mohd Noah
Nazlia Omar
Center for Artificial Intelligence Technology
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia
modarwish@hotmail.com, shahrul@ukm.edu.my, nazlia@ukm.edu.my