Multimedia Tools and Applications (2020) 79:22355–22373
https://doi.org/10.1007/s11042-020-09030-1

Automatic construction of domain sentiment lexicon for semantic disambiguation

Yanyan Wang[1] · Fulian Yin[1] · Jianbo Liu[1] · Marco Tosato[2]

Received: 13 May 2019 / Revised: 14 February 2020 / Accepted: 5 May 2020 /
Published online: 22 May 2020

Abstract
Sentiment lexicon is used to judge the sentiments of words and plays a significant role in sentiment analysis. Existing sentiment lexicons ignore the sentimental ambiguity of words in different contexts and only assign sentiment positive or negative polarity for words. In this paper, we propose an automatic method for the construction of the domain-specific sentiment lexicon (SDS-lex) to avoid sentimental ambiguity, which incorporates the sentiment information not only from the existing lexicons but also from the corpus by using our improved TF-IDF algorithm (ITF-IDF). The ITF-IDF algorithm calculates the sentiment of words by considering both the importance of words and the distribution of different part-of-speech (POS) in a corpus labeled with different sentiment tendencies. Experiments on real-world datasets show that our constructed lexicon improves the sentimental ambiguity and outperforms many existing lexicons in terms of the coverage and the accuracy when performing text sentiment classification tasks.

Keywords Improved TF-IDF · Sentiment lexicon · Sentiment classification · Word sense disambiguation

1 Introduction

Sentiment analysis (SA) has recently received more and more attention as an important method to get information from texts with the continuous rise of social networks [5, 7, 13, 20, 33, 41]. Lexicon-based [33] and machine learning-based [12, 28] methods are two main ways for sentiment analysis. Especially, the lexicon-based approach, which assigns sentiment values for words, is still the mainstream for its simplicity [34].

Earlier researches on SA have been focus on building general sentiment lexicons [6, 18, 23], and SentiWordNet [2, 8] is the most famous one. It assigns the positive and negative sentiment values for each concept of synsets through the world knowledge base WordNet,

✉ Fulian Yin
yinfulian@cuc.edu.cn

1 Information Engineering Institute, Communication University of China, Beijing 100024, China

2 Laboratory for Industrial and Applied Mathematics, York University, Toronto, M3J 1P3, Canada

and the POS of each word is marked in different sentimental expressions. In addition, in 2005, Wilson et al. [39] constructed a subjective lexicon for the specific corpus of twitter, named MPQA Subjectivity Lexicon, which labels words with sentimental polarity, strength, and POS. These sentiment lexicons give corresponding sentiment polarities or values for many words, which will be used as prior sentimental knowledge in our paper. However, most of existing lexicons are general [2, 8, 15, 31, 39, 42] and have limited flexibility that cannot be applied in some specific domains directly. Hence there still exist some factors that are worth to be considered in further studies on this topic, such as the diversity of the sentiment of words in different contexts, the volatility of the sentimental polarity of words in different fields and the identification of the sentimental polarity for polysemous words.

The construction of sentiment lexicons for specific domains has become the key for solving the issues stated previously. This work usually requires the resources of domain-specific corpora, and the existing sentiment lexicons form the basis of it. In 2013, Mohammad et al. [24] transformed the sentiment lexicon construction problem into word classification by using SVM classifiers. Two sentiment lexicons, named NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon, were constructed in their paper using a large number of twitter corpora. Later, Vo and Zhang (2016) [36] used a simple neural network architecture to reassign word sentiments, which outperforms the NRC [24]. In 2014, Tang et al. [32] proposed word sentiment training based on neural networks. They extended the sentimental words using seed sets and constructed a large Twitter-based sentiment lexicon, TS-Lex. But their method requires to manually select the seed set in advance. Kiritchenko et al. [17] automatically generated sentiment lexicon from corpora with hashtagged emotion words such as #joy, #sad, and #angry, but they did not filter the training data carefully and ignored some cases of irony and bluntness. Deng et al. [5] proposed a model of hierarchical supervision to construct a topic-adaptive sentiment lexicon, taking into account topic and sentimental words. The above methods only distinguish different sentiments of words according to different corpora.

However, a word may still have multiple meanings even under the same corpus. We take the word "patient" as the example, Fig. 1 shows that the word "patient" expresses not only the meaning of a sick person but also a good character used to describe a person. To address the problem of semantic disambiguation, in 2019, Wu et al. [41] constructed a target-specific sentiment lexicon considering that a sentiment word may express different sentiment orientations when describing different targets. For example, in the sentence "The screen is too thin.", the "screen" is the target, and "thin" is the sentiment words to describe the target "screen". The applicability of the method in [41] is limited for reasons that it ignored that some sentences are incomplete and have grammatical structure problems.

Therefore, how to distinguish the sentiment expression of words is an issue worth to be considered when constructing a sentiment lexicon. From Fig. 1, the POS of the word "patient" is a noun with the meaning of a sick person and an adjective with the meaning of a person's good character. Hence, the POS of the word can be an indicator to discriminate different sentimental expressions. Motivated by this fact, we use the ITF-IDF algorithm in

Sentences	Sentiment word	Part-of-speech	Sentiment polarity
Jane is patient to children.	patient	adjective	
Now there is a patient in the class.	patient	noun	

Fig. 1 Take the word "patient" as an example, different representations of the same word

this paper and consider the distinction of sentimental words in different polarity corpora at the same time.

In our paper, we propose an automatic construction of the domain-specific sentiment lexicon, which integrates two parts of sentimental information: the sentiment values calculated by corpus-based ITF-IDF algorithm and the prior sentimental knowledge obtained from existing sentiment lexicons. Our ITF-IDF algorithm uses different polarity labels and POS factors to obtain a more refined sentimental value of words in a specific domain, while the traditional TF-IDF only focuses on the importance of words in different documents. Then we combine the prior sentimental knowledge with the sentimental information of words calculated from domain-specific corpus to overcome the serious dependence of words' sentiment on the corpus and to supplement sentiments of words not included in the training corpus. Finally, the experimental results show that the sentiment lexicon constructed in this paper outperforms the existing sentiment lexicons under sentiment classification task, and our method is more suitable for long text corpus than short text.

The rest of the paper is organized as follows: we introduce construction approaches of sentiment lexicons in Section 2, and describe our model in Section 3. Then, we present the results and performance comparisons in Section 4 followed by the conclusions and next research plan in Section 5.

2 Related works

In order to conduct sentiment analysis, researchers proposed many general or domain-specific sentiment lexicons using manual or automatic methods.

2.1 Resources of existing sentiment lexicons

The existing sentiment lexicons consist mostly of the general and the domain-specific lexicons. The general lexicons include General Inquirer (GI) [31], SentiWordNet [2, 8], Opinion Lexicon [14], etc. GI [31] is the earliest sentiment lexicon and labels each word with polarity, intensity, POS and so on. After that, the Opinion Lexicon proposed by Hu et al.[14] consists of many sentiment words and has wide applications in industry. Two years later, the SentiWordNet, derived from the WorldNet knowledge base, was proposed by Esuli and Sebastiani [2, 8]. In this lexicon, three scores are assigned for each synset of WordNet to describe the positive, negative and objective degree, which are further used to determine the precise sentiment of each word. However, most of the general lexicons are labeled manually, which not only is time-consuming, but also requires a lot of manpower.

Most of the domain-specific lexicons are constructed using the co-occurrence of words in domain-specific corpora, such as product reviews [7, 15],social platforms [11, 24, 32, 43, 44]. Wilson et al. [39] extended the lexicon MPQA Subjectivity Lexicon, which was firstly collected by Riloff and Wiebe et al. [29] from Multi-perspective Question Answering (MPQA) Opinion Corpus. The extended lexicon makes contribution to disambiguate the polarity of the sentiment expressions and marks each word with sentimental intensity and part of speech. Although this kind of sentiment lexicon considers the domain or topic of words, it ignores the distribution of words in different sentimental polar corpora and the fact that some words on the same topic may have different sentimental polarities, such as the word “pride”, which may have two meanings: proud and complacent under the same topic, leading to different sentiment expressions. In addition, since different languages have different syntactic structures, constructing sentiment lexicons based on different languages has become an important research trend, including Vietnamese [34], Malay [43], Korean [26], Chinese [42, 44], Arabic [1, 25], Slovene [3]. Note that the possible errors existed in corpus itself have serious impacts on the accuracy of the constructed domain-specific lexicons, that is, the domain-specific lexicons usually depend on the selected corpora. Hence we use the prior information from existing lexicons to solve the dependence problem.

2.2 Construction approaches of sentiment lexicons

Lexicon-based, corpus-based and hybrid methods are three main ways to construct sentiment lexicons [37].

The first category is based on some existing sentiment lexicons or knowledge bases. WordNet is one of the most commonly used lexicons, which builds sentiment lexicon through various synonyms and antonyms [8, 10, 16]. Some following researches construct sentiment lexicons based on WordNet. Kamps et al. [16] proposed a method using WordNet to measure semantic orientations of words, while they only considered the adjectives rather than the common words with other POS. Esuli et al. [2, 8] calculated the sentiments of concepts based on synonym relations and POS information in WordNet, and then obtained the sentiment of words by integrating sentiment values for the same word of different concepts. Their method is less currently used for reasons that it requires a lot of manpower and is inefficient. In addition, they did not make proper use of domain-specific corpus, leading to the failure of the sentiment analysis on words in a specific domain. For instance, the word "big" is a positive vocabulary when describing a display screen, but might have a negative sentiment when describing clothes.

The second method builds sentiment lexicons automatically by analyzing the association of words in the domain-specific corpus [4, 32, 36, 38, 42]. Tang et al. [32] first learned sentiment-specific phrase embedding (SSPE), and then converted the construction of sentiment lexicon into classification problem based on neural network training without manual operation. However, it highly depends on the corpus. Yang et al. [42] trained a large-scale corpus through a neural language model and built a sentiment lexicon with 10 dimensions of sentimental labels. This sentiment lexicon presents a more detailed judgement for the sentiment of words by extending the original three classes to the current ten classes, but it does not consider the disambiguation of the same word.

The last method combines the corpora with the existing sentiment lexicon to jointly learn sentiment lexicons [9, 12, 19, 40, 44]. Liu et al. [19] constructed a basic micro-blog sentiment lexicon by integrating the current sentiment lexicon and extracted new words based on the CHI square and improved SO-PMI algorithms. However, this method is not robust because SO-PMI algorithms need to manually select seeds. Wu et al. [40] constructed the domain-specific sentiment lexicon by integrating information from multiple resources, including domain-specific word-sentiment knowledge extracted from corpus, sentiment similarity knowledge extracted from all the messages, and prior sentiment knowledge extracted from existing sentiment lexicons. However, they did not consider that a word in different contexts may have different polarities. Recently, Han et al. [12] used mutual information with POS to generate a sentiment lexicon for the specific domain, and achieved good results in sentiment analysis tasks. However, the mutual information needs to manually set the seed set of emotion, and this limitation increases the instability of this model. The improved TF-IDF method adopted in our model solves this problem by considering the different distribution of the POS for each word in the same sentimental category. Our method provides a more detailed and stable sentiment analysis of words than the mutual information method.

Some attempts have been made to prevent the sentimental ambiguity problem in the construction of sentiment lexicon. Lu et al. [21] integrated information from different sources to learn a context-based sentiment lexicon. This method can make different words have different sentimental representations under different topics. Saif et al. [30] used context and semantic information extracted from specific domains to update the sentiment tendency of words together to alleviate the difference in lexical sentiment when context changes. Deng et al. [5] proposed topic-adaptive sentiment lexicon (TaSL) for higher-level classification tasks, which jointly considers the topics and sentiments of words to capture different sentiment expressions of a word under different topics. However, since one word may have different polarities and intensity expressions under different contexts and topics, it is not reasonable enough to do sentiment analysis based on context or topic alone.

3 Constructing a domain-specific sentiment lexicon based on improved TF-IDF algorithm

The existing sentiment lexicons fail to express precise sentiments of words which have different meanings in different contexts, and a simple polarity label without the intensity of sentiment contains limited information. For example, if the word "flag" is used as a noun, it means a banner (referring to a country or organization and its beliefs and values) and expresses a positive tendency. However, when it is used as a verb, it means attenuating fatigue and expresses a negative sentiment. In addition, some words just appear in specific domains, and we cannot find the sentimental intensities of these words in general sentiment lexicons.

To solve these problems, we propose a framework to automatically construct domain-specific sentiment lexicon, which jointly utilizes specific domain corpus and existing sentiment lexicons. Figure 2 shows the overall framework, consisting of three parts: corpus-based sentiment calculations based on improved TF-IDF algorithm, extracting prior sentimental knowledge from existing sentiment lexicons and the construction of domain-specific sentiment lexicon. For each word, we tag the fine-grained word meanings using POS method, and this step can reduce the semantic disambiguation. Then, the different sentiment values of a word with different POS are calculated through the ITF-IDF algorithm. Due to vocabulary limitations of the specific domain corpus, we also include the existing

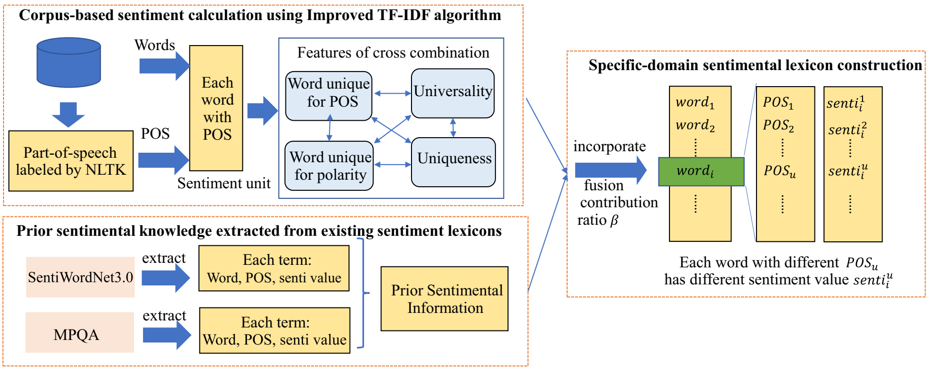

Fig. 2 The overall framework of our model

general sentiment lexicon to construct a domain-specific sentiment lexicon, which aims to increase the word coverage and fine-tune the sentimental values of words.

3.1 Corpus-based sentiment calculations

One of the most common methods to construct sentiment lexicons is the pointwise mutual information (PMI) [35]. Despite being widely used, the clear disadvantage of this method is that some words need to be manually selected as positive or negative seed sets. In this paper, we propose the improved TF-IDF (ITF-IDF) algorithm and incorporate the POS information with the distribution of words in different sentimental corpora to calculate the importance of feature words and reduce uncertainties in the training process.

Assuming that the word w_i and its corresponding POS p_u construct a new sentimental unit $w_i^{p_u}$, x_j is the corresponding sentimental polarity and has two states: the positive polarity x_p and the negative polarity x_n. When calculating the sentiment of $w_i^{p_u}$, we set four hypotheses:

- H1: Word unique for POS: The sentimental intensity of $w_i^{p_u}$ is positively correlated to the ratio of the frequency of $w_i^{p_u}$ and the frequency of w_i in x_j corpus.
- H2: Word unique for polarity: The sentimental intensity of $w_i^{p_u}$ is positively correlated to the ratio of the frequency of $W_i^{p_u}$ and the number of x_j corpus in each polarity corpus.
- H3: Uniqueness: The sentimental intensity of $W_i^{p_u}$ is positively correlated to the absolute value of the frequency difference of the words in positive and negative corpora.
- H4: Universality: The sentimental intensity of $W_i^{p_u}$ is negatively correlated to the total frequencies of $W_i^{p_u}$ in positive and negative corpora.

These properties are key in determining the sentimental intensity of words. H1 is translated mathematically into:

$$F_1(w_i^{p_u}, x_j) = N(\frac{c(w_i^{p_u}, x_j)}{\sum_u c(w_i^{p_u}, x_j)}), \quad (1)$$

where $c(w_i^{p_u}, x_j)$ is the frequency of $w_i^{p_u}$ in the corpora labeled x_j, the denominator is the total frequency of w_i with different POS in x_j corpus. $N(\cdot)$ represents that the results are normalized, which is used to stabilize the value of F_1 between 0 and 1. Formula (1) measures the sentimental intensity within a specific POS.

H2 is formalized as:

$$F_2(w_i^{p_u}, x_j) = N(\frac{c(w_i^{p_u}, x_j)}{c(x_j)}), \quad (2)$$

where $c(x_j)$ is the number of corpora labeled x_j, $N(\cdot)$ represents that the results are normalized, which is used to stabilize the value of F_2 between 0 and 1. Formula (2) indicates the distribution of the $w_i^{p_u}$ in the x_j corpora, which is used to measure the importance of the word in a certain polarity category.

H3 is stated as:

$$F_3(w_i^{p_u}) = N(\frac{|c(w_i^{p_u}, x_p) - c(w_i^{p_u}, x_n)|}{\sum_j c(w_i^{p_u}, x_j)}), \quad (3)$$

where $c(w_i^{p_u}, x_p)$, $c(w_i^{p_u}, x_n)$ and $\sum_j c(w_i^{p_u}, x_j)$ are respectively the frequencies of $w_i^{p_u}$ in positive, negative and both corpora, and $N(\cdot)$ represents that the results are normalized, which is used to stabilize the value of F_3 between 0 and 1. The absolute value of the frequency difference is useful to remove some of the words which come up frequently in corpora and to select some of the most relevant words for each of the polarity categories. Formula (3) guarantees the uniqueness of the distribution of a sentimental word with a specific POS in corpora. H4 is defined as:

$$F_4(w_i^{p_u}, x_j) = N(\log(\frac{\sum_i c(w_i^{p_u}, x_j)}{c(w_i^{p_u}, x_p) + c(w_i^{p_u}, x_n) + \varepsilon})), \tag{4}$$

where ε is a constant to guarantee that the denominator is not zero; $N(\cdot)$ represents that the results are normalized, which is used to stabilize the value of F_4 between 0 and 1. Formula (4) shows the universality of the distribution of a sentimental word within a specific POS in corpora. The sentimental value of any $w_i^{p_u}$ in x_j corpora is calculated as:

$$S_I(w_i^{p_u}, x_j) = \prod_{n=1}^{4} F_n, \tag{5}$$

where F_n expresses the n-th result in the above four hypotheses, $S_I(w_i^{p_u}, x_j)$ indicates sentimental intensity of $w_i^{p_u}$ with respect to x_j tendency and we adopt multiplication to represent the influence of each factor on the sentimental outcome. Each sentimental value is then obtained by subtracting the sentimental value of $w_i^{p_u}$ in the positive and negative polarity categories:

$$S_I(W_i^{p_u}) = S_I(W_i^{p_u}, x_p) - S_I(W_i^{p_u}, x_n), \tag{6}$$

where $S_I(w_i^{p_u}, x_p)$ is the sentimental value in positive category, $S_I(w_i^{p_u}, x_n)$ is the sentimental value in negative category and $S_I(w_i^{p_u})$ is the sentimental value of the $w_i^{p_u}$.

3.2 Constructing a domain-specific sentiment lexicon based on multiple POS

We introduce the ITF-IDF algorithm to solve word sense disambiguation and discover new words, but it has a poor coverage for general sentimental words when using small training corpus. In order to solve this issue, we propose a domain-specific sentiment lexicon by integrating prior sentiment knowledge with the corpus-based sentiment information.

We first collect the existing lexicon MPQA and SentiWordNet3.0 to acquire prior sentimental knowledge. The former contains word polarities and intensity labels without specific sentimental values. For any sentiment word W_i and POS p_u, the prior sentimental value $S_M(W_i^{p_u})$ is defined similar to that of in MPQA

$$S_M(W_i^{p_u}) = \begin{cases} \alpha(x) & \text{, if } W_i^{p_u} \text{ is postive in MPQA} \\ -\alpha(x) & \text{, if } W_i^{p_u} \text{ is negative in MPQA} \end{cases}, \tag{7}$$

$$\alpha(x) = \begin{cases} 1 & \text{, if } W_i^{p_u} \text{ is strongsubj in MPQA} \\ 0.5 & \text{, if } W_i^{p_u} \text{ is weaksubj in MPQA} \end{cases}, \tag{8}$$

where $\alpha(x)$ is used to express the intensity value of each word since the MPQA contains only the sentiment intensity labels. Its value is determined by the intensity label, where the label "strongsubj" denotes the word has a strong sentiment, "weaksubj" denotes the word has a weak sentiment.

SentiWordNet3.0 is a concept-based sentiment lexicon where each concept includes a POS, a sentiment value, a similar word set, etc. But it ignores the fact that some words may have multiple sentiment values though they have the same POS. To address this problem, we calculate the mean for multiple sentiments and obtain the prior sentiment value:

$$S_W(W_i^{p_u}) = \frac{\sum_d^D s(W_i^{p_u}, v_d)}{D}, \tag{9}$$

where $S_W(W_i^{p_u})$ is the prior sentiment value of any $W_i^{p_u}$ using SentiWordNet3.0, D represents the number of times $W_i^{p_u}$ appeared in SentiWordNet3.0, $s(W_i^{p_u}, v_d)$ indicates the sentimental value of any $W_i^{p_u}$ contained in the similar words of the concept. For each $W_i^{p_u}$, the prior sentimental value $S_P(W_i^{p_u})$ is calculated as:

$$S_P(W_i^{p_u}) = \frac{S_M(W_i^{p_u}) + S_W(W_i^{p_u})}{2}, \tag{10}$$

Finally, the prior sentimental knowledge is combined with corpus-based ITF-IDF sentimental information to build our domain-specific sentiment lexicon:

$$S_U(W_i^{p_u}) = (1 - \beta) \times S_I(W_i^{p_u}) + \beta \times S_P(W_i^{p_u}). \tag{11}$$

where the $S_U(W_i^{p_u})$ represents the sentimental value obtained by incorporating the prior knowledge and corpus-based ITF-IDF sentimental information, β is the fusion contribution ratio, varying between 0 and 1, which is used to adjust the proportion between prior knowledge and corpus-based sentimental knowledge.

3.3 Evaluation methods

We choose different measure standards defined as follows, including the precision (P), recall (R), F1, accuracy (Acc) and coverage(Cov), to test the efficiency of our proposed sentiment lexicon in performing text sentiment classification tasks.

For a text sequence $x = w_1, ..., w_k, ..., w_K$, we firstly label corresponding POS tags using NLTK a python tool and obtain the new sequence expression $x(w, p) = w_1^{p_1}, ..., w_k^{p_k}, ..., w_K^{p_K}$, where $w_k^{p_k}$ indicates that the k-th word in sequence s with POS tag p_k, K is the total word number for the sequence s. Then we obtain the corresponding sentimental values $Sen_x = (S_U(w_1^{p_1}), ..., S_U(w_k^{p_k}), ..., S_U(w_K^{p_K}))$ for each $w_k^{p_k}$ in x to synthesize the final sentence sentiment value S_x:

$$S_x = \sum_{k}^{K} S_U(w_k^{p_k}), \tag{12}$$

where $S_U(w_k^{p_k})$ is the sentiment value of the k-th word in sequence x with POS tag p_k calculated by our method, S_x is the final sentiment value of the sequence x. We turn S_x into polarity value T_x as following for the convenience of the judgement of the performance of our sentiment classification task.

$$T_x = \begin{cases} 1 & , \text{if } S_x > 0 \\ 0 & , \text{if } S_x = 0 \\ -1 & , \text{if } S_x < 0 \end{cases}, \tag{13}$$

We further count the sum of T_x of each case, denoted by NP, NN, and ZN, respectively.

We finally adopt the precision (P), the recall (R), F1, the coverage (Cov) and the accuracy (Acc) to evaluate our proposed method. For simplicity, we only present detailed definitions of the measures P(pos), R(pos), and F1(pos) on the positive tendency.

$$P(pos) = \frac{TP}{PN}, \tag{14}$$

$$R(pos) = \frac{TP}{N_{pos}}, \tag{15}$$

$$F1(pos) = \frac{2 \times P(pos) \times R(pos)}{P(pos) + R(pos)}, \tag{16}$$

Table 1 Experimental datasets

Dataset	Positive	Negative	Average length of per sentence	level
LMRD	5000	5000	231	Long text
MRD	5000	5000	21	Short text

where TP is the number of correct positive sentences using our proposed lexicon, N_{pos} is the true positive sentences in evaluation dataset. In particular, F1 is a trade-off between precision and recall. The measures P(neg), R(neg) and F1(neg) on the negative tendency can be defined in a similar way. The Acc and Cov are the measures of the accuracy and coverage of the sentiment classification task performed in our experiment, which are defined as following:

$$Acc = \frac{TP + TN}{N}, \tag{17}$$

$$Cov = \frac{NP + NN}{N}. \tag{18}$$

where TN is the number of correct negative sentences using our proposed lexicon, N is total number of evaluation datasets. The Acc reflects the performance of our model on sentiment classification, and the Cov indicates the word coverage of the sentiment lexicon we constructed.

4 Experiments

In this section, we conduct comparative experiments to test the efficiency of our constructed sentiment lexicon.

4.1 Experimental data

Our experimental data sets are from the Large Movie Review Dataset(LMRD) [22] and the Movie-review-data (MRD) [27]. The Large Movie Review Dataset[1] provided by Maas et al. [21] is extracted from IMDB. It contains 50000 labeled samples (positive and negative corpora, each of 25000 cases) and 50000 unlabeled samples. The Movie-review-data[2] provided by Pang and Lee [26] is extracted from Rotten Tomatoes website, consisting of 5331 positive and negative short text corpora.

To improve the robustness of our results, we use the method of ten-fold cross-validation in our experiments. In particular, in order to keep a balance distribution between training and test corpus, we select equal numbers (5000 of each) of the positive and negative reviews in each experiment by selecting respectively from each dataset randomly, respectively. Hence, our training dataset has 4500 positive and negative sentences, and the test dataset contains 500 positive and negative sentences in each experiment. Table 1 reports some detailed features of the datasets.

[1] http://ai.stanford.edu/~amaas//data/sentiment/

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data

4.2 Experiments on construction of our SDS-lex

In this section, we analyze the important steps in the process of the construction of the domain-specific sentiment lexicon. One step is the determination of the parameter β, which weights the corpus-based sentiment value and prior knowledge-based sentiment value. The four factors defined in formulas (1)-(4) have different influences on the sentiment calculation of words. Formula (5) multiplies these factors to measure the total impact. Another step is to explore how these factors influence the construction of sentiment lexicon. As shown in Table 2, a total of 15 features are obtained by considering each hypothetical factor separately or combining some of them together.

4.2.1 Determination of adjustment factors β

The sentiment lexicon constructed in this paper use both the prior knowledge from the existing lexicons and the sentiment values calculated based on the corpus, where β and 1-β are the weights put on this two values. Hence the determination of β is important in applications.

We take different values for β to perform sentiment classification tasks under different features in order to find if there is an optimal β that applies to the case of each or most of the features. The MRD corpus is used to train our model using the method of ten-fold cross-validation in our experiment. Figure 3 plots the accuracy of our experiment under different features with different values of β. We find that the accuracy achieves the maximum value when β is around 0.2, and this conclusion is valid for the cases of most different features except two features F_1 and $F_1 * F_4$. Hence we select 0.2 as the value of β in the following experiments.

4.2.2 Exploring the importance of different features

The construction of our SDS-lex depends on the selected features. We explore how these different 15 features reported in Table 2 influence the sentiment classification task.

Based on the short and long text corpora, Tables 3 and 4 respectively report the performance of our method under different features, where the measures Cov, Acc, P(pos), R(pos), F1(pos), P(neg), R(neg), F1(neg) defined in previous section denote word coverage, classification accuracy, and the effects on positive and negative text, respectively, the entries in bold denote the measure that achieves the optimal value among different features.

Table 2 15 features obtained from hypothetical factors and their combination

No.	Features	No.	Features
1	F1	9	F2*F4
2	F2	10	F3*F4
3	F3	11	F1*F2*F3
4	F4	12	F1*F2*F4
5	F1*F2	13	F1*F3*F4
6	F1*F3	14	F2*F3*F4
7	F1*F4	15	F1*F2*F3*F4
8	F2*F3		

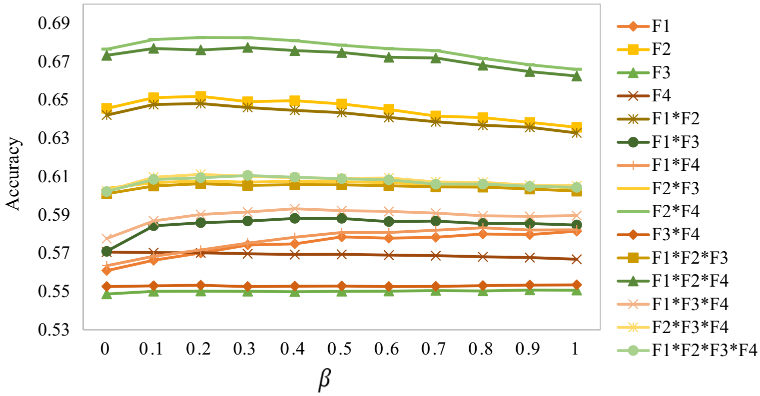

Fig. 3 Accuracy of the sentiment classification task under different features using different values of β based on Movie-review-data (MRD) corpus

From Table 3, we know that the Cov based on short text corpus under most features is greater than 98%, P and R have different values for the positive and negative texts, and the feature $F_2 * F_4$ achieves the best performance in terms of its F1 and Acc values. From Table 4, the performance of the sentiment classification task based on the long text has obvious advantages in terms of all the measures than that of based on the short text.

Therefore, the performance of the sentiment classification task is clearly dependent on the selected features, that is, the importance of different features is quite different. In order to further explore the performance of each feature, we conduct the sentiment classification tasks under different features using the MRD dataset and the LMRD dataset, and show

Table 3 Performance of the sentiment classification task measured by F1(pos), F1(neg) and Acc under different features using MRD dataset

Features	Cov	P(pos)	R(pos)	F1(pos)	P(neg)	R(neg)	F1(neg)	Acc
F_1	0.9902	0.5683	0.6568	0.6093	0.5864	0.4834	0.5297	0.5701
F_2	0.9961	0.6422	0.6996	0.6695	0.6696	0.6042	0.6350	0.6519
F_3	0.9157	0.5819	0.6866	0.6299	0.6353	0.4138	0.5010	0.5502
F_4	0.9157	0.6006	0.6940	0.6439	0.6611	0.4464	0.5327	0.5702
F_1*F_2	**0.9963**	0.6379	0.6982	0.6665	0.6665	0.5980	0.6300	0.6481
F_1*F_3	0.9867	0.5769	0.7154	0.6386	0.6227	0.4564	0.5265	0.5859
F_1*F_4	0.9875	0.5699	0.6574	0.6104	0.5918	0.4860	0.5336	0.5717
F_2*F_3	0.9959	0.5842	0.7766	0.6664	0.6634	0.4388	0.5269	0.6077
F_2*F_4	**0.9963**	**0.6966**	0.6568	0.6759	0.6752	**0.7084**	**0.6912**	**0.6826**
F_3*F_4	0.9157	0.5849	0.6852	0.6310	0.6392	0.4214	0.5077	0.5533
$F_1*F_2*F_3$	0.9959	0.5831	0.7746	0.6650	0.6611	0.4380	0.5257	0.6063
$F_1*F_2*F_4$	**0.9963**	0.6883	0.6540	0.6706	0.6700	0.6982	0.6837	0.6761
$F_1*F_3*F_4$	0.9867	0.5809	0.7148	0.6409	0.6271	0.4656	0.5342	0.5902
$F_2*F_3*F_4$	0.9959	0.5826	**0.8154**	**0.6790**	**0.6900**	0.4070	0.5096	0.6112
$F_1*F_2*F_3*F_4$	0.9959	0.5813	0.8120	0.6770	0.6863	0.4068	0.5087	0.6094

Table 4 Performance of the sentiment classification task measured by F1(pos), F1(neg) and Acc under different features using LMRD dataset

Features	Cov	P(pos)	R(pos)	F1(pos)	P(neg)	R(neg)	F1(neg)	Acc
F_1	1.0000	0.5745	0.7700	0.6579	0.6527	0.4294	0.5175	0.5997
F_2	1.0000	0.7305	0.7908	0.7592	0.7722	0.7070	0.7378	0.7489
F_3	1.0000	0.5924	0.8380	0.6940	0.7241	0.4220	0.5329	0.6300
F_4	1.0000	0.6056	0.8404	0.7038	0.7403	0.4512	0.5602	0.6458
F_1*F_2	1.0000	0.7303	0.7746	0.7515	0.7601	**0.7126**	0.7353	0.7436
F_1*F_3	1.0000	0.5804	0.7606	0.6583	0.6536	0.4498	0.5326	0.6052
F_1*F_4	1.0000	0.5738	0.7610	0.6541	0.6465	0.4344	0.5192	0.5977
F_2*F_3	1.0000	0.5808	0.9536	0.7216	0.8690	0.3092	0.4536	0.6314
F_2*F_4	1.0000	**0.7323**	0.8884	**0.8015**	0.8588	0.6690	**0.7488**	**0.7787**
F_3*F_4	1.0000	0.5928	0.8384	0.6944	0.7250	0.4226	0.5336	0.6305
F_1*F_2*F_3	1.0000	0.5837	0.9508	0.7231	0.8654	0.3194	0.4642	0.6351
F_1*F_2*F_4	1.0000	0.7297	0.8738	0.7940	0.8437	0.6706	0.7441	0.7722
F_1*F_3*F_4	1.0000	0.5827	0.7702	0.6634	0.6619	0.4480	0.5341	0.6091
F_2*F_3*F_4	1.0000	0.5973	**0.9588**	0.7356	**0.8951**	0.3494	0.4974	0.6541
F_1*F_2*F_3*F_4	1.0000	0.5995	0.9548	0.7360	0.8892	0.3578	0.5048	0.6563

their performances reported by three measures F1(pos), F1(neg) and Acc in Figs. 4 and 5, respectively.

It can be seen from Figs. 4 and 5 that (1) the tasks conducted by using two different datasets have similar performances; (2) the optimal performance is achieved under the feature F_2 among the four separate features, which means that the distribution of vocabulary in different sentiment polarity categories is of great significance in the construction of sentiment lexicons; (3) the feature $F_2 * F_4$ always guarantees the optimal values of F1(pos), F1(neg) and Acc among all the features;(4) the results of $F_2 * F_4$ and $F_1 * F_2 * F_4$ have balanced performance in both positive and negative corpora, while other features are generally useful to single oriented corpora.

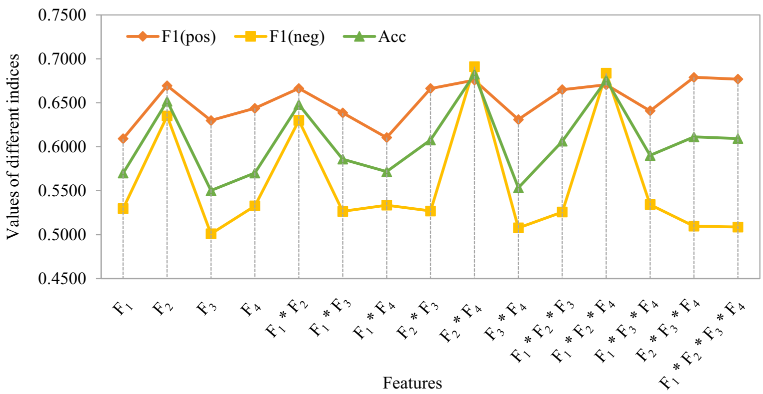

Fig. 4 The effects of the three indicators of accuracy, F1(pos), and F1(pos) on Movie-review-data (MRD)

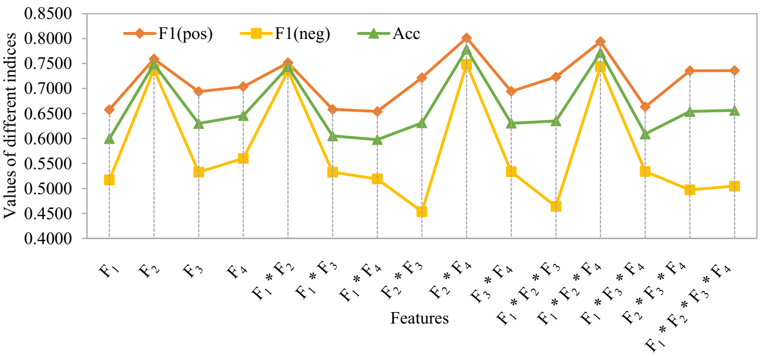

Fig. 5 The effects of the three indicators of accuracy, F1(pos), and F1(pos) on Large Movie Review Dataset(LMRD)

Therefore, we use $F_2 * F_4$ to calculate corpus-based sentiment values and combine it with the prior sentiment lexicons to build our SDS-lex.

4.3 Comparative experiment

We conduct sentiment classification tasks and compare our proposed method with some existing sentiment lexicons, including GI [31], MPQA [39], SW [2], NRC [24], S140 [24], ETSL [17], HIT [32], NN [36], and HSSWE [38].

Tables 5 and 6 respectively give the performance of the text sentiment classification tasks conducted by using different lexicons based on the short text corpus in MRD and the long text corpus in LMRD. It is clear that the Acc of our method is highest, which improves by

Table 5 Effects of text sentiment classification task using different features based on long text corpus MRD

Methods	Cov	P(pos)	R(pos)	F1(pos)	P(neg)	R(neg)	F1(neg)	Acc
General lexicon								
GI	0.7127	0.6055	0.5858	0.5952	0.6806	0.3118	0.4275	0.4488
Lexicons with part-of-speech								
MPQA	0.7202	0.6486	0.5104	0.5711	0.6455	0.4214	0.5097	0.4659
SW	0.9485	0.5787	0.6430	0.6090	0.6096	0.4788	0.5361	0.5609
Domain-specific lexicons								
NRC	0.9969	0.6730	0.4902	0.5671	0.5990	0.7580	0.6691	0.6241
S140	0.9958	0.5696	**0.8324**	**0.6764**	**0.6885**	0.3652	0.4771	0.5988
ETSL	0.7747	0.5349	0.6364	0.5812	0.6160	0.2216	0.3259	0.4290
HIT	0.9799	0.6268	0.6376	0.6321	0.6329	0.5966	0.6142	0.6171
NN	**0.9984**	0.6569	0.5192	0.5799	0.6025	0.7268	0.6588	0.6230
HSSWE	0.9955	0.5804	0.7550	0.6562	0.6491	0.4478	0.5297	0.6014
Our domain-specific lexicon								
Our SDS-lex	0.9963	**0.6966**	0.6568	0.6759	0.6752	**0.7084**	**0.6912**	**0.6826**

6% and 9% over the optimal results of the existing sentiment lexicons in both the short and long text corpora .

For short text corpora, Table 5 shows that our method has similar coverage to the other methods. However, our method requires less corpus than the methods of NN, NRC and S140. An additional observation from Table 5 is that the sentiment lexicons S140 and NN are only efficient for certain sentiment corpus, and the sentiment lexicons our SDS-lex, MPQA and SW have the potential for both the positive and negative corpora. This is because the latter three lexicons include part-of-speech information. In particular, our constructed lexicon outperforms MPQA and SW in terms of all the measures.

For long text corpora, Table 6 shows that the Cov of all sentiment lexicons are greater than 90% based on the long text corpora. We find again that the use of the POS information improves the applicability of sentiment lexicon.

In summary, the sentiment lexicon constructed by our method not only has the same coverage as the existing sentiment lexicons under large-scale corpus data, but also has high accuracy and good stability for both positive and negative corpora.

4.4 Effect of the dataset size of corpus on the construction of our SDS-lex

We choose different dataset sizes to test their effects on the construction of the SDS-lex. The selected dataset sizes of the positive and negative corpus are 1000, 3000, 5000, 7000, 9000, 11000, and 12500 from LMRD, respectively.

Figure 6 plots the detailed results and shows that the datasize has no effect on the coverage of the constructed sentiment lexicon. We find that the indicators P and R are fluctuating and the indicators F1 and Acc are increasing with the increase of the dataset size. Since F1 is a trade-off between P and R, and Acc is the accuracy of the sentiment lexicon in terms of performing classification task, we conclude that the performance of our constructed SDS-lex can be improved by increasing the dataset size.

Table 6 Effects of text sentiment classification task using different features based on long text corpus LMRD

Methods	Cov	P(pos)	R(pos)	F1(pos)	P(neg)	R(neg)	F1(neg)	Acc
General lexicon								
GI	0.9258	0.5759	0.8390	0.6829	0.7460	0.2940	0.4214	0.5665
Lexicons with part-of-speech								
MPQA	0.9358	0.6608	0.7120	0.6852	0.7074	0.5610	0.6253	0.6365
SW	0.9991	0.6007	0.8086	0.6892	0.7084	0.4616	0.5587	0.6351
Domain-specific lexicons								
NRC	**1.0000**	0.8891	0.2482	0.3876	0.5633	0.9688	0.7123	0.6085
S140	**1.0000**	0.5889	0.9024	0.7126	0.7915	0.3690	0.5027	0.6357
ETSL	0.9990	0.5209	**0.9736**	0.6787	**0.8025**	0.1034	0.1829	0.5385
HIT	0.9994	0.6640	0.7692	0.7126	0.7262	0.6098	0.6627	0.6895
NN	**1.0000**	0.8068	0.4522	0.5794	0.6196	0.8912	0.7309	0.6717
HSSWE	**1.0000**	0.6228	0.8374	0.7143	0.7525	0.4924	0.5951	0.6649
Our domain-specific lexicon								
Our SDS-lex	***1.0000***	**0.7323**	0.8884	**0.8015**	0.8588	**0.669**	**0.7488**	**0.7787**

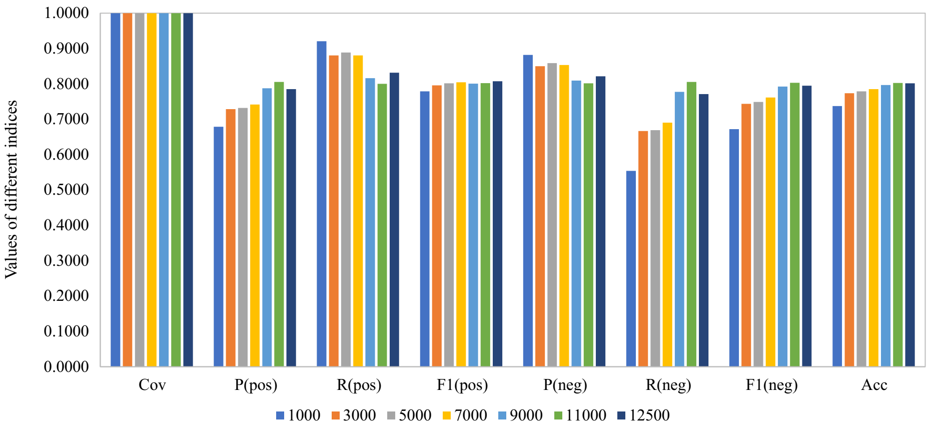

Fig. 6 The effects of the three indicators of accuracy, F1(pos), and F1(pos) on Large Movie Review Dataset(LMRD)

4.5 Case analysis

An advantage of our constructed SDS-lex is to identify different sentiments of the same word. We present a case study of the two words “flag” and “accent” contained in different sentences to analyze this advantage. Table 7 shows the sentiment polarity and the POS of the two words provided by our constructed sentiment lexicon. The word “flag” (“accent”) is a noun as well as a verb, which has different sentiment values in different sentences. We find from Table 7 that our constructed sentiment lexicon has the potential to identify this fact.

Note that since the same word usually has various forms, and it can increase the computation burden greatly if different forms are marked by different POS. Our constructed sentiment lexicon avoids this problem and tags the same word with different forms in the same way. The verbs,adjectives,adverbs and nouns including different forms are marked as “v”,“a”,“r”, and “n”, respectively, the POS of other infrequent words remain unchanged.

Table 7 Take the two words as examples to analysis the performance of word semantic disambiguation

Word	Sentiment	Part-of-speech	Sentences
flag	positive	n	mel gibson fights the good fight in vietnam in director randall wallace’s flag-waving war flick with a core of decency.
	negative	v	the film sometimes flags....
accent	positive	v	hoffman notches in the nuances of pain, but his smart, edgy voice and waddling profile accent the humor of wilson’s plight...
	negative	n	thoughtless, random, superficial humour and a lot of very bad scouse accents

5 Conclusion

In this paper, we constructed a sentiment lexicon by using both the prior information from existing lexicons and the corpus information in specific domains. Our ITF-IDF algorithm uses different polarity labels and POS factors to obtain a more refined sentimental value of words in a specific domain, which prevents the sentimental ambiguity problem appeared in many existing lexicons. Then we combine the prior sentimental knowledge with the sentiment values calculated from domain-specific corpus to overcome the serious dependence of words' sentiment on the corpus. This method also has quite effectiveness to supplement sentiments of words not included in the training corpus. We finally conducted text sentiment classification tasks on real datasets to test the efficiency of our constructed sentiment lexicon and found that our method has obvious advantages in comparison with some commonly-used methods.

For our future works, we would like to further study the tense and some other factors to construct a more fine-grained sentiment lexicon and diversify its evaluation methods.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 61801440), the High-quality and Cutting-edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China) and the Fundamental Research Funds for the Central Universities.

References

1. Assiri A, Emam A, Al-Dossari H (2018) Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. J Inf Sci 44(2):184–202
2. Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An Enhanced Lexical Resource For Sentiment Analysis and Opinion Mining. In: International conference on language resources and evaluation, Valletta
3. Bucar J, Znidarsic M, Povh J (2018) Annotated news corpora and a lexicon for sentiment analysis in Slovene. Lang Resour Eval 52(3):895–919
4. Cambria E, Poria S, Hazarika D, Kwok K (2018) Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings In: Proc. 32th Int. Conf. Assoc. Adv. Artif. Intell, pp 1795–1802
5. Deng D, Jing LP, Yu J, Sun S. L., Ng MK (2019) Sentiment lexicon construction with hierarchical supervision Topic Model. IEEE-ACM Trans Audio Speech Lang Process 27(4):704–718
6. Denecke K (2008) Using sentiwordnet for multilingual sentiment analysis. In: 2008 IEEE 24th international conference on data engineering workshop, pp 507–512
7. Dey A, Jenamani M, Thakkar JJ (2018) Senti-n-gram: An n-gram lexicon for sentiment analysis. Expert Syst Appl 103:92–105
8. Esuli A, Sebastiani F (2006) SentiWordNet: a publicly available lexical resource for opinion mining. In: International Conference on Language Resources and Evaluation (LREC-2006), pp 417-422
9. Feng J, Gong C, Li XD, Lau RYK (2018) Automatic approach of sentiment lexicon generation for mobile shopping reviews. Wireless Communications and Mobile Computing
10. Gatti L, Guerini M, Turchi M (2016) Sentiwords: deriving a high precision and high coverage lexicon for sentiment analysis. IEEE Trans Affect Comput 7(4):409–421
11. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. In: Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group
12. Han HY, Zhang JP, Yang J, Shen YR, Zhang YS (2018) Generate domain-specific sentiment lexicon for review sentiment analysis. Multimed Tools Appl 77(16):21265–21280
13. Hegazy AE, Makhlouf MA, El-Tawel GS (2019) Feature selection using chaotic salp swarm algorithm for data classification. Arab J Sci Eng 44(4):3801–3816
14. Hu MQ, Liu B (2004) Mining and summarizing customer reviews. In: ACM SIGKDD, pp 168-177
15. Khoo CSG, Johnkhan SB (2018) Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. J Inf Sci 44(4):491–511

16. Kamps J, Marx M, Mokken RJ, De RM (2004) Using wordNet to measure semantic orientations of adjectives. In: Proc. 4th Int. Conf. Lang. Resour. Eval, vol 4, pp 1115–1118
17. Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762
18. Kumari P, Haider MTU (2020) Sentiment analysis on aadhaar for Twitter data-a hybrid classification approach. Proceeding of International Conference on Computational Science and Applications: ICCSA 2019. Springer Nature, pp 309–318
19. Liu J, Yan M, Luo J (2016) Research on the construction of sentiment lexicon based on chinese microblog. In: International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, pp 56–59
20. Liu J, Fu X, Liu J et al (2018) Analyzing and assessing reviews on JD.com. Intell Autom Soft Comput 24(1):73–79
21. Lu Y, Castellanos M, Dayal U, Zhai C (2011) Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proc. 20th Int. Conf. World Wide Web (WWW), pp 347–356
22. Maas AL, Daly RE, Pham PT et al (2011) Learning word vectors for sentiment analysis. In: Meeting of the association for computational linguistics: Human language technologies, Portland, pp 142–150
23. Mandal S, Singh GK, Pal A (2020) Text summarization technique by sentiment analysis and cuckoo search Algorithm. Computing in Engineering and Technology. Springer, Singapore, pp 357–366
24. Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-canada: building the state-of-the-art in sentiment analysis of tweets. In: Second Joint Conference on Lexical and Computational Semantics, (SEM). Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol 2. Atlanta, pp 321–327
25. Al-Moslmi T, Albared M, Al-Shabi A, Omar N, Abdullah S (2018) Arabic senti-lexicon: constructing publicly available language resources for Arabic sentiment analysis. J Inf Sci 44(3):345–362
26. Chul-won NA, Choi M (2018) KNU Korean Sentiment lexicon: bi-LSTM-based method for building a Korean sentiment lexicon. J Intell Inf Syst 24(4):219–240
27. Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Meeting on Association for Computational Linguistics. Ann Arbor, pp 115–124
28. Rani S, Kumar P (2019) Deep learning based sentiment analysis using convolution neural network. Arab J Sci Eng 44(4):3305–3314
29. Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: EMNLP, pp 105–112
30. Saif H, Fernandez M, Kastler L, Alani H (2017) Sentiment lexicon adaptation with context and semantics for the social web. Semant Web 8(5):643–665
31. Stone PJ, Dunphy DC, Smith MS (1966) The general inquirer: a computer approach to content analysis. Inf Storage Retriev 4(4):375–376
32. Tang D, Wei F, Qin B et al (2014) Building large-scale Twitter-specific sentiment lexicon: a representation learning approach. In: COLING, pp 172–182
33. Tao W, Liu T, Yu W et al (2018) Building ontology for different emotional contexts and multilingual environment in opinion mining. Intell Autom Soft Comput 24(1):65–71
34. Tran TK, Phan TT (2018) A hybrid approach for building a Vietnamese sentiment dictionary. J Intell Fuzzy Syst 35(1):967–978
35. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, pp 417–424
36. Vo DT, Zhang Y (2016) Don't count, predict! An automatic approach to learning sentiment lexicons for short text. In: Proc 54th Annual. Meeting Assoc Comput. Linguist, pp 219–224
37. Wang K, Xia R (2016) A survey on automatical construction methods of sentiment lexicons. Acta Autom Sin 42(4):495–511
38. Wang Y, Zhang Y, Liu B (2017) Sentiment lexicon expansion based on neural PU learning, double dictionary lookup, and polarity association. In: Proc. Conf. Empirical Methods Natural Lang Process, pp 553–563
39. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, pp–354
40. Wu F, Huang Y, Song Y et al (2016) Towards building a high-quality microblog-specific Chinese sentiment lexicon. Decis Support Syst 87(C):39–49
41. Wu S, Wu F, Chang Y, Wu C, Huang Y (2019) Automatic construction of target-specific sentiment lexicon. Expert Syst Appl 116:285–298

42. Yang XP, Zhang ZX, Wang L et al (2017) Automatic construction and optimization of sentiment lexicon based on Word2Vec. Comput Sci 74(1):42–47
43. Zabha NI, Ayop Z, Anawar S, Hamid E, Abidin ZZ (2019) Developing cross-lingual sentiment analysis of Malay twitter data using lexicon-based approach. Int J Adv Comput Sci Appl 10(1):346–351
44. Zhao CJ, Wang SG, Li DY (2019) Exploiting social and local contexts propagation for inducing Chinese microblog-specific sentiment lexicons. Comput Speech Lang 55:57–81

Yanyan Wang received a Bachelor's degree in engineering from the Communication University of China in 2016, then decided to study text sentiment analysis during her Master. She is currently working toward the Ph.D. degree at College of Information and Communication Engineering, Communication University of China, Beijing, China. Her current research interests include representation learning and sentiment analysis.

Fulian Yin received Bachelor, Master and PhD. degrees from Harbin engineering university in 2005, 2007 and 2010. She is currently an Associate Professor and Master Tutor in the College of Information and Communication Engineering, Communication University of China. Her current research interests natural language processing, data analysis and data mining.

Jianbo Liu received the Bachelor's degree in Department of Radio Electronics in Tsinghua University, Beijing, China, in 1985 and then received Master's degree in Communication University of China, Beijing, China, in 1988. He is a Professor at the College of Information and Communication Engineering, Communication University of China. His research interests include Cable TV and broadband network technology.

Marco Tosato completed his Bachelor of Science in pure mathematics (algebraic field) and then decided to switch to Applied (Biological) Mathematics during his master. Both the bachelor and the master were completed at the University of Trento, Italy with a short parenthesis as an exchange student at Utrecht University in the Netherlands. He is currently a phD student under the supervision of Professor Jianhong Wu at York University, Toronto, Canada. His research interests include Neural Networks,Delay Differential Equations.