

---

# INDUCING A DOMAIN-INDEPENDENT SENTIMENT LEXICON IN MALAY



Mohammad Darwich, Shahrul Azman Mohd Noah, Nazlia Omar  
Center fo Artificial Intelligence Technology (CAIT),  
Faculty of Information Science and Technology (FTSM),  
Universiti Kebangsaan Malaysia, Bangi, Selangor (UKM), Malaysia

# Introduction

---

Sentiment analysis (SA) is a discipline that involves the detection of user sentiment, emotion and opinion within natural language text.

Popular in important domains such as commercial, financial and governmental

Lack of resources for this task in non-English languages such as Bahasa Malaysia

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# Sentiment Analysis

---

SA model involves determining whether a document carries a positive or negative sentiment polarity, or no polarity at all.

two main approaches:

- 1) (unsupervised) lexicon-based approach: involves employing a sentiment lexicon to compute the overall polarity of a document (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011)
- 2) (supervised) classification-based approach, which involves a supervised classifier provided with manually labelled training data

# Lexicon-Based Sentiment Analysis

---

Lexicon-based SA models make use of a sentiment lexicon for SA tasks

Lexicon:

- a linguistic resource that comprises a priori knowledge about subjective words tagged with their underlying sentiment polarity
- the most important element that impacts the performance of such lex-based models
- typically consists of subjective words that deviate from neutrality, towards positivity or negativity
- degree of deviation represents the intensity of a sentiment word

# Sentiment Lexicon

---

A sentiment lexicon can be formulated in one of two ways:

Manually or automatically

Manual compilation is a tedious task; entire span of terms in a language must be marked for optimal effectiveness

Automated methods use either a dictionary or a corpus to generate a sentiment lexicon.

# Sentiment Lexicon

---

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.

Benefits: a dictionary play the role of a semantic, lexical knowledge base that has extensive coverage of words defined within a natural language

Drawbacks: generates a domain independent lexicon only

# Related Work

---

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).

# Methodology

---

WordNet Bahasa (WNB; Noor, Sapuan, & Bond, 2011) is the formally standardized Malay version of WordNet.

We map the Malay and Bahasa version senses to the English WordNet senses using their offset values.

We extract only the adjectives from WNB, and map them onto their linked English versions.



# Methodology

---

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.

English translation:

**Sp = {good, nice, excellent, positive, fortunate, correct, superior}**

and

**Sn = {bad, nasty, poor, negative, unfortunate, wrong, inferior}**

# Methodology

---

WordNet Synonym and Antonym Propagation Algorithm: We use the seed set to propagate through WordNet synonymy and antonymy relations.

Intuition: synonymous words do not only have similar meanings, but also generally have similar semantic orientations, while antonyms have opposing meanings.

# Methodology

---

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 ( $S_i^-$ ). Same is applied for the negative set. This is iteratively run for three rounds (S3).

Objective class  $S_i^0$  is formulated by adding to it all of the terms not included in the expanded positive or negative seed sets.

# Methodology

---

The words labelled by the propagation algorithm were used to train a classifier to label unseen words with a polarity

No preprocessing

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

Ternary naïve Bayes classifier for classification, which can be defined as follows:

$$\operatorname{argmax}_{C_{Polarity}} P(C_{Polarity} | w) = \operatorname{argmax}_{C_{Polarity}} P(C_{Polarity}) P(w | C_{Polarity})$$

C = any one of three classes – positive, negative, objective

# Findings & Discussion

---

General Inquirer used as a gold standard: 1,915 positive words, 2,291 negative, and 7,583 objective terms. Intersecting words between words labelled by my proposed model, and GI words used to compute accuracy

The classifier achieved an accuracy of 0.894 overall.

This demonstrates that the classifier is able to label words with an accuracy that outperforms that of humans, which is about 82% (Wilson et al. 2005).

This indicates that the ability to accurately label words greatly relies on the quality of the training data used.

Since we only use three iterations for WordNet expansion, this provides useful training data with minimal noise, since the closer the distance between words in WordNet, the stronger their semantic relations.


However, accuracy is preferred over coverage in this case.



# Conclusion

---

We proposed an automated sentiment induction algorithm for the Malay language:

- Mapped WordNet Bahasa onto the English WordNet to formulate a multilingual word network
  - Used the WordNet Synonym and Antonym Propagation Algorithm and a supervised classifier to mark words with a polarity
  - Evaluation of the algorithm demonstrates that it performs with reasonable accuracy
- 

# Contributions

---

This work provides a foundation for further progress on sentiment lexicon generation algorithms in this target language.

It also defines a baseline that can be used as a benchmark in future work.

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# Future Work

---

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, since a subjective word may also contain subjective words within its gloss.

Also, this work only considered adjectives, but other word classes would provide for additional coverage in the lexicon such as nouns and verbs.

We plan to employ this lexicon, and the incorporation of in-context rules such as intensifiers and negation words, in a phrase labelling task to demonstrate its ability to classify full-texts based on polarity.



---

Thank you

