Domain-specific Sentiment Dictionary Construction for Sentiment Classification

Aye Aye Mar, Nyein Thwet Thwet Aung, Su Su Htay
Faculty of Information Science, University of Information Technology, Myanmar
ayeayemar, suhtay, nyeinthwet}@uit.edu.mm

Abstract

Sentiment dictionaries are commonly used to solve the problem of sentiment classification for customer reviews. The number of sentiment words in the generalized dictionaries such as SentiWordNet is limited and lack of many sentiment words especially domain-specific sentiment words. Different domains have different sentiment words and the sentiment of a word depends on the domain in which it is used. In this paper, an approach based on Point-wise Mutual Information (PMI) is proposed to construct a domain-specific sentiment dictionary effectively and automatically. The proposed system is evaluated on three diverse datasets from different domains by using 10-fold cross validation. Accordingly to the experimental results, the goodness of the extracted dictionary is relatively high and significantly improves the performance of sentiment classification. The experimental results show that the extracted domain-specific dictionary outperforms the generalized dictionary, SentiWordNet. The proposed method learns the domain-specific sentiment words efficiently and it is domain adaptable.

Keywords- Sentiment Analysis, Polarity Classification, Sentiment Dictionary, Domain-specific Sentiment Words, Point-wise Mutual Information

1. Introduction

The amount of user generated data on the web is increasing more and more during the last few years. As a consequence of this, sentiment analysis from these data has become a prominent research area. Sentiment analysis is a kind of text mining combined with the natural language processing and computational linguistics. It is the task of extracting valuable information from a collection of documents containing opinions, feelings and attitudes.

Sentiment analysis is applied in three levels of granularity, which are document level, sentence level and aspect level also called feature level. The key factor of all these levels is to identify the polarities of the sentiment words. The polarities of some sentiment words vary based on the domain in which it is used. "Unpredictable" may have a negative sentiment in a car review as in "unpredictable steering," but it could have a positive sentiment in a movie review as in "unpredictable plot" [1]. Most of the existing sentiment dictionaries specify the polarity of such words generally instead of considering the polarity of these words for each specific domain. Therefore, the methods that are able to construct domain-specific sentiment dictionaries are essential for an accurate sentiment classification system.

This paper proposes an approach for constructing a domain-specific dictionary by using labelled review datasets as the training. The approach considers the probability distribution of the sentiment words with the class labels which is computed by Point-wise Mutual Information (PMI). The proposed method solves the following three main problems of sentiment analysis: (1) the need of human effort to construct a domain -specific dictionary manually (2) missing the polarities of domain-specific sentiment words when the generalized sentiment dictionaries are lack of them (3) the problem to identify the right polarities for domain dependent sentiment words.

The rest of the paper is organized as follows: Section 2 summarizes the related work. The proposed system is presented in Section 3. In Section 4, the experimental evaluations are described. Section 5 concludes the paper and the future work of the proposed method is presented in Section 6.

2. Related Work

There are three common approaches for generating sentiments of words: manual, dictionary-based and corpus-based approaches. The manual approach simply uses human knowledge to decide the sentiment of a word. Meanwhile, dictionary based and corpus-based approaches automatically generate sentiments of words using dictionaries and corpuses respectively [2]. Dictionary-based approach is one of the main approaches to extract sentiment words in sentiment analysis [1]. Due to the importance of sentiment words within sentences, many approaches have been proposed to predefine the polarity of sentiment words.

A manually built lexicon has been used to classify the text as the positive or negative. In [3], Hatzivassiloglou and Wiebe assumed that adjectives are the clues to trace the sentiment orientation of a given text. Based on the manually created lexicon for adjectives and their semantic orientation values (SO),

for any given text, all adjectives are extracted and associated with their dictionary SO values. The overall sentiment score is obtained by summing up all adjective SO scores within the given text. The given text is then classified as bearing a positive or negative sentiment based on the overall score for the obtained adjectives within it.

The need of creating new hand-built lexicons for the new domains is very labor intensive. In order to create a sentiment dictionary efficiently and escape any manual effort, Turney [4] proposed a simple promising approach to create a sentiment dictionary in an automatic way. The dictionary was built by using the positive and negative seed words. In order to find the correlation between a seed word and the target word, the author used mutual information approach which is based on statistical data extracted from the web with the aid of AltaVista search engine. The target word is passed as a query to the search engine i.e., either with the word "excellent" or the word "poor". The semantic orientation is then acquired based on the mutual information between the target word with the word "excellent" or with the word "poor". If the attained mutual information score for the target word with the word "excellent" is greater than the one with the word "poor", the target word will be classified as positive, otherwise it will be classified as negative.

In [5], Hu and Liu claimed that the created dictionary list can be further expanded by utilizing synonym and antonym sets in WordNet [6]. The polarities of groups of synonyms are assumed to be similar e.g., "beautiful" and "pretty" while the polarities of antonyms are supposed to be opposite e.g., "excited" and "bored". However, Leung et al., argue that 580 semantic similarities do not necessarily employ sentimental similarity accordingly to statistical evidence obtained from movie review data [7].

Manually created dictionary would be convenient to detect a sentiment only for a given domain. Therefore, researchers created publicly available lexical resources such as SentiWordNet. SentiWordNet is a lexical resource of sentiment information for terms in English language designed to assist in opinion mining tasks. Each term in SentiWordNet is associated with numerical scores for positive and negative sentiment information [9] [10]. SentiWordNet can be used for sentiment analysis of all domains. However, the number of terms defined in SentiWordNet is limited [2].

Different domains have different kinds of sentiment words. Although sentiment words can be the common words among different domains, all of the sentiment words cannot be the same. Even though the same sentiment word is contained in the sentiment words lists of many domains, its polarity will be changed depending on the domain that is associated with. A positive or negative sentiment word may have opposite orientations in different application domains. For example, "suck" usually indicates negative sentiment, e.g., "This camera sucks," but it can also imply positive sentiment, e.g., "This vacuum cleaner really sucks." [1].

An automatic approach for constructing domain-specific sentiment dictionary is necessary for improving the sentiment classification. The number of terms in SentiWordNet is limited and usually lack of many sentiment words, especially domain specific sentiment words [2]. This is the motivation of our research to propose an approach for building the domain-specific dictionary by using a labelled training dataset instead of using generalized SenitWordNet dictionary for sentiment classification.

The main goal of the proposed method is to seek the relevant sentiment words for a given domain effectively and automatically. The system is also aimed to develop robust classification approach of customer reviews based on domain-specific labelled training datasets by applying statistical approach. Moreover, the system analyses the performance of Point-wise Mutual Information (PMI) method by using different review datasets from different domains.

3. Proposed System

The proposed system is mainly composed of three components: preprocessing, constructing sentiment dictionary and classifying the reviews by utilizing the extracted dictionary. Firstly, all of the review datasets are preprocessed. Secondly, domain-specific dictionary is constructed by computing sentiment orientation of these subjective words based on Point-wise Mutual Information (PMI). Finally, review documents are classified by applying the extracted domain-specific sentiment dictionary. Figure 1 shows the system flow of the proposed system.

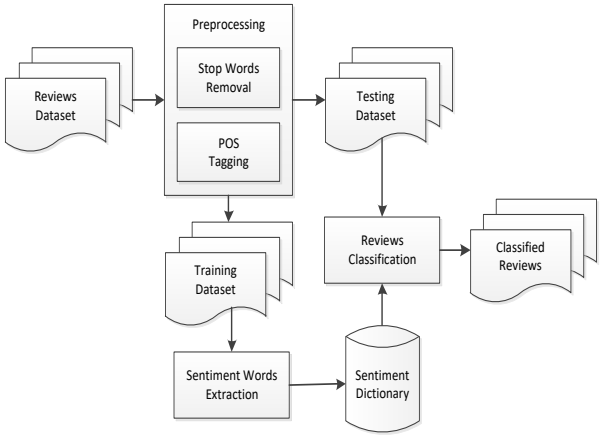

Figure 1. System flow of the proposed system

3.1 Preprocessing

Preprocessing is necessary before extracting the sentiment features. It includes two parts: POS tagging and stop words removal. First and foremost, POS tagging is done by applying the Standford POS tagger tool[1]. POS tagging means labelling each word in a sentence with its appropriate part of speech such as noun, adjective, adverb etc.

Stop words such as verb to be, pronouns, prepositions and conjunctions do not give meaningful information for sentiment analysis. So, the stop words[2] are removed to save the processing time.

3.2 Constructing Sentiment Dictionary

Most of the previous works consider only adjectives as the sentiment words. In similar to adjectives, adverbs and verbs also describe sentiments as the adjectives. The experimental results shows that that polarity classification is more accurate by considering the polarities of adjective, adverb and verb instead of adjective alone. Therefore, not only adjective but also adverb and verb are considered as the sentiment words in this system. The sentiment dictionary is constructed by using the Algorithm 1.

The positive and negative sentiment scores of each sentiment word are computed based on the Point-wise Mutual Information (PMI) [11]. If a word is occurred frequently and predominantly in one class (positive or negative), then that word would have high polarity. If the positive PMI score of a sentiment word is greater than its negative PMI score, it indicates that the word has occurred mostly in positive documents. Alternatively, it indicates that the word has occurred mostly in negative documents if the negative PMI score of a word is greater than its positive PMI score. Point-wise Mutual Information (PMI) is used to calculate the strength of association between a word and positive or negative documents in sentiment analysis. The positive PMI-Score and negative PMI-Score of each sentiment word w is computed by Eq. (1) and by Eq. (2) respectively. In this system, both the positive PMI-Score and negative PMI-Score of each sentiment word are taken into account for computing the polarity of the review document.

$$PMI(w, Positive) = \log_2 \frac{P(w, Positive)}{P(w)P(Positive)} \quad (1)$$

$$PMI(w, Negative) = \log_2 \frac{P(w, Negative)}{P(w)P(Negative)} \quad (2)$$

[1] https://nlp.stanford.edu/software/tagger.shtml
[2] http://xpo6.com/list-of-english-stop-words/

Where, P(w , Positive) is the joint probability of co-occurrence of sentiment word w found together with the class Positive and P(w) and P(Positive) are the probability of occurrence of sentiment word w and class Positive independently.

After computing the positive PMI-Score and negative PMI-Score for each sentiment word, the system constructs the sentiment dictionary which includes the sentiment words together with their respective POS tag (Part of Speech tag), positive PMI-Score and negative PMI-Score.

Algorithm 1. Algorithm for constructing sentiment dictionary

Input : A given training reviews set S with the POS tagged words ← { $s_1(w_1,w_2,....,w_n$: label),…....,s_n ($w_1,w_2,....,w_n$: label)}

Output: sentiment dictionary which contains sentiment words together with their respective POS tag, positive score and negative score

counts the occurrence frequency of each word $w \in s \in S$
1: **for** each training review $s \in S$ **do**
2: **for** each word $w \in s$ **do**
3: **if** (w is adjective or adverb or verb) **then**
4: **if** (w is not in sentiment word list L) **then**
5: $L \leftarrow w$
6: **end if**
7: **if** (label of s is positive) **then**
8: increase one to *pos_count* of s
9: **else** (label of s is negative)
10: increase one to *neg_count* of s
11: **end if**
12: **end if**
13: **end for**
14: **end for**
15: create the frequency table by using *pos_count* and *neg_count* of each $w \in L$
16: compute the probability table from the frequency table
compute positive PMI-Score and negative PMI-Score for each $w \in L$
17: **for** each $w \in L$
18: *pos_score* ← *PMI (w, Positive) by Eq.1*
19: *neg_score* ← *PMI (w, Negative) by Eq.2*
20: save w into sentiment dictionary D together with *POS tag, pos_score, neg_score*
21: **end for**

Algorithm 2 classifies the different testing datasets by utilizing the domain-specific sentiment dictionary extracted from the respective training review datasets.

Algorithm 2. Algorithm for classifying reviews documents by using the extracted domain-specific dictionary

---

Input : testing reviews set T with the POS tagged words ← {t_1 (w_1,w_2,....,w_n),....., t_n (w_1,w_2,....., w_n)}, sentiment dictionary D

Output : testing reviews set T with assigned labels

classify the testing reviews set by using the extracted sentiment dictionary

1: **for** each testing review $t \in T$
2: **for** each $w \in t$
3: *total_pos_score* ← *pos_score* of w in D
4: *total_neg_score* ← *neg_score* of w in D
5: **end for**
6: **if** (*total_pos_score* > *total_neg_score*) **then**
7: t ← Positive
8: **end if**
9: **if** (*total_neg_score* > *total_pos_score*) **then**
11: t ← Negative
12: **end if**
13: **if** (*total_neg_score* == *total_pos_score*) **then**
14: t ← Neutral
15: **end if**
16: **end for**

---

4. Experiment Evaluation

4.1 Dataset Description

The publicly available three diverse review datasets are used to evaluate the domain adaptability of the proposed method. These three domains are movie, product and hotel. We used publicly available Cornell movie review dataset[3] of Peng and Lee for movie domain. Some of the hotel reviews from tripadvisor are taken for hotel review dataset[4]. For product domain, the beauty product reviews[5] of amazon product review datasets are used. All of the three datasets contain 50% positive reviews and 50% negative reviews to maintain the class distribution. The description of these three diverse datasets is shown in Table 1.

[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[4] http://sifaka.cs.uiuc.edu/~wang296/Data/index.html
[5] http://www.ilabsite.org/?page_id=1091

Table 1. Dataset description

Domain	Positive	Negative	Total
Movie	1000	1000	2000
Product	4500	4500	9000
Hotel	12500	12500	25000
Total	18000	18000	36000

4.2 Preparing Training and Testing Data

In our evaluation, splitting the datasets into training and testing involves the k-fold cross validation method [8]. In k-fold cross validation method, the data is split into k folds where k-1 folds is used for training the algorithm and the remaining one fold is used for testing the algorithm. The final measure of performance takes the average of the results of all folds. In this work, we used 10-fold cross validation to make robust evaluation.

4.3 Evaluation Metrics

Five evaluation metrics, which are precision, recall, F-measure, accuracy and failure-ratio, are used to evaluate the effectiveness of the system. These are calculated by using Eq. (3)-(7) respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F - measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Failure - Ratio = \frac{NumberOfMisclassified\ Reviews}{TotalNumberOf\ Reviews} \tag{7}$$

Where:

TP refers to the number of true positive reviews.
TN refers to the number of true negative reviews.
FP refers to the number of false positive reviews.
FN refers to the number of false negative reviews.
Number of Misclassified Reviews refers to the reviews labelled to the class label which was not included in the actual class labels.
Total Number of Reviews refers to the number of all reviews.

4.4 Experimental Results

This section analyses the experimental results of the proposed method and the baseline SenitWordNet dictionary on three diverse datasets. Table 2 shows the evaluation results of baseline SentiWordNet and the proposed method by using the movie, hotel and product review datasets. Among the five performance measures, the proposed method has significant results than the baseline in precision, F-measure and accuracy.

A few explanations concerned with the failure-ratio should be made here. The failure-ratio in Table 2 means that the error that is occurred when a review document is labelled with the class labels that is not really present. As shown in Table 1, the datasets contain only two class labels, positive and negative. There were no neutral review documents in the datasets. However, both the baseline method SentiWordNet and the proposed method make classification to a few documents as the neutral documents (which is not present in the actual class labels). Labelling the documents as the neutral is happened when the total positive score is equal to total negative score of the review document.

Table 2. Experimental results in % of SentiWordNet (SN) and the proposed method (PM)

Dataset	Product Dataset		Movie Dataset		Hotel Dataset	
Method	SN	PM	SN	PM	SN	PM
Precision	58.05	**85.46**	57.75	**78.16**	76.75	**81.69**
Recall	**88.11**	84.55	**88.15**	76.64	**96.86**	92.54
F-measure	69.94	**85.00**	69.67	**77.28**	85.63	**86.76**
Accuracy	62.25	**85.20**	62.05	**77.58**	77.10	**85.71**
Failure-Ratio	**0.18**	0.6	**0**	0.12	**0.33**	0.54

For product review dataset, the experimental results show that proposed method (PM) has significant high results than the baseline SentiWordNet in precision, F-measure and accuracy except low result in recall. In the proposed model, precision is improved dramatically by 27.41%, the F-measure is increased significantly by 15.06 % and the accuracy is improved inevitably by 22.95% except the decline of 3.56% in recall. Both the baseline method and the proposed method have failure-ratio of 0.18% and 0.6% respectively.

As in the product domain, the performance of proposed method in movie domain has a visible improvement in precision, F-measure and accuracy with the increment of 20.41%, 7.61% and 15.53% respectively. The recall of the proposed method is decreased by 11.51%. In the view point of failure-ratio, the baseline method is failure free in this domain although the proposed method has a slight failure-ratio of 0.12%.

In hotel domain, both the methods have a pretty good evaluation results and the proposed method is improved in terms of precision, F-measure and accuracy. The proposed method has higher precision by 4.94 %, increased F-measure by 1.13% and raised accuracy by 8.61% than the baseline except the low recall by 4.32%.

To sum up about the comparative results of the baseline method and the proposed method, the proposed method is improved dramatically although it has low recall and a slight failure-ratio than the baseline.

5. Conclusion

This paper proposed an approach to solve the problem of automatic construction of domain-specific sentiment dictionary. The extracted dictionary is evaluated by using 10-fold cross validation to be robust evaluation of the system. The experimental results demonstrate that the proposed method efficiently learns domain-specific sentiment words. The precision, F-measure and accuracy of the proposed system have a significant result than the baseline generalized dictionary, SentiWordNet.

6. Future Work

As the future work of the system, negation case will be considered to improve the performance of the system. Currently, the system is able to analyze sentiments in document level. To get the sentiments of customers in more details, aspect level sentiment analysis will have to be done. The sentiment targets of the sentiment words will be detected in order to make the right decision what the customers like and dislike.

7. References

[1] B. Liu, "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1, 2012, pp. 1-167.

[2] N.H.Nguyen, T.V.Le, H.S.Le, T.V.Pham, "Domain specific sentiment dictionary for opinion mining of vietnamese text." International Workshop on Multi-disciplinary Trends in Artificial Intelligence. Springer, Cham, 2014.

[3] V.Hatzivassiloglou, J.M.Wiebe, “Effects of adjective orientation and gradability on sentence subjectivity”, In Proceedings of the 18th conference on Computational linguistics-Volume 1 , Association for Computational Linguistics, 2002,pp. 299-305.

[4] P.D.Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In Proceedings of the 40th annual meeting on association for computational linguistics,. Association for Computational Linguistics, 2002, pp. 417-424.

[5] M.Hu, B.Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 168-177.

[6] G.A.Miller. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[7] C.W.Leung, S.C.Chan, F.Chung. "Integrating collaborative filtering and sentiment analysis: A rating inference approach." Proceedings of the ECAI 2006 workshop on recommender systems. 2006, pp. 62-66.

[8] R.Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *Ijcai*, 1995,vol. 14, no. 2, pp. 1137-1145.

[9] S.Baccianella, A.Esuli, and S|.Fabrizio "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, 2010, pp. 2200-2204.

[10] A.Esuli , F.Sebastiani, "SENTIWORDNET:A publicly available lexical resource for opinion mining", In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), 2006, pp. 417–422, Genova, IT.

[11] K.W.Church, and H.Patrick "Word association norms, mutual information, and lexicography." Computational linguistics 16, no. 1, 1990, pp. 22-29.