INDUCING A DOMAIN-INDEPENDENT SENTIMENT LEXICON IN MALAY

Mohammad Darwich[1], Shahrul Azman Mohd Noah[2], Nazlia Omar[3]
[1,2,3]*Center fo Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia*
modarwish@hotmail.com, shahrul@ukm.edu.my, nazlia@ukm.edu.my

Abstract: Sentiment analysis (SA) is a discipline that involves the detection of user sentiment, emotion and opinion within natural language text. Lexicon-based SA models make use of a sentiment lexicon for SA tasks, which is a linguistic resource that comprises a priori knowledge about subjective words tagged with their underlying sentiment polarity. A sentiment lexicon and greatly contributes to SA tasks. This is evident in the emergence of the large number of research works that have aimed to develop automated sentiment lexicon induction algorithms. However, most works primarily consider the English language; this is attributable to the availability of a sufficient amount resources and tools for this language. On the other hand, this is not the case for low-resource languages such as Malay. Research focused on sentiment lexicon induction algorithms in particular, and SA in general, in the Malay language, is lacking. This has brought up the motivation to develop a sentiment lexicon induction algorithm for the Malay language. We first map WordNet Bahasa onto the English WordNet to construct a multilingual word network, and then use a dictionary-based approach and a supervised classifier for classifying words with their sentiment polarities. The algorithm was evaluated against the General Inquirer lexicon, demonstrating that it performs with accuracy that is comparable to human accuracy.

Keywords: Sentiment analysis, opinion mining, sentiment lexicon, subjectivity detection

INTRODUCTION

Sentiment analysis (SA), or opinion mining (OM), is a discipline that involves the detection of user sentiment and opinion embedded within natural language text. It has been a rapidly emerging area of focus during the past several years, and has been significantly driven by practical application within almost all domains, by commercial and non-commercial organizations.

In principle, an SA model involves determining whether a document carries a positive or negative sentiment polarity, or no polarity at all. There are two main approaches utilized by SA models. The first is the (unsupervised) lexicon-based approach, which involves employing a sentiment lexicon to compute the overall polarity (or *semantic orientation*) of a document based on the polarities of independent word matches within the document (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). The second is the (supervised) classification-based approach, which involves employing a supervised classifier that is provided with manually labelled training data for the sentiment classification task. Current state-of-the-art SA models sometimes use hybrid of both the lexicon-based and the classification-based approaches in combination (e.g. use the a priori information about the polarity of individual sentiment words within a document as features to train a classifier), with the aim of exploiting the benefits of both approaches (Peng & Park, 2011).

SA models that use the lexicon-based approach make use of a sentiment lexicon, which is a linguistic resource that consists of sentiment words labelled with their underlying sentiment polarity. Information about sentiment words and phrases greatly contributes to SA tasks. This is evident in the emergence of an extensive number of research works pace attention on the SA subtask of tagging single words with a polarity. In principle, positive sentiment-bearing words convey desired emotive states, while negative sentiment-bearing words convey undesired emotive states. A sentiment lexicon can be formulated in one of two ways: manually or automatically. Automated methods use either a dictionary or a corpus to generate a sentiment lexicon.

Manually tagging words to induce a sentiment lexicon is prohibitively costly in terms of annotator time and effort, and often results in subjective bias in annotation, since the judgment of annotators varies to a particular degree (Andreevskaia & Bergler 2006; Dragut, Wang, Yu, Sistla, & Meng, 2012). The increasing popularity of sentiment analysis has resulted in the high demand of automatic sentiment lexicon induction algorithms that require minimal human intervention.

A large volume of research works in the literature have focused on the induction of automatic sentiment lexicons. Automated lexicon induction algorithms take two directions: (1) the dictionary-based approach and (2) the corpus-based approach. The intuition that lies in making use of an online dictionary is that words are not only semantically related in terms of meaning, but to a certain extent, are related in terms of their sentiment properties as well. An online dictionary in this approach takes the role of a semantic, lexical knowledge base that has extensive coverage of words defined within a natural language. Kamps et al. (2004) and Williams and Anand (2009) propose WordNet distance-based semantic similarity measures to tag words with their underlying sentiment polarity. Hu and

Liu (2004) proposed a bootstrapping algorithm that uses an initial set of manually labeled seed words and WordNet synonym and antonym semantic relations for this task. The occurrence of a synset's synonym members (Kim and Hovy 2004) and gloss information (Esuli and Sebastiani 2006b) in WordNet were used as features for supervised classification. WordNet subgraphs that use label propagation were exploited (Rao and Ravichandran 2009; Blair-Goldensohn 2008). Hassan and Radev (2010) proposed a Monte Carlo random walk model in which seed words played the role of absorbing boundaries. Morphological (affix) features of terms were exploited to automatically derive new terms, while preserving the sentiment features of the original (Mohammad et al. 2009; Neviarouskaya 2009).

Conversely, the corpus-based approach uses statistical and syntactic patterns in text corpora, as well as a seed set of manually annotated words for marking words with a sentiment polarity. Peng and Park (2011) used both a dictionary and a corpus extracted from a social media platform in combination to mark words with a polarity and strength. The entire Web was used as a corpus to compute the polarity of a word based on the condition that its co-occurrence with a group of positive seed words is greater than its co-occurrence with a group of negative seed words (Turney and Littman 2003). An approach that extracts adjective-pairs conjoined by the coordinating conjunctions *and* and *but* from a corpus was proposed; the intuition is that conjoined adjectives abide to linguistic constraints, where *and* conjoins an adjective pair of equal polarities, while *but* conjoins an adjective pair with opposing polarities (Hatzivassiloglou and mcKeown 1997). The research works mentioned generally focus on the English language. There is no prior work that focuses on automatically marking words in the Malay language with a polarity, hence the motivation to propose an algorithm specifically tailored for the this target language.

We utilize the dictionary-based approach for this task, and map WordNet Bahasa (WNB) onto the English WordNet to construct a multilingual word network. Our approach is minimally supervised in that we utilize a seed set comprising a bipolar adjective pair, and automatically propagate via WordNet's synonymy and antonymy relations to automatically tag words with their underlying sentiment polarity. We then use the expanded seed set to train a ternary (positive-negative-objective) classifier to classify the remaining words that were not picked up via the WordNet propagation process. The proposed algorithm is evaluated against the intersecting words in the General Inquirer (GI) lexicon (Stone, Dunphy, & Smith, 1966) to demonstrate that it performs relatively good accuracy.

METHODS

We present the proposed model that automatically tags words in the Malay language with a polarity. The methodology is conducted in multiple stages. In the first step, we map WordNet Bahasa (WNB) onto the English WordNet 3.0 to construct a Malay word network, while preserving the semantic relations between words in the process. This is followed by constructing a seed set of manually annotated sentiment-bearing words to propagate through the constructed network and automatically tag words using WordNet synonym and antonym relations. The final expanded seed set is employed to train a ternary (positive-negative-objective) classifier to classify the remaining words that were not classified via network propagation in the previous step.

Mapping WordNet Bahasa to English WordNet

Since the proposed model uses the dictionary-based approach, and focuses on the Malay language, there is a requirement for a lexical knowledge in this target language. WordNet Bahasa (WNB; Noor, Sapuan, & Bond, 2011) is the formally standardized Malay version of WordNet. It comprises 49,668 synsets, 145,696 unique senses and 64,431 terms. We map the Malay and Bahasa version senses to the English WordNet senses using their offset values. 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. We extract only the adjectives from WNB, and map them onto their linked English versions. This induces an adjectives subgraph, keeping all of the original lexical relations and paths of WordNet intact. The final adjectives subgraph consists of 10,716 senses and 7,371 terms.

Seed Set

We use the seed sets: Sp = {*baik, bagus, cemerlang, positif, bernasib baik, betul, unggul*} and Sn = {*buruk, jahat, miskin, negatif, malang, salah, rendah*} to define the positive and negative classes respectively ($S_i = S_i^+ \cup S_i^-$), where i represents the number of iterations of WordNet propagation. This seed set is adopted from Turney and Littman's (2003) equivalent English seed sets: Sp = {*good, nice, excellent, positive, fortunate, correct, superior*} and Sn = {*bad, nasty, poor, negative, unfortunate, wrong, inferior*}.

WordNet Synonym and Antonym Propagation Algorithm

We use the seed set 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 labeled semantic orientation, and propagating through WordNet's synonymy relations, the semantic orientation properties of words are preserved. In contrast, propagating through antonymy relations flips a word's semantic orientation. 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, all of its synonyms are also added to the negative set, while all of its antonyms are added to the positive set. The objective class S_i^O is formulated by adding to it all of the terms not included in the expanded positive or negative seed sets. Since this set is extensively larger than the positive and negative seed sets, we trim it down to about 1000 terms.

Ternary Positive-Negative-Objective Classification

The words labelled by the propagation algorithm were used to train a classifier to label unseen words with a polarity. We do not carry out pre-processing or cleaning of data, since the training data contains single terms only, as opposed to larger pieces of text, for which additional pre-processing steps such stemming may increase performance. For features extraction, for each word sense, we extract all of its synonym members out of its synset, and insert them into the corresponding class. Since a synset contains words that all refer to the same meaning or concept, they generally have the same semantic orientation. This enriches the training data used to train the learner model compared to using the individual senses alone. We use a multiclass naïve Bayes classifier for the classification task, which can be defined as follows:

$$\underset{\mathrm{C_{Polarity}}}{\operatorname{argmax}} P(C_{Polarity} \mid w) = \underset{\mathrm{C_{Polarity}}}{\operatorname{argmax}} P(C_{Polarity}) P(w \mid C_{Polarity})$$

Where $C_{Polarity}$ is either the positive class, the negative class or the neutral class.

Using this model, we classify all of the remaining words to generate the resultant lexicon, in which all of the adjectives in the extracted adjective subgraph are labelled with a polarity. The findings in the following section demonstrate the effectiveness of this model to generate a Malay sentiment lexicon.

FINDINGS AND ARGUMENT

The GI was manually annotated by classifying terms based on their semantic characteristics. It consists of 1,915 words categorized as 'Positiv' and 2,291 words categorized as 'Negativ' to denote terms marked as carrying positive and negative polarities respectively. The remainder of the terms were not labelled neither as positive nor as negative, and thus we define these terms as the objective test set (a total of 7,583 terms). A term may be included in several categories, each one defining a certain trait associated with the term. The GI has been employed as a reliable benchmark in many of works in this field. Both the accuracy of WordNet propagation and the accuracy of the word-polarity classification model are estimated independently by comparing the resultant set of words of each algorithm against the intersecting words in the GI.

WordNet Propagation Results

After a total of three iterations, 2779 unique terms were generated in the positive set, and 931 unique terms were generated in the negative seed set. The synonymy and antonymy propagation algorithm after four iterations yielded an accuracy of 0.574 to label positive terms, while it yielded an accuracy of 0.645 to label negative terms. The overall accuracy of this algorithm is 0.61. This demonstrates that using synonymy and antonymy relations is not enough. If augmented by using terms' gloss information, this could potentially increase accuracy. For example, for multiple iterations, if a term is found in the glosses of the terms in the positive (resp. negative) seed set, that term can also be added to the positive (resp. negative) set.

Ternary Positive-Negative-Objective Classification Results

The classifier achieved an accuracy of 0.894 overall. This demonstrates that the classifier is able to label words as positive, negative and neutral with an accuracy that outperforms that of humans, which is about 82% (Wilson et al. 2005). This demonstrates that its ability to accurately label words greatly relies on the quality the training data used. Since we only use three iterations for WordNet expansion, this provide useful training data with minimal noise, since the closer the distance between words in WordNet, the stronger their semantic relations. It is important to note that a term's gloss information may also be used as features for a classifier, which may potentially improve the accuracy. This is a direction to take in further research.

CONCLUSIONS

This study proposed an automated sentiment induction algorithm for the Malay language. We first mapped WordNet Bahasa onto the English WordNet to formulate a multilingual word network, and then used a dictionary-based approach and a supervised classifier to mark words with a positive, negative or neutral polarity. Evaluation of the algorithm against the manually annotated General Inquirer lexicon demonstrates that it performs with reasonable accuracy. The contribution of this paper to the development of SA models that are specifically constructed to work with 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. We plan to further expand the proposed algorithm by considering other word classes such as nouns and verbs, which also carry sentiment. Along with synonyms of an input word as features for the classifier, augmenting the gloss of the input words may 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 in the target language such as WNB and a corpus-based approach for polarity classification, which may exploit the benefits of both approaches, and in turn improve performance.

REFERENCES

Andreevskaia, A., & Bergler, S. (2006, April). Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In *EACL* (Vol. 6, pp. 209-216).

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

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

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417-422).

Hassan, A., & Radev, D. (2010, July). Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 395-403). Association for Computational Linguistics.

Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In *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* (pp. 174-181). Association for Computational Linguistics.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In*Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004, May). Using WordNet to Measure Semantic Orientations of Adjectives. In *LREC* (Vol. 4, pp. 1115-1118).

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

Mohammad, S., Dunne, C., & Dorr, B. (2009, August). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2* (pp. 599-608). Association for Computational Linguistics.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009, September). Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1-6). IEEE.

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

Peng, W., & Park, D. H. (2011, May). Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Rao, D., & Ravichandran, D. (2009, March). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 675-682). Association for Computational Linguistics.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315-346.

Williams, G. K., & Anand, S. S. (2009). Predicting the polarity strength of adjectives using wordnet.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... & Patwardhan, S. (2005, October). OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations* (pp. 34-35). Association for Computational Linguistics.