Journal of Computer Science

Original Research Paper

Integrating a Lexicon Based Approach and K Nearest Neighbour for Malay Sentiment Analysis

Ahmed Alsaffar and Nazlia Omar

Center for AI Technology, FTSM University Kebangsaan Malaysia, UKM 43000 Bangi Selangor, Malaysia

Article history
Received: 06-05-2015
Revised: 10-06-2015
Accepted: 16-06-2015

Corresponding Author:
Ahmed Alsaffar
Center for AI Technology,
FTSM University Kebangsaan
Malaysia, UKM 43000 Bangi
Selangor, Malaysia
Email: ahmed_saffar5@yahoo.com

Abstract: Sentiment analysis or opinion mining refers to the automatic extraction of sentiments from a natural language text. Although many studies focusing on sentiment analysis have been conducted, there remains a limited amount of studies that focus on sentiment analysis in the Malay language. In this article, a new approach for automatic sentiment analysis of Malay movie reviews is proposed, implemented and evaluated. In contrast to most studies that focus on supervised or unsupervised machine learning approaches, this research aims to propose a new model for Malay sentiment analysis based on a combination of both approaches. We used sentiment lexicons in the new model to generate a new set of features to train a k-Nearest Neighbour (k-NN) classifier. We further illustrated that our hybrid method outperforms the state of-the-art unigram baseline.

Keywords: Malay Sentiment Analysis, Feature Extraction, Machine Learning, Combinations Techniques

Introduction

Opinions are playing a primary role in decision-making processes. Whenever people need to make a choice, they are naturally inclined to hear others' opinions. In particular, when the decision involves consuming valuable resources, such as the time and/or money, people strongly rely on their peers' past experiences. On the other hand, customers could also learn about positivity or negativity of different features of products/services according to users' opinions, to make an educated purchase. Furthermore, applications like rating movies based on online movie reviews (Pang *et al*., 2002) could not emerge without making use of these data.

The topic of sentiment analysis has become extremely popular in the last couple of years. There has been a tremendous amount of research on this topic. There are several names for this topic, including opinion mining and sentiment classification. Generally, sentiment analysis is a unique case of text classification, which aims to classify sentiments for subjective texts, usually customer reviews for some product or service.

The organizations are looking for opportunities to analyze the personal opinions that are gathered online about their services and products to develop their businesses outcomes. However, there is difficulty in classifying the large volume of online users' information in order to reflect the users' opinions accurately. Additionally, the users' express their opinions based on free texts i.e., unstructured methods which maximize the difficulty of analyzing the opinions polarity from these texts (Puteh *et al*., 2013).

The majority of studies concerns with analyzing the users' opinions based on English language. There has been a very limited amount of research that focuses on sentiment analysis in the Malay language (Samsudin *et al*., 2013).

The main goal of this work is to identify an optimized set of features that enhance the Malay sentiment analysis and classifications. We consider the bag-of-words (unigrams) as a baseline for sentiment classification. We train the k-Nearest Neighbour (k-NN) classifier based on the unigram feature set and compare them against our new proposed model which combines lexicon knowledge and a supervised machine learning approach for Malay sentiment analysis and classification.

There are multiple approaches to sentiment analysis (SA), which may be separated into three main categories: Firstly, supervised machine learning approach that has been implemented in numerous studies (Balahur *et al*., 2014; Pang *et al*., 2002; Greaves *et al*., 2012; Kang *et al*., 2012; Turney, 2002) Secondly, unsupervised machine learning approach is also a popular technique for sentiment analysis (Gezici *et al*., 2013).

Science Publications

On the other hand, combination of both approaches. There is a limited amount of research that has been carried out in order to extract opinions from posts that are available online in the Malay language.

Supervised Machine Learning Approach

When it comes to classification using Machine Learning (ML), there are usually two different collections of documents that are needed, which are a training collection and a test collection. The training collection is used by the classifier in order for it to learn how to differentiate between the different features of the text. The test collection is used in order to estimate the overall efficiency of the classifier. Pang *et al*. (2002) were one of the very first to perform SA on online English movie reviews. They tested ML approaches, namely Naïve Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (ME) classifiers and trained them on different feature sets including unigrams. Their findings showed that an SVM trained on a unigram bag-of-words feature set outperforms all other approaches presented in their work. Samsudin *et al*. (2012) reported the use of the Artificial Immune System (AIS) technique in classifying Malaysian online movie reviews into negative and positive sentiments. The AIS performance was compared with other traditional machine learning techniques. The performance of other conventional text classification techniques to mine opinion such as the Naïve Bayes, k-Nearest Neighbour and Support Vector Machine were better than the performance of the AIS. Isa *et al*. (2013) studied the effect of pre-processing methods for Malay sentiment classification texts. The researcher also used machine learning by Artificial Immune Network (AIN) for the sentiment classification of newspaper articles in the Malay language.

Unsupervised Machine Learning Approach

Unsupervised machine learning approaches are mostly based on a Sentiment Lexicon (SL), in which each sentiment-bearing word is associated with either a sentiment score or a set of sentiment-bearing seed words. These methods use different algorithms to compute a sentiment score for a given document. Zamani *et al*. (2013) studied how to identify opinion mining and sentiment analysis components to extract both English and Malay words in Facebook. Their work focused on quantifying Facebook sentiments using a lexicon-based approach for both English and Malay texts. Liau and Tan (2014) used a lexicon-based approach to study consumer opinion towards the low-cost airlines or Low-Cost Carriers (LCCs) industry in Malaysia. Zamani *et al*. (2013) and Liau and Tan (2014) created a lexicon that can be considered fairly small due to the number of sentiment words it consists of. Also, they used a very simple classifier to calculate the polarity of each comment.

Combination of Supervised and Unsupervised Approach

In this approach, a combination of both of the above mentioned categories was adopted to perform opinion mining. In most of such systems, the sentiment lexicon is used to capture the sentiment features in the document. The process converts each review into a vector of feature values by means of the sentiment lexicon. Then these features are used for training a machine learning classifier. In fact, the use of the lexicon based approach followed by a machine learning classification can be regarded as a hybrid method (Patil and Deshmukh, 2014). One combined approach was done by Liu *et al*. (2004). They started with two English word lexicons and unlabelled data. With the two discriminatory-word lexicons (negative and positive), they created pseudo-documents containing all the words of the chosen lexicon. After that, they computed the cosine similarity between these pseudo-documents and the unlabelled documents. Based on the cosine similarity, a document is assigned either a positive or negative sentiment. Then they used these to train a Naive Bayes classifier. Another combined approach was done by Agarwal *et al*. (2011); they adopted an English sentiment lexicon to count the number of occurrences of positive and negative words in tweets and used these two measures along with other features to train their classifier. However, no previous research in Malay sentiment classification has attempted to combine both machine learning and Malay sentiment lexicon approaches to take advantage of the benefits of both approaches. Therefore this research aims to focus on analysing the Malaysian users' opinions by using a combined approach.

Materials and Methodology

The dataset used in this research was gathered from several online forums and blogs of Malaysian websites. The dataset contains 2,000 movie reviews. 1,000 of these movie reviews were deemed to be positive, while the remaining 1,000 were deemed to be negative. Also, for the sentiment lexicon, a total of 2478 Malay sentiment lexicon words and phrases were used in this study. Basically, the English WorldNet was employed to collect the more accurate sentiment words that are used in English. These words were then translated into the Malay language for adoption as sentiment words in the lexicon. A synonym was stored for each word and phrase and the polarity was manually assigned a score by a Malay native speaker; each word was assigned a value ranging from -5 (strongly negative) to 5 (strongly positive). This section presents the methodology that has been used in the proposed classification model. The methodology employed by this research covers the following steps: Pre-processing phase, Feature Extraction and Classification phase.

Pre-Processing

The dataset obtained from social media sites was regarded as the raw data, meaning that it contained noisy data. For this reason, it was necessary to pre-process the data before they could be classified by means of machine learning methods. The steps of pre-processing phase are illustrated as follow:

The tokenization process, text of a document is split into a sequence of tokens. Words are confined by whitespace and parentheses, quotes or punctuation marks may appear before or after these words. As such, tokenization is used to break down the sequence of characters according to the locations of the whitespace or punctuation marks between the words that make up a sentence.

Stop Word Removal Task, the process of eliminating words that have a high frequency but which have no bearing on the sentiment of the sentence is known as stop word removal. Listed words such as ‘a’, ‘‘the” and ‘or’ as likely to be regarded as stop words. The stop words in the Malay reviews in this study, such as ‘adalah’, ‘itu’, ‘selepas’ and ‘mereka’ were matched and defined against a list of 475 Malay stop words before being removed from the text.

Feature Extraction

In this phase, a Malay sentiment lexicon is used to capture the sentiment features in the document. The main purpose of this phase is to convert each review into a vector of feature values by means of the Malay sentiment lexicon. A set of features were defined for sentiment classification and different features could be used in this case. For example, the keywords lexicon can be used as features. This study incorporated the stylistic features of sentiment classification as these features are more general and can be used for all languages. The dataset can then be represented by these features using the value of their presence or frequency. An extensive set of 11 features were used and these were grouped under four categories: Features based on the presence and frequency of sentiment words, features based on the level of the sentence, features based on the polarity level of sentiment words and features based on the conditional probability of subjective words. Table 1 illustrates these features.

Some of the features proposed in previous literatures, as well as a proposed new set of features were used. The features from F1 to F8 are adopted with a small modification from (Agarwal *et al*., 2011; Gezici *et al*., 2013; Balahur *et al*., 2014) whilst the new feature set is of subjective words conditional probability features (F9, F10 and F11).

The rest of the features mentioned in Table 1 are mostly self-explanatory. The sentiment word polarity level features include two features that were involved in the sentiment words polarity level features include two features were involved in the conversion of a review into a features vector. The first feature was the weighted probabilities of positive in a review R, It can be formulated as follows Eq.1:

$$F_7(R) = p*(1-P_+)*m_{p+} \tag{1}$$

Where p is the number of positive words in R and P_+ is the probability of seeing positive words in a review. m_{p+}, is the manual assigned scores of positive words from the lexicon.The second feature was the weighted probabilities of negative in a review R, It can be formulated as follows Eq.2:

$$F_8(R) = n*(1-N_-)*m_{n-} \tag{2}$$

However, Subjective words conditional probability level features include three features. Firstly, the average conditional probability of the positive subjective words feature was calculated as follows Eq.3:

$$f_9(R) = \frac{1}{n-1}\sum_{i=1}^{n} p(w_i \mid positive\,class) \tag{3}$$

Where *n* denotes the number of subjective words in the review and *p* (w_i|*pos class*) signifies the probability of the positive words and can calculated as follows Eq.4:

$$p\left(w_i|positive\,class\right) = \frac{f(w_i, pos)+1}{\sum_{w' \in v} f\left(w', pos\right)+v} \tag{4}$$

Secondly, the average conditional probability of the negative subjective words was calculated as follows Eq.5:

$$f_{10}(R) = \frac{1}{n-1}\sum_{i=1}^{n} p(w_i \mid negative\,class) \tag{5}$$

Where n denotes the number of subjective words in the review and *p*(w_i|*negative class*) signifies the probability of the positive words and can calculated as follows Eq.6:

$$p\left(wo_i|negative\,class\right) = \frac{f\left(w_i, negative\right)+1}{\sum_{w' \in v} f\left(w', negative\right)+v} \tag{6}$$

Table 1. Features extracted for each review

Feature set name		Feature name
Sentiment words presence-level features	F1.	Presence of positive words
	F2.	Presence of negative words.
	F3.	Presence of positive words in proportion to the presence of negative words.
	F4.	Frequency of positive words in proportion to the frequency of negative words.
Sentence-level features	F5.	Cumulative frequency of positive words in the first three sentences.
	F6.	Cumulative frequency of negative words in the first three sentences.
Sentiment words polarity level features	F7.	Weighted probabilities of a positive review
	F8.	Weighted probabilities of a negative review.
Subjective words conditional probability features	F9.	Average conditional probability of positive subjective words.
	F10.	Average conditional probability of negative subjective words.
	F11.	Standard deviation of the conditional probability of the subjective words.

The standard deviation of the conditional probability of the subjective words was then calculated as follows Eq.7:

$$f_{11}(R) = \frac{1}{n-1}\sum_{i=1}^{n} p\left(w_i | \frac{postive}{class} \right) - p\left(w_i | \frac{negative}{class} \right) \tag{7}$$

Classification Phase

This study aims to take the advantage of k-Nearest Neighbour (k-NN) classifier in order to combine it with the lexicon based. The k-Nearest Neighbour (k-NN), a popular example-based classifier, is also known as lazy learning because it postpones the decision to make generalizations beyond the training data until it has located every single new incidence. In order to classify a review, the k-NN classifier roughly ranks the review among the training reviews, before classifying it according to the k most similar neighbours.

When presented with a test review d, the classifier will locate the k nearest neighbours among training reviews. The score of each nearest neighbour review that is the most similar to the test review is used as the weight of the classes of the neighbour reviews. The weighted sum in k-NN classification can be represented as follows Eq.8:

$$score(d,t_i) = \sum\nolimits_{d_j = kNN(d)} sim(d,d_j)\,\delta(d_j,c_i) \tag{8}$$

Where, k *NN*(*d*) denotes the set of k-Nearest Neighbours for the review (d). If d_j belongs to c_i, then δ ($d_{j,}$ c_i) could either be equal to 1 or 0. In the case of the test review, *d*, it should fit into the class with the highest resulting weighted sum.

Results

First, the effects of the unigram baseline model are evaluated on the k-NN classifier, by using a unigram feature set to train and test the classifier. The purpose of this experiment was to define a baseline for evaluating the Malay sentiment classification model. The results of this experiment are presented in Table 2.

In the second experiment, the k-NN classifier was applied to the test set using a 10 -fold cross-validation procedure. (F1 to F11) were used to symbolise the extracted features from each review, with F1 being in reference to the presence of positive words, F2 to the presence of negative words and so on in sequence, as demonstrated in Table 1.

Also as shown in Table 1, there were about 11 features, which mean that 2^{11} different experiments could be performed. However, the results that were obtained were for selected experiments from among these 2^{11} experiments. The purpose was to show that the best results were obtained when the k-NN classifier was applied. The accuracy of the k-NN classifier with regards to the precision, recall and F-measure of the Malay sentiment analysis through the application of the k-NN classifier to the different sets of features is shown in Table 3. The selected feature is ticked by "√" symbol and the obtained accuracy is shown for the combined ticked features consecutively. For example, for run number (2), a set of sentiment word presence-level features (Fl, F2, F3 and F4) were tested by applying the k-NN classifier. The result obtained was only 77.40%.

Discussion

It As can be seen in Table 3, the highest accuracy is shown in run number (12), when all features have been implemented, including the new feature set which is Subjective words conditional probability features (F9, F10 and F11), except for Sentiment words polarity level features (F7, F8), to improve the performance of the k-NN classifier in terms of the result obtained, at 86.43%. Meanwhile, performance accuracy, in terms of the average F-measures of the first seven runs (i.e., number 1 -7), was implemented without using the new feature set (F9, F10 and F11). These runs gave results that did not exceed 85.88%. The newly proposed feature set (F9, F10 and F11), gave the highest result and outperformed the first seven runs (i.e., number 1-7), that were implemented without using the new features and gave a 0.55% increase in accuracy against the other feature sets.

The results illustrated in Tables 2 and 3 show that using feature sets has a clear positive effect on the quality of the Malay sentiment analysis performed by the k-NN classification model. Generally, the results illustrated in Table 3 show the use of feature sets was a marked improvement over the baseline model. Obviously, the new features (F9, F10 and F11), had a greater impact on the performance of the k-NN classification model compared to other feature sets. Finely, the highest result that were implemented in Table 3, without using the Sentiment words polarity level features (F7, F8) and gave a 24.43% increase in accuracy against the unigram baseline model.

Table 2. Results obtained on the application of unigram features

Classification method	Macro F1-Measure	F1 Measure for the positive class	F1 measure for the negative class
k-NN	62.00	61.81	62.80

Table 3. Performance of k-NN classifier for Malay sentiment analysis with different feature sets

Run no.	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	Precision%	Recall%	F-Measure%
1.	√	√										69.47	75.98	67.59
2.	√	√	√	√								77.76	80.21	77.40
3.					√	√						69.17	79.36	66.43
4.			√	√	√	√						83.07	83.17	83.07
5.							√	√				84.35	84.64	84.35
6.	√	√					√	√				85.77	86.25	85.77
7.					√	√	√	√				85.90	86.50	85.88
8.									√	√	√	86.29	86.82	86.28
9.					√	√	√	√	√	√	√	84.76	84.87	84.76
10.	√	√	√	√	√	√	√	√	√	√	√	85.40	85.58	85.41
11.			√	√					√	√	√	86.31	86.51	86.32
12.	√	√	√	√	√	√			√	√	√	86.43	86.78	86.43

Conclusion

In this study a Malay sentiment classification model was proposed, whereby this research described several experiments that were conducted to study the effects of four feature sets (based on sentiment words presence and Frequency, sentence level, sentiment word polarity and subjective word conditional probability features) on the performance of the classification method for Malay sentiment classification. In addition, the feature set size was very small, which means the system has a low time and memory complexity and at the same time outperforms unigram feature sets in terms of classification accuracy. Also we found that unlike the unigram feature model, our method would not be dramatically impacted if the similarity between the training and testing data fluctuates. This is because our model captures the underlying sentiment patterns of the documents, rather than exact words. Observed our system with the newly proposed features obtains one of the best results obtained among of all experiments.

Funding Information

The authors have no support or funding to report.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Agarwal, A., B. Xie, I. Vovsha, O. Rambow and R. Passonneau, 2011. Sentiment analysis of twitter data. Proceedings of the Workshop on Languages in Social Media, (LSM' 11), Association for Computational Linguistics Stroudsburg, USA, pp: 30-38.

Balahur, A., R. Mihalcea and A. Montoyo, 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. Comput. Speech Language, 28: 1-6. DOI: 10.1016/j.csl.2013.09.003

Gezici, G., R. Dehkharghani, B. Yanikoglu, D. Tapucu and Y. Saygin, 2013. Su-sentilab: A classification system for sentiment analysis in twitter. Proceedings of the 7th International Workshop on Semantic Evaluation, Jun 14-15, Association for Computational Linguistics, Atlanta, pp: 471-477.

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi and L. Donaldson, 2012. Machine learning and sentiment analysis of unstructured free-text information about patient experience online. Lancet, 380: S10-S10. DOI: 10.1016/S0140-6736(13)60366-9

Isa, N., M. Puteh and R.M.H.R. Kamarudin, 2013. Sentiment Classification of Malay Newspaper Using Immune Network (SCIN). Proceedings of the World Congress on Engineering, Jul. 3-5, London, UK, ISBN: 978-988-19252-9-9

Kang, H., S.J. Yoo and D. Han, 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst. Appl., 39: 6000-6010. DOI: 10.1016/j.eswa.2011.11.107

Liau, Y.B. and P.P. Tan, 2014. Gaining customer knowledge in low cost airlines through text mining. Industrial Managem. Data Syst., 114: 1344-1359. DOI: 10.1108/IMDS-07-2014-0225

Liu, B., X. Li, W.S. Lee and P.S. Yu, 2004. Text classification by labeling words. Proceedings of the 19th National Conference on Artificial Intelligence, (CAI' 04), AAAI Press, Cambridge, pp: 425-430.

Molina-González, M.D., E. Martínez-Cámara, M.T. Martín-Valdivia and J.M. Perea-Ortega, 2013. Semantic orientation for polarity classification in Spanish reviews. Expert Syst. Appl., 40: 7250-7257. DOI: 10.1016/j.eswa.2013.06.076

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (NLP' 02), Association for Computational Linguistics, USA, pp: 79-86. DOI: 10.3115/1118693.1118704

Patil, S.S.D. and R.R. Deshmukh, 2014. Review of Twitter sentiment analysis. Int. J. Scientific Engg. Res., 5: 1616-1623.

Puteh, M., N. Isa, S. Puteh and N.A. Redzuan, 2013. Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. Proceedings of the World Congress on Engineering, Jul. 3-5, London, ISBN: 978-988-19252-9-9

Samsudin, N., A.R. Hamdan, M. Puteh and M.Z.A. Nazri, 2013. Mining Opinion in Online Messages. Int. J. Adv. Comput. Sci. Appl., 4: 19-24.

Samsudin, N., M. Puteh, A.R. Hamdan and M.Z.A. Nazri, 2012. Is artificial immune system suitable for opinion mining?. In Proceedings of the 4th Conference on Data Mining Optimization, Sept. 2-4, IEEE Xplore press, Langkawi, pp: 131-136. DOI: 10.1109/DMO.2012.6329811

Turney, P.D., 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL' 02), Association for Computational Linguistics, USA, pp: 417-424. DOI: 10.3115/1073083.1073153

Xu, T., Q. Peng and Y. Cheng, 2012. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowledge-Based Syst., 35: 279-289. DOI: 10.1016/j.knosys.2012.04.011

Zamani, N.A.M., S.Z.Z. Abidin, N. Omar and M.Z.Z. Abiden, 2013. Sentiment Analysis: Determining People's Emotions in Facebook. ISBN: 978-960-474-368-1.