A construction and self-learning method for intelligent domain sentiment lexicon

Shaochun Wu*, Qifeng Xiao, Ming Gao and Guobing Zou

Department of Intelligent Information Processing,
Shanghai University,
Shanghai 200444, China
Email: scwu@shu.edu.cn
Email: xqf_shu@163.com
Email: qywtgm950120@foxmail.com
Email: gbzou@shu.edu.cn
*Corresponding author

Abstract: A new method of building intelligent sentiment lexicon based on LDA and word clustering is put forward in this paper. In order to make seed words more representative and universal, this method uses LDA topic model to build the term vectors and select seed words. The improved SO-PMI algorithm has been used to calculate the emotional tendency of each sentiment word. In addition, the domain sentiment lexicon's automatic extension and update method is designed to deal with dynamic corpus data. Experiments show that the proposed method can build the sentiment lexicon with higher accuracy, and can reflect the change of words' emotional tendency in real time. It is proved in this paper that this method is more suitable for processing a large number of dynamic Chinese texts.

Keywords: sentiment lexicon; SO-PMI algorithm; seed words; LDA topic model; word clustering; incremental text processing.

Reference to this paper should be made as follows: Wu, S., Xiao, Q., Gao, M. and Zou, G. (xxxx) 'A construction and self-learning method for intelligent domain sentiment lexicon', *Int. J. Information Technology and Management*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes:

This paper is a revised and expanded version of a paper entitled 'A construction and self-learning method for intelligent domain sentiment lexicon' presented at ICSS2018, Shanghai University, 13 May 2018.

1 Introduction

During the past decade, with the widespread use of web sites and social media, a large number of comment texts are generated. These comments reflect the commenters' feelings about certain person, event and commodity. They can be a reflex of the orientation of public opinion clearly. Sentiment analysis of these text data is of great scientific and commercial value.

Sentiment analysis is refered to the analysis and reasoning of subjective text with emotional content. According to the size of the processing target, the text sentiment analysis can be divided into word, phrase, sentence and chapter sentiment analysis (Pang et al., 2008; Cambria et al., 2013). Word sentiment analysis is the basis of sentence and chapter sentiment analysis. A broad and precise sentiment lexicon could be used to improve the accuracy of the sentiment analysis method (Turney and Littman, 2003; Kennedy and Inkpen, 2006).

The most common way to build a sentiment lexicon is to calculate the pointwise mutual information (PMI) between unknown emotional words and known words, and then using the calculated PMI to measure the emotional tendency of each unknown emotional word. Using this method to construct the sentiment lexicon can match the basic demand for sentiment analysis. Yet there are some problems which could always lead to low accuracy such as seed words selection strategies, the complex association relationship between words, data sparse and so on (Turney and Littman, 2003). Furthermore, the traditional methods of building sentiment lexicon are mostly used to train samples in static historical data monolithically to obtain emotional information of unknown sentiment words. When sentiment lexicon requires update, we need to add new training data into the original corpus and train the whole corpus again to establish the sentiment model. This approach has the following limitations:

- Commentary data on TV series usually comes in batches. It is difficult to collect enough samples at once. The traditional method of training and learning for historical data needs to be repeated, this would be a waste of time.
- For sentiment words in some areas like sports and entertainment news, it is a common case that a positive word mutates into a negative one. Their emotional tendency often changes rapidly and intensely. Traditional training methods based on the historical data cannot to adjust the model for dynamic data immediately, so that the sentiment lexicon cannot adapt to the changes of the network language environment, and result in declining accuracy.

- In some professional fields like medical science, some words' semantic and emotional tendencies are usually different from those in normal language environment (for example: 'hyperactive' in common text such as sports news tend to show a positive emotion, but it tends to show negative sentiment in most medical text). Sentiment lexicon needs to reflect the emotion tendency in specific domain for the accuracy.

To overcome these problems, in this paper, we provide a construction and self-learning method for intelligent domain sentiment lexicon. First, the seed words set is constructed by word clustering methods based on LDA topic model. Second, synonym group mining is imported to improve the accuracy of traditional SO-PMI algorithm. Then the initial domain sentiment lexicon is constructed. And as the expanding of corpus, the lexicon can constantly adapt to the new text data, maintaining high coverage and accuracy.

2 Related work

Lexicons play an important role in the study of textual sentiment analysis. There are existing a large number of researches about the topic of lexicons. Deng et al. (2017) who designed methods to improve the accuracy of sentiment lexicon in certain domain by training domain corpus repeatedly. Park and Kim (2016) selected several synonyms of certain candidate from different dictionaries, and judged the emotional tendency of emotional words according to the distribution of positive and negative words among these synonyms. Kawabe et al. (2015) classified texts with LDA topic model, and they constructed a sentiment lexicon based on this topic to proceed sentiment analysis on texts under a specific topic. Then they use this method for Twitter sentiment analysis. Because sentiment lexicon resources under English, Japanese is relatively complete, scholars can focus on improving the method of sentiment mining and promoting the efficiency and accuracy of sentiment lexicon. However, the sentiment dictionaries are insufficient for the Chinese context. Some researchers start to expand existing sentiment lexicon to construct a high-performance sentiment lexicon. Wang and Ku (2016) has constructed the largest Chinese general sentiment lexicon, ANTUSD by expanding NTU. Liu et al. (2013) offered a word sentiment tendency calculation method with a combination of HowNet and PMI. They used HowNet to expand the synonyms set so that problems caused by low frequency of the emotional words in the corpus could be reduced. The other group of researchers hope to build sentiment lexicon by the statistical characteristic of the emotional words in domain corpus. Fu et al. (2017) combined HowNet lexicon and word emotion tendency calculation to train phrase recursive autoencoder to solve the insufficient of sentiment analysis in the context relationship, then improved the accuracy of sentiment recognition.

The researchers above mostly focused on calculation of words' sentiment tendency. Most of them did not improve sentiment lexicon's ability of reflecting the change of sentiment words in dynamic corpus, and the specific sentiment tendency of uncommon sentiment words in certain domain areas.

In order to handle these shortcomings, in this paper, we present a construction and self-learning method of sentiment lexicon for dynamic data and specific domain. In the case of high accuracy, self-learning and self-expansion are implemented for dynamic data and in specific domain language environments.

3 Extraction of seed words based on word clustering

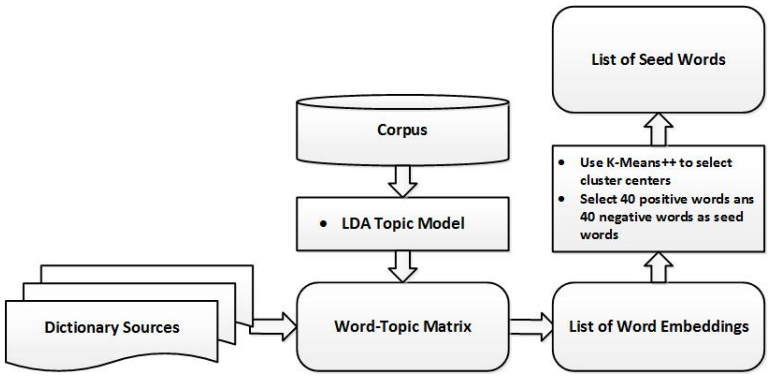
In natural language processing, PMI is mainly used to measure the correlation between two specified sample points (such as two words). The sentiment tendency of a word could be calculated by using the PMI scores between the word and the seed words. Then, the selection of seed words becomes the key to building a high-precision initial sentiment lexicon.

Seed words refer to several words with strong commendatory or derogatory tendency in known vocabulary. Seed words are often very representative among words of the same emotional tendency. A word set's coverage, emotional intensity and distinguishing ability must be considered before being selected as seed words set. A seed words set must have some characteristics including a wide range of uses, small similarity between different members and strong emotional tendencies (Zagibalov and Carroll, 2008; Ju et al., 2012; Yu et al., 2013; Duyu et al., 2013). Traditional SO-PMI algorithm usually selects seed words by the frequency of the sentiment words in the corpus. This method of selection can reduce the deviation of sentiment calculation caused by the low frequency of seed words. However, seed words selected through frequency cannot reflect seed words' ability of representation in theory.

In order to make sure that the seed words have a very obvious commendatory or derogatory tendency and strong representative characteristics without affecting the co-occurrence statistics, we use word clustering to cluster the known sentiment words and select several clusters' centre as seed words.

In this section, we mainly introduce the selection process of seed words with positive emotion words as an example. The selection process is shown in Figure 1.

Figure 1 Flow chart of seed words selection



First of all, we combined HowNet, NTUSD and some other dictionaries as the basic sentiment lexicon D1. Then, we deduced the word frequency of the word in the corpus, and selected 3,000 words with the maximum frequency as the basic seed words set D2.

To characterise the emotional words set, we need to get the feature space and eigenvector of each word. Here we use the LDA topic model to assign m topics $(p_topic_1, p_topic_2, p_topic_3, ..., p_topic_m)$to the original corpus. Then we choose a word's topic probability distribution from the resulting 'word-subject' matrix as its m-dimensional eigenvector (Laohakiat et al., 2017; Cai et al., 2014; Li et al., 2017; Onan et al., 2017; Fries, 2015).

After we get the eigenvectors, we used word clustering to select seed words. First, we calculate the similarity between two eigenvectors by cosine similarity, as follows:

$$\overline{word} = (p\_topic_1, p\_topic_2, p\_topic_3, ..., p\_topic_m) \tag{1}$$

$$D(x) \leftarrow \frac{\overline{word_1} * \overline{word_2}}{\| \overline{word_1} \| * \| \overline{word_2} \|} \tag{2}$$

Then, we used the K-means++ algorithm to select the centre of each cluster. The main ideas of K-means++ algorithm is that assuming that n initial clustering centres have been selected ($0 < n < K$), when the $n+1$ clustering centre is selected, the further away from the current n clustering centres, the higher the probability will be selected as the $n+1$ clustering centre and the first clustering centre ($n = 1$) is also selected by random method. So, K-means++ algorithm mainly focuses on the improvement of the clustering centre initialisation. It has a more scientific method to select the cluster centre instead of random selection which K-means algorithm to select the centre of each cluster by for the distinguishing ability of seed words. Since we choose k cluster centres as k seed words, we only need to complete the process of clustering centre initialisation of K-means++ algorithm (Arthur and Vassilvitskii, 2007), then we can get k positive seed words, and construct a positive seed words set. Pseudo-code of seed words selection method is as follows:

Algorithm 1 Seed words selecting algorithm

Input: a set of documents D, basic seed set S, the count of seed words k;
Output: a set of seed words $Seed = \{s_1, s_2, ..., s_k\}$;
do run LDA Topic on D and get word-topic matrix M_1;
foreach $s_i \in Seed$ do
    find the probability distribution of s_i among all topics in M_1 as v;
    add v into V
do select a random point from V as the first initial centre;
foreach $j \in k$ do
    foreach $x \in S$ do
        select its closest centre point C;
        $Sum(D(x)) \leftarrow \sum D(x)$;
    repeat
        $Random- = Sum(D(x))$
    until $Random \leq 0$;$s_i \leftarrow current\_centre$;
$Seed \leftarrow \{s_1, s_2, ..., s_k\}$;
return $Seed$

4 Construction of initial sentiment lexicon based on improved SO-PMI algorithm

After selecting enough reasonable seed words, we need to use the SO-PMI algorithm to calculate the emotional tendency of unknown sentiment word.

The SO-PMI algorithm mainly evaluates the similarity between two words by calculating the PMI value between them and evaluates the emotional tendency of the candidate word by whether the PMI between it and the commendatory is larger than that between it and the derogatory seed words. For the words w1 and w2, their PMI can be expressed as:

$$PMI(w_1, w_2) = \log \frac{p(w_1 \& w_2)}{p(w_1) \times p(w_2))} \approx \log \frac{N \times C(w_1 \& w_2)}{C(w_1) \times C(w_2))} \tag{3}$$

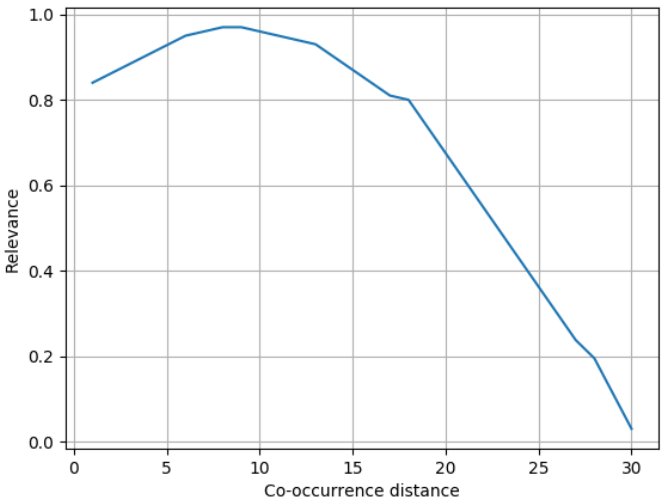
where N is the total number of documents, $C(w_1)$ is the word frequency of w_1, $C(w_1 \& w_2)$ is the number of simultaneous co-occurrences of words w_1 and w_2 in the same document.

The formula above gives the basic method of calculating the mutual information between two words. However, the accuracy of the result is not satisfactory enough in practice. The reason is that the degree of contact between two words cannot be determined only by whether they appear in a single sentence together. In actual calculation process, we often face the problem of data sparseness, where we use the co-window and synonym group to improve the accuracy of PMI calculation.

A co-occurrence window d refers to the two words which appear in the window with specified size of d. This means that we treat those two words as co-occurrence words while the distance between them in a single sentence is less than or equal to d.

The effect of co-occurrence window in PMI calculation (Damani, 2013; Wawer, 2012; Paperno and Baroni, 2016) is shown in Figure 2. It can be seen that, once the distance is crossed by a specific threshold, the correlation is linear with the co-distance, and the correlation decreases as the covariance distance increases.

Figure 2 The effect of co-window on the accuracy of sentiment calculation (see online version for colours)



Not only that, the co-distance between two words can reflect the degree of correlation between those two words, so this paper imports co-window and co-distance into SO-PMI

algorithm to improve the accuracy. After the co-distance is added to the calculation, the formula for calculating the emotional predisposition is as follows:

$$SO\_PMI(w) = \sum_{pword_i \in Pwords}^{i=1,2...m} \frac{1}{d_i} PMI(cw, pword_i) - \sum_{pword_i \in Nwords}^{j=1,2...m} \frac{1}{d_j} PMI(cw, nword_j) \tag{4}$$

where d denotes the co-distance between the candidate word cw and the seed word, cw denotes terms in N, pwords and nwords denotes the positive seed words and negative seed words respectively.

A lot of candidate words meet the problem of low frequency and data sparse (Cheng et al., 2016; De Vylder and Tuyls, 2005). Such as '偲', means versatile, with a clear tendency to commend, but the number of occurrences in the corpus is almost 0, which leads to the final yield of SO-PMI close to 0. This would make the original strong positive word into a 'neutral word'. However, '偲' and '多才多艺' is synonymous, if we can treat '多才多艺' as a consent word of '偲' during calculation of PMI, this would be able to greatly improve the accuracy rate. We use synonyms to improve the lexicon construction method. For the unknown sentiment word w, whose occurrences are fewer than the threshold m, we find the synonym group of w in the synonym sources, and convert the SO-PMI calculation of w to the SO-PMI calculation of its synonym group.

Wu et al. (2013) proposed the use of synonym lexicon provided by Harbin Institute of Technology Social Computing and Information Retrieval Research Center to find the synonym group of sparse words, but in the actual use of the process, there are many professional words and network emerging words do not have synonyms that could be found in the synonym lexicon. Therefore, this paper proposes a method to construct the synonym group by LDA topic model.

Because the construction of the synonym group is very similar with the word clustering process, we generate the word eigenvector with the LDA 'word-topic' matrix. Then we use cosine similarity to calculate the similarity between different words and select the n words most similar with each word as its synonym group. The construction of synonym group is as follows:

Algorithm 2 Synonyms mining algorithm

Input: a set of word-topic vectors V, a candidate word c;
Output: a set of c's synonyms $S = \{w_1, w_2, \ldots, w_k\}$;
do find the word-topic vector of c in V as $\vec{c}$;
foreach $w \in V$ do
    calculate similarity between w and c;
do select top k words with highest similarity as s;
$S \leftarrow \{s_1, s_2, ..., s_k\}$;
return S

While we calculate the SO-PMI value of w, we calculate the SO-PMI value of the synonym group N of w instead of w itself. The improved SO-PMI formula is as follows:

$$SO\_PMI(w) = \frac{1}{N}\sum_{cw\in N}^{N}\Big(\sum_{pword_i\in Pwords}^{i=1,2...m}\frac{1}{d_i}PMI(cw,pword_i) - \sum_{pword_i\in Nwords}^{j=1,2...m}\frac{1}{d_j}PMI(cw,nword_j)\Big) \tag{5}$$

where d denotes the co-distance between the candidate word cw and the seed word, N denotes the synonyms group of w, and cw denotes terms in N, $pwords$ and $nwords$ denotes the positive seed words and negative seed words respectively. After using the improved SO-PMI algorithm to calculate the emotional tendency of all the candidate words, we integrate the external sentiment lexicon resource D1 with the calculated candidate word set, which can obtain the high-accuracy basic sentiment lexicon domain basic sentiment lexicon (DBSL).

5 The self-learning process of intelligent domain sentiment lexicon

In order to make the domain sentiment lexicon (DSL) be able to adapt to the rapidly changing language environment in the big data environment, the intelligent domain sentiment lexicon (IDSL) needs to be able to perceive the language changes from the newly loaded new corpus data, such as the change of certain word's emotional tendency, as well as the emergence of new emotional words, etc. The IDSL needs to have the ability to adjust the emotional value of existing words under dynamic data and add new words through self-learning.

During the update process, assume that there are n new documents updated, we preprocess these n documents, figure out candidate words' frequency and update co-occurrence information between candidate words and seed words. While the data update is completed, the emotional polarity of the candidate words is updated by the following formula to update the emotional tendencies in the sentiment lexicon:

$$PMI(w_1,w_2) = \log\frac{(N+\Delta N)\times(C(w_1\&w_2)+\Delta C(w_1\&w_2))}{(C(w_1)+\Delta C(w_1)\times(C(w_2)+\Delta C(w_2)))} \tag{6}$$

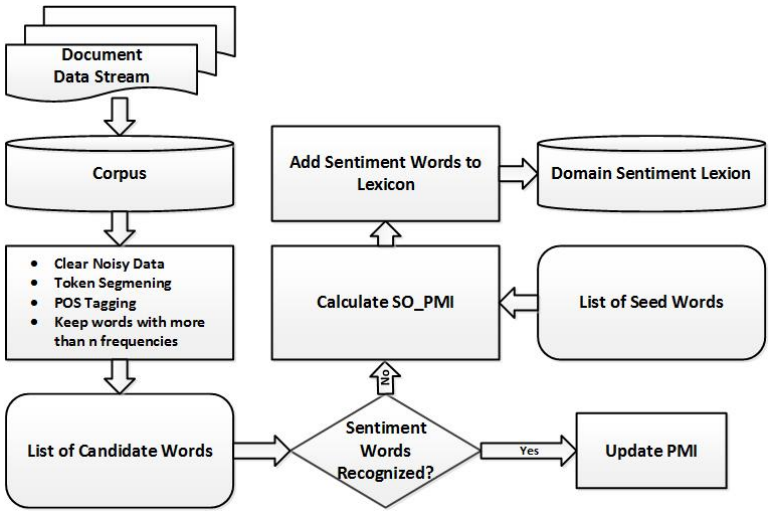
where ΔN represents the number of new text, $\Delta C(w_1)$, $\Delta C(w_2)$, $\Delta C(w_1\&w_2)$ are the growth of candidate words, seed words and the co-occurrences of them, respectively.

In the process of updating the sentiment lexicon, not only the emotional tendency of the words in the emotional lexicon will change with the change of the time node, but also the construction of the sentiment lexicon will encounter a variety of new words, or some cold old words whose frequency would improve greatly because of the emergence of new interpretations of these words. Our sentiment lexicon needs to archive expansion, self-renewal and self-improvement with the dynamic inflow.

While updating the lexicon, we need to preprocess the incoming text and count candidate word frequency again. In the statistical process, we select emotional words with frequency over a certain threshold, but the emotional words are not included in the pre-update sentiment lexicon, then we figure out their co-occurrence with the seed words

and calculate their SO-PMI value. Then those new words are added to the sentiment lexicon, so as to realise the real-time active learning mechanism of the DSL. The active expansion process is shown in Figure 3.

Figure 3 Self-learning process of the IDSL



6 Experimental results and evaluation

The above section details the process of improving the accuracy of traditional SO-PMI algorithm by constructing the word clustering and the synonym group and adding the co-distance parameter. The improved algorithm takes full account of the influence of seed words selection, data sparse and co-distance on the calculation of emotional words' predisposition. Theoretically, the improved SO-PMI algorithm should be better than the traditional algorithm in performance. In this section, we compare the performance of the different SO-PMI algorithms and IDSL's ability of self-learning. The experiment proves the effectiveness of the algorithm.

6.1 Experimental data

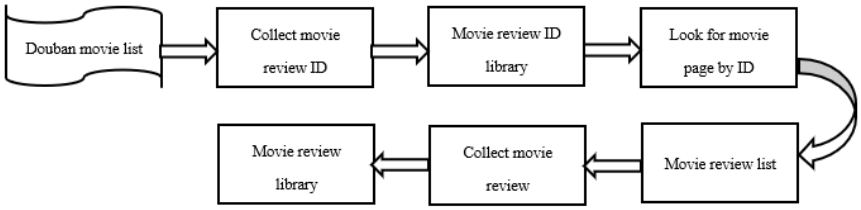
The experimental data of this paper is crawled from Douban film by Scrapy. The process of data crawling is shown in Figure 4. More than 520,000 film review text data are crawled for the experiment. There are several types of dirty data that exist in the data collection of movie comments:

- There are some comments without any content about the film, such as advertisements mixed in the comments. Such data have no value on the construction of sentiment lexicon, but also recede the program performance and algorithm accuracy.
- Comments with less than five words. Most of this kind of film critics carries a lot of information, but they are useless for the construction of emotional lexicon.

- Critics written in traditional Chinese. Traditional Chinese text needs to be translated into simplified Chinese text, and then be proceeded to the next step.

After data preprocessing of the initial film critics dataset, we have got more than 25,000 film review data, which are written in simplified Chinese, have a length greater than 50 words and have a clear meaning. We will use these film review data as test set filmReview_T1.

Figure 4 Review text crawler process



In this paper, we use 60 artificial tagged network buzzwords and 60 professional words for film critics. Words those appeared more than 20 times in the film review data are randomly selected into three groups as a test set. The composition of sentiment lexicon test set can be seen in Table 1.

Table 1 The design of contrast experiments

Set	*Positive*	*Negative*	*Total*
Common test set 1 (CT_1)	2,150	1,875	4,025
Common test set 2 (CT_2)	1,465	1,962	3,427
Domain test set (DT)	40	40	80

This paper sets up two test sets: generic emotional words test set and domain emotional words test set. The universal emotional words test set is a combination of NTUSD, Student Sentiment Lexicon, HowNet emotional lexicon which contains self-marked data. We select words whose frequency is between 20 and 50 and whose frequency is greater 50 as common test set 1 and 2. And then we select both 40 positive words and 40 negative words from network buzzwords and professional words of film critics with artificial tags. They constitute the domain emotional words test set.

The emotional words in the common test set will be labelled according to the emotional polarity in the basic sentiment lexicon. They are the basis for calculating the accuracy of our final emotional word classification algorithm. The domain-related emotional words selected from the network buzzwords will be labelled artificially as the domain test set. Table 1 shows the composition of the specific test set for the classification of emotional words.

6.2 Experimental methods and results

The construction process of DSL includes data preprocessing, seed words selection, candidate words emotional tendency calculation.

After the completion of the data preprocessing, we use jieba (https://pypi.python.org/pypi/jieba/) as a word segmentation tool. In the process of word segmentation, Sogou laboratory Internet thesaurus and manual collection of the domain thesaurus are imported as supplementation lexicon to improve the accuracy of word segmentation.

Then, those words are arranged according to their frequency. Words with the known emotional tendencies in the sentiment lexicon resources are removed. Nouns and adjectives with frequency of more than 10 are extracted as candidate words.

6.2.1 Comparison of seed words selection methods

In this paper, we use the word clustering method based on LDA topic model and K-means++ algorithm to select seed words. In this section, we select seed words in three different ways in order to compare their performances:

- Select both 40 positive and negative words with highest frequency of in the corpus respectively.
- Search for the terms in Google. Select both 40 positive and negative words with highest result respectively.
- According to the word clustering method proposed in this paper, we select both 40 positive and negative words respectively.

We select seed words in the test datasets CT_1, CT_2 and DT respectively by the above three methods, and use the selected seed words to calculate the emotional tendency of each word in the test set with traditional SO-PMI algorithm. Table 2 gives the results of the experiment using different methods and different test sets. The last column in the table gives the weighted average accuracy for each method.

Table 2 Contrast of selection methods of seed words

Method	*CT_1*	*CT_2*	*DT*	*AP*
M_1	0.786	0.792	0.788	0.789
M_2	0.766	0.800	0.863	0.783
M_3	0.776	0.809	0.900	0.792

It can be found that M_3 has improved the average accuracy of the construction of the sentiment lexicon, which proves that the method of seed words selection proposed in this paper is reasonable and effective. And in the process of longitudinal analysis of experimental data, we can find that the accuracy rate of M_3 in the domain emotional word set DT has reached a certain height, and the increase is much more outstanding than which in the general emotional word. It is obvious that this method is also appropriate for domain emotional words classification. This is mainly because the word clustering method proposed in this paper not only takes the representation and emotional polarity of the seed words in the general language environment into account, but also considers the representation and emotional polarity in corpus of certain domain.

6.2.2 Improved SO-PMI algorithm performance test

In this paper, in order to solve the problem of data sparse, the original SO-PMI algorithm is improved by using the LDA-based synonym group construction and co-distance. This section will compare the performance of SO-PMI algorithms which are improved in different aspects. First of all, we construct test sets for word classification based on the emotional words in Table 1. We use three different SO-PMI-based algorithms to test the performance separately:

- $SO\text{-}PMI_0$ represents the traditional SO-PMI algorithm
- $SO\text{-}PMI_1$ represents the SO-PMI algorithm with synonym group selection using the synonym word forest of HITO
- $SO\text{-}PMI_2$ represents SO-PMI algorithm proposed in this paper.

Table 3 shows the performance for each emotional word classification method, including the accuracy rate P, the recall rate R, the F value, the average accuracy rate AP, the average recall rate AR and the average F value AF.

Table 3 Improved SO-PMI algorithm performance test

Test set	*Evaluation*	*SO-PMI$_0$*	*SO-PMI$_1$*	*SO-PMI$_2$*
CT_1	P	0.864	0.883	0.898
CT_1	R	0.681	0.812	0.91
CT_1	F1	0.762	0.877	0.904
CT_2	P	0.812	0.824	0.921
CT_2	R	0.713	0.846	0.883
CT_2	F1	0.759	0.854	0.902
DT	P	0.83	0.851	0.907
DT	R	0.736	0.798	0.897
DT	F1	0.780	0.874	0.902
AVG	AP	0.674	0.681	0.931
AVG	AR	0.714	0.812	0.874
AVG	AF	0.693	0.772	0.902

It can be found that the average accuracy of $SO\text{-}PMI_2$ on the construction of emotional lexicon has a certain degree of improvement, which proves that the proposed method of SO-PMI can actually improve the performance of sentiment analysis. Moreover, in the process of analysing the experimental data vertically, the accuracy rate of $SO\text{-}PMI_2$ is higher than that of the $SO\text{-}PMI_1$ in the domain emotional word set DT, and the difference is more distinct than that in common sentiment words. It shows that $SO\text{-}PMI_2$ can promote the ability of processing emotional words in certain domain. This is mainly because the domain emotional words and emerging popular words are generally rarer in the traditional lexicon resources, so their synonyms groups are more difficult to be found. The proposed synonym group construction method can find synonyms from the corpus environment. It has better processing capacity to the emerging words and domain emotional words. $SO\text{-}PMI_1$ and $SO\text{-}PMI_2$ have little difference in the processing capacity of CT_1 to CT_2, which indicates that we can use the method of

combining synonym lexicon and LDA-based synonyms construction in dealing with general emotional words. Constructing the synonym group for words hard to be found in the synonym lexicon can improve the efficiency of lexicon construction.

6.2.3 Domain sentiment lexicon expansions and emotional tendency changes

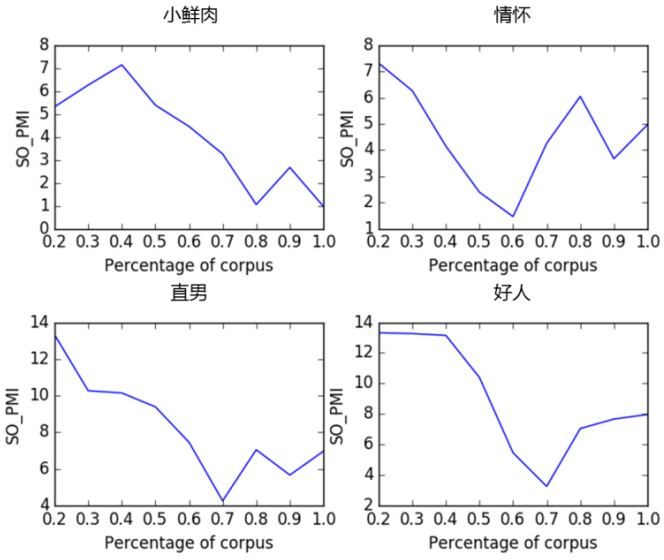
In order to test the process of IDSL's ability of self-expansion and adjustment, we divide the film review data into two parts according to the date of film and comment, and uses the film and television comment data before 2013 as the initial evaluation dataset, which is used to construct the initial lexicon, data after 2014 are divided into 3,000 texts per batch for experiment of the self-learning process of DSL.

Table 4 shows the accuracy of the construction method of DSL under different batches of data in the overall test set. The result includes TPR, FNR and AP. The experimental results show that the intelligent constructing and self-expansion method of DSL can improve the accuracy rate and the coverage of the words according to the growth of the corpus data.

Table 4 The accuracy of DSL varies with the corpus growth

Lexicon	*TPR*	*FNR*	*AP*
L	0.891	0.883	0.886
$L_{0.8}$	0.787	0.656	0.719
$L_{0.5}$	0.527	0.478	0.502
$L_{0.2}$	0.265	0.193	0.228

Figure 5 Change in emotional tendencies (see online version for colours)



In the experimental results, we can find some words such as '小鲜肉', '情怀', '直男', '好人' and so on. As time goes on, the emotional tendencies of these words tend to be neutral and even derogatory from the previous commendation, as Figure 5 shows that. This also reflects the Chinese language environment, especially in the buzzword environment. Due to changes in the network language environment and some other reasons, some words and phrases often produce different emotional tendencies at different time. This also shows that the method of constructing the lexicon proposed in this paper is more sensitive to the change of the word's emotion tendency, which is also difficult for the traditional emotional lexicon resources.

7 Conclusions

In this paper, LDA is used to construct the eigenvector of seed words, and then the seed words are selected by the cluster centre selection of K-means++ algorithm. Not only that, the original SO-PMI algorithm is improved by introducing the co-distance and the synonym group mining method based on LDA word similarity calculation. The initial domain sentiment lexicon is constructed with the improved SO-PMI algorithm and corpus. And the DSL's self-learning method is designed for the dynamic data. The experiment is conducted based on the film and television evaluation data from http://www.douban.com.

The experimental results show that the proposed method has a great improvement on the DSL construction, and has good processing ability for dynamic data. In the future work, we will improve the existing IDSL construction method, in order to achieve a fuzzy function, and the output of specific emotional values, and import the concept of time window, making the sentiment lexicon keen to reflect the change of current words' emotion.

This theory in this paper can be used to determine the emotional inclination of some short texts, which can be applied to the field of service or public opinion analysis.

References

Arthur, D. and Vassilvitskii, S. (2007) 'k-means++: the advantages of careful seeding', *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp.1027–1035.

Cai, Y., Chen, X., Peng, P.X. and Huang, J.Z. (2014) 'A LDA feature grouping method for subspace clustering of text data', *Pacific-Asia Workshop on Intelligence and Security Informatics*, pp.78–90.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013) 'New avenues in opinion mining and sentiment analysis', *IEEE Intelligent Systems*, Vol. 28, No. 2, pp.15–21.

Cheng, W., Song, Y., Zhu, Y. and Jian, P. (2016) 'Dimensional sentiment analysis for chinese words based on synonym lexicon and word embedding', *2016 International Conference on Asian Language Processing (IALP)*, pp.312–316.

Damani, O.P. (2013) *Improving Pointwise Mutual Information (PMI) by Incorporating Significant Co-occurrence*, arXiv preprint arXiv:1307.0596.

De Vylder, B. and Tuyls, K. (2005) 'Towards a common lexicon in the naming game: The dynamics of synonymy reduction', *BNAIC*, pp. 112–119.

Deng, S., Sinha, A.P. and Zhao, H. (2017) 'Adapting sentiment lexicons to domain-specific social media texts', *Decision Support Systems*, Vol. 94, pp.65–76.

Duyu, T., Bing, Q., LanJun, Z., KamFai, W., Yanyan, Z. and Ting, L. (2013) *Domain-Specific Sentiment Word Extraction by Seed Expansion and Pattern Generation*, arXiv preprint arXiv:1309.6722.

Fries, T.P. (2015) 'Fuzzy clustering of network traffic features for security', *2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*, pp.127–128.

Fu, X., Liu, W., Xu, Y. and Cui, L. (2017) 'Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis', *Neurocomputing*, Vol. 241, pp.18–27.

Ju, S., Li, S., Su, Y., Zhou, G., Hong, Y. and Li, X. (2012) 'Dual word and document seed selection for semi-supervised sentiment classification', *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp.2295–2298.

Kawabe, T., Namihira, Y., Suzuki, K., Nara, M., Sakurai, Y., Tsuruta, S. and Knauf, R. (2015) 'Tweet credibility analysis evaluation by improving sentiment dictionary', *2015 IEEE Congress on Evolutionary Computation (CEC)*, pp.2354–2361.

Kennedy, A. and Inkpen, D. (2006) 'Sentiment classification of movie reviews using contextual valence shifters', *Computational intelligence*, Vol. 22, No. 2, pp.110–125.

Laohakiat, S., Phimoltares, S. and Lursinsap, C. (2017) 'A clustering algorithm for stream data with lda-based unsupervised localized dimension reduction', *Information Sciences*, Vol. 381, pp.104–123.

Li, C., Yang, C. and Jiang, Q. (2017) 'The research on text clustering based on lda joint model', *Journal of Intelligent & Fuzzy Systems*, Vol. 32, No. 5, pp.3655–3667.

Liu, L., Lei, M. and Wang, H. (2013) 'Combining domain-specific sentiment lexicon with hownet for chinese sentiment analysis', *Journal of Computers*, Vol. 8, No. 4, pp.878–884.

Onan, A., Bulut, H. and Korukoglu, S. (2017) 'An improved ant algorithm with lda-based representation for text document clustering', *Journal of Information Science*, Vol. 43, No. 2, pp.275–292.

Pang, B., Lee, L., et al. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends® in Information Retrieval*, Vol. 2, Nos. 1–2, pp.1–135.

Paperno, D. and Baroni, M. (2016) 'When the whole is less than the sum of its parts: how composition affects PMI values in distributional semantic vectors', *Computational Linguistics*, Vol. 42, No. 2, pp.345–350.

Park, S. and Kim, Y. (2016) 'Building thesaurus lexicon using dictionary-based approach for sentiment classification', *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, pp.39–44.

Turney, P.D. and Littman, M.L. (2003) 'Measuring praise and criticism: Inference of semantic orientation from association', *ACM Transactions on Information Systems (TOIS)*, Vol. 21, No. 4, pp.315–346.

Wang, S-M. and Ku, L-W. (2016) 'Antusd: A large chinese sentiment dictionary', *LREC*.

Wawer, A. (2012) 'Mining co-occurrence matrices for so-pmi paradigm word candidates', *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp.74–80.

Wu, H-H., Tsai, A.C-R., Tsai, R.T-H. and Hsu, J.Y-J. (2013) 'Building a graded chinese sentiment dictionary based on commonsense knowledge for sentiment analysis of song lyrics', *J. Inf. Sci. Eng.*, Vol. 29, No. 4, pp.647–662.

Yu, H., Deng, Z-H. and Li, S. (2013) 'Identifying sentiment words using an optimization-based model without seed words', *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp.855–859.

Zagibalov, T. and Carroll, J. (2008) 'Automatic seed word selection for unsupervised sentiment classification of chinese text', *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1, pp.1073–1080.