IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2020, accepted March 25, 2020, date of publication March 30, 2020, date of current version April 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984284

The Construction of Sentiment Lexicon Based on Context-Dependent Part-of-Speech Chunks for Semantic Disambiguation

FULIAN YIN[1], YANYAN WANG[1], JIANBO LIU[1], AND LISHA LIN[2]
[1]Information Engineering Institute, Communication University of China, Beijing 100024, China
[2]College of Mathematics, Hunan University, Changsha 410082, China

Corresponding author: Yanyan Wang (yywang@cuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61801440, in part by the High-Quality and Cutting-Edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China), in part by the State Key Laboratory of Media Convergence and Communication (Communication University of China), and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT Sentiment lexicon, which provides sentiment information for words, plays an important role in sentiment analysis task. Currently, most of sentiment lexicons have only one sentiment polarity for each word and ignore sentimental ambiguity. In this paper, we propose to construct the sentiment lexicon based on context-dependent part-of-speech (POS) chunks, namely CP-chunks, which aims at solving the ambiguity of lexical sentiments. Given that the POS of context has impact on the word polarity and intensity, we take CP-chunks as an unit to do sentiment calculation. Our method is evaluated through the classification task of text sentiment. The experiment results indicate that, in comparison with the existing methods, the applicability of our method is more stable and balanced for both the positive and negative polarities corpora, and the accuracy of our method reaches 82% for the sentiment classification of a domain-specific corpus.

INDEX TERMS Part-of-speech, ambiguity, sentiment lexicon, sentiment analysis.

I. INTRODUCTION

With the fast development of computer technology, it has become a popular trend to express opinions containing different sentiment information through the network platform [1]–[3]. This sentiment information has great commercial value and relevant influence on public opinion or merchandise sales. Many techniques of text sentiment analysis, such as sentiment lexicons [4]–[8], word representation [9], [10], deep learning [3], [11], [12] and capsule networks [13], [14] have been studied to explore the values enclosed in the comments.

The constructed sentiment lexicon has been widely applied in sentiment analysis, especially in performing sentiment classification tasks in the internet industry. In 2013, Mohammad *et al.* [15] transformed the sentiment lexicon construction problem into the classification of word 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 [16] used a simple neural network architecture to reassign word sentiments, which outperformed the NRC [15]. In 2014, Tang *et al.* [17] 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 required to manually select the seed set in advance. Kiritchenko *et al.* [18] 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. Cambria *et al.* [8] employed recurrent neural networks to construct SenticNet5 sentiment lexicon, which provided the sentiment polarities of 100,000 commonsense concepts. Deng *et al.* [19] proposed a model of hierarchical supervision to construct a topic-adaptive sentiment lexicon, taking into account topics and sentimental words. The above methods only distinguish different sentiments of words according to different corpora. Zhao *et al.* [20] proposed to construct domain-specific sentiment lexicons, which considered user

The associate editor coordinating the review of this manuscript and approving it for publication was Shangce Gao.

relationship and topic characteristics etc. Despite the fact that many specific sentiment lexicons have been provided, there still have some limitations in the existing researches, for example, the sentimental ambiguity is usually neglected. Therefore, our work mainly focuses on the disambiguation of lexical sentiments in specific fields.

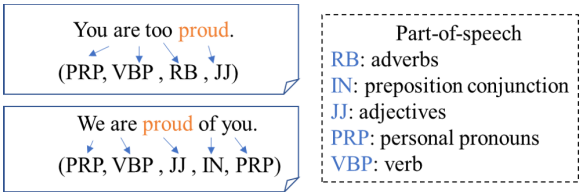

FIGURE 1. An illustrative example of two reviews that all include the word "proud". The two sentences have labeled each word with the part of speech.

To the best of our knowledge, most researchers only consider the positive and negative polarities of words and ignore the fact that the same word with the same POS may have different sentiments and meanings in different contexts. For example, the word "proud" in two different contexts has the same POS, which is shown in Fig. 1. If we judge the sentiment tendency of this word by using the traditional method, it would be difficult to identify the sentiment difference between this two reviews. As illustrated in the two sentences, we observe that (1) the surrounding words of the "proud" are different, (2) the POS information of these surrounding words is also different and (3) the varied POS of these surrounding words can be used to determine sentiments. In particular, if we consider the context of the current word to construct lexicon, it will produce massive results and lead to information redundancy. Therefore, we take the POS of surrounding words into account.

Inspired by the idea that POS can capture semantic information to enrich the sentiment of words, we propose a construction method of the specific sentiment lexicon, which considers the POS obtained from surrounding words as an unit to distinguish each word with the same POS but different concepts. It means that we decide the sentiment of a word by taking into account the POS of its previous words and following words. The new unit CP-chunk is built by combing the word with the POS of its surrounding words. Then several CP-chunks for each sentimental word are composed by the different POS of context words, and we assign different sentimental values to these CP-chunks using CHI algorithm. Finally, through the text sentiment classification experiments, our model achieves the state-of-the-art results in sentiment classification tasks compared to the existing lexicons, thus showing its great potential in studying sentimental disambiguation.

II. RELATED WORKS

A. RESOURCES OF EXISTING SENTIMENT LEXICONS

Sentimental lexicons play an important role in the task of sentiment analysis. SentiWordNet is the most famous affective lexicon constructed on the basis of WordNet in English language [6], [21]. In this lexicon, three scores were assigned for each synset of WordNet to describe the positive, negative and objective degree, which were further used to determine the precise sentiment of each word. However, most of the general lexicons were labeled manually, which not only were time-consuming, but also required a lot of manpower. General Inquirer (GI) [22] is broadly recognized as the first English lexicon, which collected 1914 positive words and 2293 negative words in the early stage, and labeled each word with polarity, intensity and lexical characteristics. About four decades later, Wilson *et al.* [23] proposed the MPQA sentimental lexicon, containing 2718 positive words and 4912 negative words. Each word was scored according to its sentimental strength from Multi-perspective Question Answering (MPQA) Opinion Corpus. SenticNet [8] is a popular sentiment lexicon for aspect-level sentiment analysis tasks [24]–[26]. It provided a set of 100,000 natural language concepts that were labeled related semantic, sentimental and polar associations. Mohammad *et al.* [15] built two large-scale sentiment lexicons for Twitter corpus. Yang *et al.* [27] constructed a 10-dimensional sentimental lexicon and used it for sentimental computing and text mining. Wu *et al.* [7] built a specific domain sentimental lexicon that contained opinion targets and sentimental words.

In addition, there are some studies of the construction of language-specific sentimental lexicons, such as Urdu [28], Malay [29] and Slovene [30]. Note that the above mentioned sentimental lexicon resources cannot always recognize sentimental words in specific areas and cannot distinguish the diversity of lexical sentiments in different contexts.

B. CONSTRUCTION APPROACHES OF SENTIMENT LEXICONS

Recently, researchers pay more and more attention to the construction of sentimental lexicon used for sentiment analysis. By using the existing sentiment lexicon as prior knowledge, some researchers constructed new sentiment lexicons with higher coverage in specific domains. In 2016, Liu *et al.* [31] integrated HowNet and NTUSD (released by University of Taiwan) to construct sentiment lexicon based on microblog. About one year later, Kimura and Katsurai [32] introduced emoji into the construction of sentiment lexicon and assigned a vector representation to each emoji by calculating the co-occurrence between an emoji and each sentimental word. The limitation of above methods is that they cannot judge the sentiment of words that do not present in sentiment lexicons.

Because of the richness of text corpora, more and more studies focus on mining new words from corpora and the construction of sentiment lexicons. Tang *et al.* [17] regarded the construction of sentiment lexicon as a word-level sentimental classification problem, and their work appeared to have a better performance than the existing Opinion Lexicon and MPQA lexicon. Huang *et al.* [33] adopted an automatic construction strategy to build a domain-specific sentiment lexicon based on the propagation of constrained labels, which used block-dependent information and existing lexicons to

extract candidate sentimental words. It showed that this method improved the performance of domain-specific sentiment lexicon significantly. After two years, Yang *et al.* [27] proposed the automatic construction method using coordinate shift. They trained the large-scale corpora through neural networks and constructed an unified form of evaluation function to study the effects of multiple constraints. Wu *et al.* [34] proposed an unified framework, which considered the lexical POS, formal and informal sentimental words and incorporated the connection between words and emotional symbols using sentimental similarity. Jin *et al.* [35] combined the existing sentiment lexicon resources with the user sentiment lexicon and constructed sentiment lexicon by rule-based fusion method applied on Twitter corpus. Currently, Zhao *et al.* [36] used the context propagation framework of sentiment unit and extracted the explicit and implicit sentiment features from Chinese microblog to construct lexicon. Its performance outperformed all the state-of-the-art baselines on sentiment classification task. These methods not only considered the information of existing sentiment lexicons, but also found some new words in specific domain corpora. However, they ignored the quality of corpora and ambiguity of words. In addition, a lot of methods manually selected positive and negative sentimental seed set which failed to consider the overall rationality of seed sets.

Some attempts have been made to prevent the sentimental ambiguity problem in the construction of sentiment lexicon. In 2017, Saif *et al.* [37] used context and semantic information extracted from specific domains to update the sentiment tendency of words and to alleviate the difference in lexical sentiment when context changes. Later, Han *et al.* [38] 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. In 2019, Wu *et al.* [7] 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 [7] is limited for reasons that it ignored some sentences are incomplete and have grammatical structure problems. Deng *et al.* [19] proposed a topic-adaptive sentiment lexicon (TaSL) for higher-level classification tasks, which jointly considered 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.

In this paper, we build the sentiment lexicon based on the CP-chunks algorithm, which takes the surrounding POS of words to construct CP-chunks and ensures the comprehensiveness of the lexicon. After that, we use the CHI algorithm to calculate the sentimental values of CP-chunks, and it avoids the difficulty of manually obtaining high-quality corpus and seed sets.

III. THE CONSTRUCTION OF SENTIMENT LEXICON BASED ON CP-CHUNKS

Most existing methods judge the sentimental polarity of the sentimental word only from the information of the vocabulary itself, such as part-of-speech or topic labels. SentiWordNet [6], [21] and MPQA [23] sentiment lexicons both considered the sentiment values of words in different parts-of-speech. Deng *et al.* [19] jointly considered the topics and sentiments of words to capture different sentiment expressions of a word under different topics. Different from the above-mentioned methods, we take the POS of the context of the word into account and determine the sentiment value and the polarity of the word in this paper. Also, we use the CP-chunks as an unit to reduce the sentimental ambiguity of words, and then we construct the sentiment lexicon by calculating the sentimental value of the chunks.

Fig.2 shows the sentiment lexicon construction process based on CP-chunks. This process involves preprocessing corpus, selecting sentimental words set, filtering candidate CP-chunks sets, calculating the sentiment of CP-chunks and evaluating method of the sentiment lexicon. We will introduce each step in the following sections.

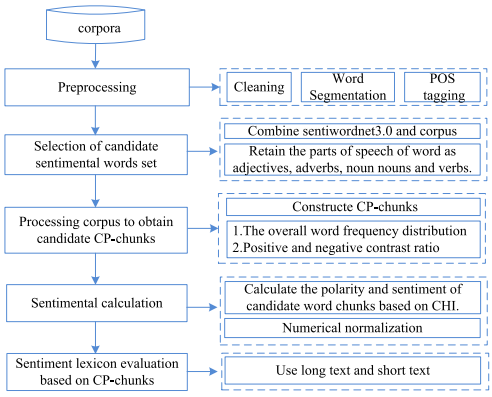

FIGURE 2. The overall framework for construction of sentiment lexicon based on CP-chunks.

A. BUILDING THE CONTEXT-DEPENDENT POS CHUNKS

First, we do word segmentation and POS tagging for each review. Then we incorporate the results with the candidate sentimental words set denoted by SW_{set} to construct the CP-chunks.

For each sentimental word sw_i, let $FP = (fp_1, \ldots, fp_i, \ldots, fp_N)$ be the POS to the corresponding N words $FW = (fw_1, \ldots, fw_i, \ldots, fw_N)$ before sw_i, let $BP = (bp_1, \ldots, bp_i \ldots, bp_M)$ be the POS to the corresponding M words $BW = (bw_1, \ldots, bw_i \ldots, bw_M)$ after sw_i. We build the CP-chunks for sw_i by combining any number between 0 and N of previous words' POS and any number between

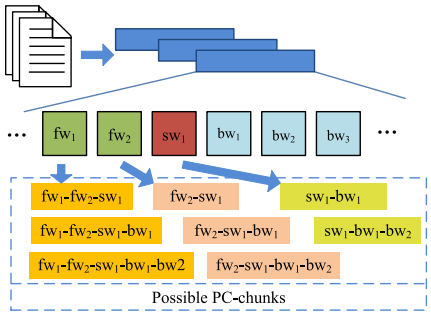

FIGURE 3. The process of the CP-chunks construction.

TABLE 1. Algorithm 1. Constraint screening algorithm for the selection of CP-chunks.

Inputs: corpora, $R(sw_i)$, T_1,T_2, $SW_{set} = \{sw_1, sw_2, ..., sw_i, ...\}$,
N:number of previous words, M: number of following words,
NLTK:part of speech tagging tools
Outputs:CP-chunks set
Initialize CP-chunks set as an empty set
For each corpus in corpora:
  Use NLTK to label the POS of each word.
  According to the size of M and N,
    obtain all CP-chunks of each corpus.
  For each CP-chunk in all CP-chunks:
    If sentimental word sw_i in the CP-chunks is in SW_{set},
    and the POS in the CP-chunks is adjectives,
    adverbs, nouns or verbs, the CP-chunk is reserved;
    Otherwise, delete this CP-chunk;
Then integrate and obtain CP-chunks of all sentimental words;
For each sentimental word sw_i in SW_{set}:
  Extract all CP-chunks of sw_i and count the frequency;
  For each CP-chunk in all CP-chunks:
    Filter by progressive conditions:
    (1)**Self-singleness**: the frequency of the CP-chunk is larger than T_1.
    (2)**Positive and negative contrast ratio of CP-chunk**:$R(sw_i) > T_2$.
  If both conditions are established:
    Add CP-chunk to CP-chunks set.
  Else:continue.
End.

0 and M of the following words' POS according to the words order in comment text. In this way, each sentimental word sw_i may have $(M + 1) * (N + 1) - 1$ types of CP-chunks. The core idea for constructing CP-chunks is shown in Fig. 3, and N and M are both taken 2 as an example. It can be seen that there are 8 cases for each sentimental word in each comment text. Since we often need to analyze thousands of corpora and each corpus contains some words, it is necessary to filter the candidate CP-chunks with the purpose of reducing computational cost.

Some CP-chunks may exist simultaneously in both positive and negative text corpora and do not have actual meanings or sentimental tendencies, such as ''book'',''table''. In order to reduce the impact of these words, we mostly take into account three factors, including POS, word frequency and distribution of words in positive and negative corpora. Our proposed constraint screening algorithm shown in Table 1 is used to select available CP-chunks.

B. CALCULATING THE SENTIMENT VALUE OF CP-CHUNKS

1) SELECTING THE SENTIMENTAL WORDS SET

Many words have no practical significance in the real review corpora, such as ''the'' and ''a'' etc., but they might affect our judgement on the polarity of sentiment words when we treat them as the previous words or following words of the CP-chunks. Therefore, it is necessary to select suitable candidate sentimental words following certain criteria. (i) If the frequency of a word in corpus is over T_1, then we add it to the candidate words set; (ii) We insert the vocabulary from the SentiWordNet3.0 lexicon [21] to the candidate words set. (iii) To select the suitable sentimental words set, for the sentimental word sw_i, we define its positive and negative contrast ratio as $R(sw_i)$:

$$R(sw_i) = \frac{|f(sw_i, x_p) - f(sw_i, x_n)|}{\sum_j f(sw_i, x_j)} > T_2, \tag{1}$$

where x_j is the corresponding sentimental polarity and has two states: the positive polarity x_p and the negative polarity x_n,$f(sw_i, x_p)$ is the frequency of the sw_i in positive corpus, $f(sw_i, x_n)$ is the frequency of the sw_i in negative corpus, $\sum_j f(sw_i, x_j)$ is the frequency of the sw_i in both positive and negative corpora, T_2 is the threshold to select the sentimental words set.

2) COMPUTING THE SENTIMENT OF CP-CHUNKS FOR EACH WORD

In this section, we introduce chi-square theory [39] to calculate the polarity and intensity of each CP-chunk. Different from traditional theory method of using a word as an unit, our method use CP-chunks in (2):

$$C(c_i, x_j) = \frac{N \times (AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}, \tag{2}$$

where c_i is i-th CP-chunk in sentimental words set, $C(c_i, x_j)$ is the sentimental value of c_i in x_j corpora, A is the number of corpora that belongs to x_j and contains c_i, B is the number of corpora that does not belong to x_j and contains c_i, C is the number of corpora that belongs to x_j and does not contain c_i, D is the number of corpora that does not belong to x_j and does not contain c_i, N is the total number of corpora. Then we calculate the strength and polarity of each word's sentiment using (3).

$$C(c_i) = C(c_i, x_p) - C(c_i, x_n), \tag{3}$$

where $C(c_i, x_p)$ is the sentimental value of c_i in the positive corpora, $C(c_i, x_n)$ is the sentimental value of c_i in the negative corpora and $C(c_i)$ is the final sentimental value of c_i. The sign of $C(c_i)$ determines the polarity of the c_i and its absolute value defines the chunk's sentimental strength.

C. 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

$x = (c_1, \ldots\ldots, c_k, \ldots\ldots, c_K)$, where c_k indicates that the k-th CP-chunk in sequence x, K is the total number of CP-chunks in the sequence x. Then we obtain the corresponding sentimental values $Sen_x = (S(c_1), \ldots\ldots, S(c_k), \ldots\ldots, S(c_K))$ for each c_k in x to synthesize the final sentence sentiment value S_x:

$$S_x = \sum_{k}^{K} S(c_k), \tag{4}$$

where $S(c_k)$ is the sentiment value of the k-th CP-chunk in sequence x 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.

$$T_x = \begin{cases} 1, & \text{if } S_x > 0 \\ 0, & \text{if } S_x = 0 \\ -1, & \text{if } S_x < 0, \end{cases} \tag{5}$$

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.

$$P(pos) = \frac{TP}{PN}, \tag{6}$$

$$R(pos) = \frac{TP}{N_{pos}}, \tag{7}$$

$$F1(pos) = \frac{2 \times P(pos) \times R(pos)}{P(pos) + R(pos)}, \tag{8}$$

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. Acc and Cov are two adopted measures of the accuracy and coverage of our method, which are defined as following:

$$Acc = \frac{TP + TN}{N}, \tag{9}$$

$$Cov = \frac{NP + NN}{N}. \tag{10}$$

where TN is the number of correct negative sentences using our proposed lexicon, N is total number of evaluation datasets.

IV. EXPERIMENTS AND ANALYSIS

In this section, we apply the proposed model to classify sentimental polarity based on film review data and compare its performance with the existing methods.

A. EXPERIMENTAL DATA

Our experimental data are real movie review data, including two datasets: one is the long text review dataset (LMRD) provided by Maas *et al.* [40], which contains 25,000 positive labeled samples, 25,000 negative labeled samples and 50,000 unlabeled samples; Another one (MRD), provided by Pang and Lee [41], is short text labeled samples and includes 5331 positive and same amount of 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 use 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 2 reports some detailed features of the datasets.

TABLE 2. Experimental datasets.

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

B. EXPERIMENTS ON THE CONSTRUCTION OF SENTIMENT LEXICON BASED ON CP-CHUNKS

As illustrated in previous sections, we construct the CP-chunks by taking into account the previous N and following M words around the current word. For simplicity, we only take $N = 1$, $M = 1$, $T_1 = 1$, and $T_2 = 0.1$. Then, the sentiment values of the CP-chunks are calculated using chi-square algorithm as shown in (2) and (3), to construct the original CP-chunks sentiment lexicon, namely OCP-Lex.

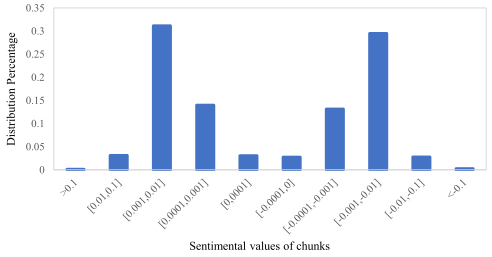

FIGURE 4. Distribution percentage of sentimental values of chunks in OCP-Lex.

Fig. 4 plots the distribution of chunks' sentimental values. We find that this distribution is uneven, and sentimental values of most chunks are distributed in the interval -0.01 to 0.01. To better reflect the influence of the sentimental value of each chunk on corpora, the sentimental values in the OCP-Lex are segmentally normalized and adjusted to get the fine-tuning CP-chunks sentiment lexicon, namely

FCP-Lex, which requires the determination of number of segments in advance. Fig. 4 shows that the distribution of positive and negative chunks in OCP-Lex is almost symmetrical, and hence it is sufficient to take positive values as an example to calculate the fine-tuning sentiment values of chunks.

For each segment $i \in [0, 1, 2, 3, 4, 5]$, the maximum and minimum of boundary are calculated in the following way:

$$v^i_{max} = (0.1)^i, \quad i = 0, 1, 2, 3, 4, 5 \tag{11}$$

and

$$v^i_{min} = (0.1)^{i+1}, \quad i = 0, 1, 2, 3, 4, 5 \tag{12}$$

The sentimental absolute value $\hat{v}$ of each chunk by fine tuning is calculated by (13):

$$\hat{v} = (0.5 - \frac{i}{10}) + \frac{F \times (|v| - v^i_{min})}{v^i_{max} - v^i_{min}}, \tag{13}$$

$$F = \begin{cases} 0.5, & \text{if } i = 0 \\ 0.1, & \text{if } i = 1, 2, 3, 4, 5, \end{cases} \tag{14}$$

where v is the sentimental value in OCP-Lex, which is obtained using (15):

$$\hat{v} = \begin{cases} \hat{v}, & \text{if } v > 0 \\ -\hat{v}, & \text{if } v < 0. \end{cases} \tag{15}$$

TABLE 3. Some CP-chunks examples in FCP-Lex.

No	word	CP-chunks	Sentiment value in OCP-Lex	Sentiment value FCP-Lex
1	great	great+N	0.2847	0.6026
2	great	N+great+N	0.0282	0.4203
3	great	V+great+N	0.0280	0.4200
4	well	R+well	0.0246	0.4162
5	excellent	excellent+N	0.0096	0.3953
6	young	young+N	0.0869	0.4855
7	film	film+IN	0.1500	0.5278
8	film	IN+film	-0.0020	-0.3110
9	bad	bad+N	-0.4356	-0.6864
10	bad	R+bad	-0.2966	-0.6092

From Table 3, we known there are some CP-chunks with undefined POS. We explain these POS below. The POS is tagged using python's natural language processing package NLTK. The arising is that some frequently used POS such as nouns (N), adverbs (R), adjectives (A) and verbs(V) have various tense forms, thus producing a large number of chunks and leading to a heavy computational burden. In order to solve this problem,we use the same tag to represent different tense forms for each frequent POS. The remaining POS are divided according to the results of NLTK, where IN expresses preposition or subordinate conjunction etc.

In order to test the quality of our constructed lexicon, we conduct a qualitative analysis via some case studies reported in Table 3, which provides the sentiment values of some CP-chunks under OCP-Lex and FCP-Lex. The positive (negative) sentiment values indicate a positive (negative) tendencies. From Table 3, we observe that:

(i) Our method can assign different sentiment strength to the same word under different POS of context, such as the word "great", "bad". This is helpful to solve the problem of the ambiguity of lexical sentiments and to obtain the fine-grained sentimental values.

(ii) Lines 7 and 8 show that the word "film" expresses two different sentimental polarities and intensities in different POS of contexts. The word "film" usually does not have an obvious sentimental tendency, but it has a certain emotional tendency in some specific contexts. Therefore, our method also achieves sentimental discrimination of some neutral words to a certain extent.

(iii) By comparing the original sentiment value in OCP-Lex and fine-tuned sentiment value in FCP-Lex, it can be seen that our numerical processing method can decrease the negative impact of some words on the result caused by the minimum sentimental intensity.

TABLE 4. Comparison of Effects of the OCP-Lex and the FCP-Lex.

Dataset	Methods	P(pos)	R(pos)	F1(pos)	P(neg)	R(neg)	F1(neg)	Acc
MRD	OCP-Lex	0.6076	0.8452	0.7062	0.7483	0.4431	0.5530	0.6441
	FCP-Lex	**0.6881**	0.7977	**0.7385**	**0.7607**	**0.6290**	**0.6879**	**0.7134**
LMRD	OCP-Lex	0.7925	0.7255	0.7572	0.7470	0.8093	0.7767	0.7674
	FCP-Lex	**0.7969**	**0.8633**	**0.8284**	**0.8511**	0.7787	**0.8128**	**0.8210**

Table 4 compares the efficiency of the OCP-Lex and the FCP-Lex in performing the text sentiment classification task. We find that our FCP-Lex has a better performance than the OCP-Lex in both datasets. Especially from the indicators of the F1, P, and Acc, the advantage is obvious in both the positive and negative tendency corpora. In addition, in the short text corpus MRD, the OCP-Lex only works well for positive corpus, and performs poorly for negative corpus. However, our FCP-Lex is efficient for both positive and negative corpus, and the highest accuracy reaches 82% for long text corpus LMRD.

C. COMPARATIVE EXPERIMENTS

In this section, we compare our proposed FCP-Lex with some existing sentiment lexicons, including GI [22], MPQA [23], SW [21], NRC [15], S140 [15], ETSL [18], HIT [17], NN [16], and HSSWE [11], in terms of performing the sentiment classification task.

Table 5 and Table 6 respectively give the performance of the text sentiment classification tasks conducted by 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 9% and 13% over the optimal results of the existing sentiment lexicons in the short and long text corpora, respectively.

For short text corpora, Table 5 shows that our method has a similar coverage to the other methods. However, our proposed FCP-Lex requires less corpus than the methods of NN, NRC and S140. In addition, the sentiment lexicons S140 and NN are only efficient for certain sentiment corpus,

TABLE 5. Effects of text sentiment classification task using different features based on short 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 FCP-Lex	0.9939	**0.6881**	0.7977	**0.7385**	**0.7607**	0.6290	**0.6879**	**0.7134**

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 FCP-Lex	**1.0000**	0.7969	0.8633	**0.8284**	**0.8511**	0.7787	**0.8128**	**0.8210**

and the sentiment lexicons our FCP-lex, MPQA and SW have the potential for both the positive and negative corpora. 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%. We find again that the use of the POS information improves the applicability of sentiment lexicon.

In summary, our FCP-Lex proves to be more effective than other sentiment lexicons since it takes the POS of the context into account. Moreover, due to similar performance for both positive and negative corpora, it also shows the applicability and stability of our model for both positive and negative corpora.

D. EXPLORING FACTOR EXPERIMENTS

We explore the optimal values of the word frequency selection factor T_1 and the positive and negative contrast ratio T_2, respectively. In experiments, we use the short text dataset MRD to conduct 10-fold cross-validation to improve the robustness of our constructed lexicon in performing sentiment classification tasks.

1) WORD FREQUENCY SELECTION

In order to determine the word frequency of sentiment words in the construction of our sentiment lexicon, we select MRD corpora for experiments to explore the performance of our lexicon under different word frequency choices. For each value of T_1, we first divide a total of 10,000 positive and negative corpora into 10 parts, taking 9/10 corpora of them to build a CP-chunk-based sentiment lexicon, and then use 1/10 corpora for text sentiment classification. Fig. 5 shows

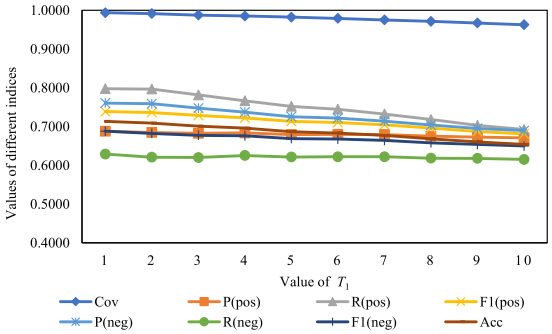

FIGURE 5. The indicators' performance of text sentiment classification under different T_1.

the impact of different word frequencies on text sentiment classification tasks. We find that the all indicators are declining with the increase of the T_1. And the running time has little difference because the amount of data used in this paper is small. Hence, the experimental results can be mainly considered to determine optimal T_1, and we take the value 1 for constructing lexicon.

2) THE POSITIVE AND NEGATIVE CONTRAST RATIO

In order to determine the value of the positive and negative contrast ratio in the construction of our sentiment lexicon, we select MRD corpora for experiments to explore the performance under T_2. Fig. 6 shows the effect of sentiment classification under different T_2. It can be seen that the vocabulary coverage is stable under different T_2. When the value of T_2 is between 0.1-0.6, all indicators have a similar effect, but the effect is very bad when the value of T_2 is greater than 0.6. Therefore, we set the optimal value of T_2 between 0.1-0.6, and we take the value 0.1 when constructing our FCP-Lex.

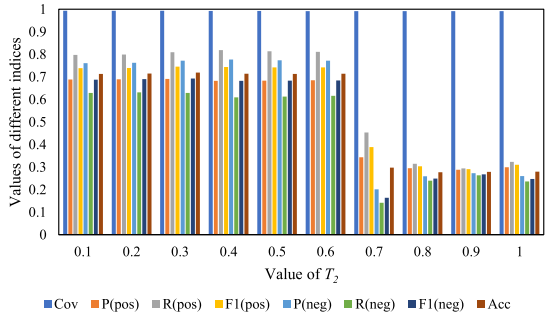

FIGURE 6. The indicators' performance of text sentiment classification under different T_2.

V. CONCLUSION AND DISCUSSION

In this paper, we propose an automatic construction method of the sentimental lexicon, named FCP-Lex, based on CP-chunks. Our method can not only reduce the ambiguity of words, but also avoid the difficulty in obtaining high-quality corpora and seed sets manually. The experiment results of text sentiment classification tasks show that our constructed

FCP-Lex is more effective in sentiment analysis than those existing sentiment lexicons, and it performs good for both positive and negative corpora with a high accuracy (over 80%). Because of its unique historical and linguistic environment, Chinese has the characteristics of implicit expression, and a word often presents many different meanings and sentiments in different contexts. For example, in the following two sentences, the word "骄傲"(pride) has two different meanings, which represent positive and negative sentiment tendencies, respectively.

1. 我们为我们取得的成绩骄傲. (We are proud of our achievements.)

2. 她平时很骄傲, 不帮助他人. (She is usually too proud and doesn't help others.)

Unlike English corpus, the continuous composition of Chinese corpus words brings challenges to process Chinese natural language. In future studies, we would like to concentrate on the issues of Chinese natural language processing, including word segmentation, semantic disambiguation, word embedding, etc.

REFERENCES

[1] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl. Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.

[2] F. Iqbal, J. M. Hashmi, B. C. M. Fung, R. Batool, A. M. Khattak, S. Aleem, and P. C. K. Hung, "A hybrid framework for sentiment analysis using genetic algorithm based feature reduction," *IEEE Access*, vol. 7, pp. 14637–14652, 2019.

[3] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, and E. Cambria, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov. 2018.

[4] T. Wilson, P. Hoffmann, and S. Somasundaran, "OpinionFinder: A system for subjectivity analysis," in *Proc. EMNLP-HLT*, 2005, pp. 34–35.

[5] L. Xu, H. Lin, Y. Pan, H. Ren, and J. Chen, "Constructing the affective lexicon ontology," *J. China Soc. Sci. Tech. Inf.*, vol. 27, no. 2, pp. 180–185, 2008.

[6] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. LREC*, 2006, pp. 417–422.

[7] S. Wu, F. Wu, Y. Chang, C. Wu, and Y. Huang, "Automatic construction of target-specific sentiment lexicon," *Expert Syst. Appl.*, vol. 116, pp. 285–298, Feb. 2019.

[8] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Proc. 32th Int. Conf. Assoc. Adv. Artif. Intell.*, 2018, pp. 1795–1802.

[9] R. Othman, Y. Abdelsadek, K. Chelghoum, I. Kacem, and R. Faiz, "Improving sentiment analysis in Twitter using sentiment specific word embeddings," in *Proc. 10th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Systems: Technol. Appl. (IDAACS)*, Sep. 2019, pp. 854–858.

[10] X. Wang, J. Chen, A. Hawbani, F. Miao, and C. Shao, "Building sentiment lexicon with representation learning based on contrast and label of sentiment," in *Proc. 4th Int. Conf. Big Data Comput. Commun. (BIGCOM)*, Aug. 2018, pp. 151–156.

[11] Y. Wang, Y. Zhang, and B. Liu, "Sentiment lexicon expansion based on neural PU learning, double dictionary lookup, and polarity association," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 553–563.

[12] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, W. H. Chan, and E. Cambria, "Segment-level joint topic-sentiment model for online review analysis," *IEEE Intell. Syst.*, vol. 34, no. 1, pp. 43–50, Jan. 2019.

[13] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2011, pp. 44–51.

[14] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. NIPS*, 2017, pp. 3856–3866.

[15] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proc. Int. Workshop Semantic Eval.*, 2013, pp. 321–327.

[16] D. T. Vo and Y. Zhang, "Don't count, predict! An automatic approach to learning sentiment lexicons for short text," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 219–224.

[17] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building large-scale Twitter-specific sentiment lexicon: A representation learning approach," in *Proc. COLING*, 2014, pp. 172–182.

[18] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, Aug. 2014.

[19] D. Deng, L. Jing, J. Yu, S. Sun, and M. K. Ng, "Sentiment lexicon construction with hierarchical supervision topic model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, pp. 704–718, Apr. 2019.

[20] C. Zhao, S. Wang, and D. Li, "Exploiting social and local contexts propagation for inducing chinese microblog-specific sentiment lexicons," *Comput. Speech Lang.*, vol. 55, pp. 57–81, May 2019.

[21] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. LREC*, 2010, pp. 83–90.

[22] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, "The general inquirer: A computer approach to content analysis," *Amer. J. Sociol.*, vol. 73, no. 5, pp. 634–635, 1968.

[23] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. EMNLP-HLT*, 2005, pp. 347–354.

[24] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, "Merging SenticNet and WordNet-affect emotion lists for sentiment analysis," in *Proc. IEEE 11th Int. Conf. Signal Process.*, Oct. 2012, pp. 1251–1255.

[25] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 31–38, Mar. 2013.

[26] D. Vilares, H. Peng, R. Satapathy, and E. Cambria, "BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1292–1298.

[27] X. Yang, Z. Zhang, and Z. Zhang, "Automatic construction and global optimization of a multisentiment lexicon," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–8, Nov. 2016.

[28] N. Mukhtar, M. A. Khan, and N. Chiragh, "Lexicon-based approach outperforms supervised machine learning approach for urdu sentiment analysis in multiple domains," *Telematics Informat.*, vol. 35, no. 8, pp. 2173–2183, Dec. 2018.

[29] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Zainal, "Developing cross-lingual sentiment analysis of malay Twitter data using lexicon-based approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 346–351, 2019.

[30] J. Bučar, M. Žnidaršič, and J. Povh, "Annotated news corpora and a lexicon for sentiment analysis in slovene," *Lang. Resour. Eval.*, vol. 52, no. 3, pp. 895–919, Sep. 2018.

[31] J. Liu, M. Yan, and J. Luo, "Research on the construction of sentiment lexicon based on Chinese microblog," in *Proc. 8th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, Aug. 2016, pp. 56–59.

[32] M. Kimura and M. Katsurai, "Automatic construction of an emoji sentiment lexicon," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, 2017, pp. 1033–1036.

[33] S. Huang, Z. Niu, and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowl.-Based Syst.*, vol. 56, pp. 191–200, Jan. 2014.

[34] F. Wu, Y. Huang, Y. Song, and S. Liu, "Towards building a high-quality microblog-specific Chinese sentiment lexicon," *Decis. Support Syst.*, vol. 87, pp. 39–49, Jul. 2016.

[35] Z. Jin, Y. Yang, X. Bao, and B. Huang, "Combining user-based and global lexicon features for sentiment analysis in Twitter," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4525–4532.

[36] C. Zhao, S. Wang, and D. Li, "Exploiting social and local contexts propagation for inducing Chinese microblog-specific sentiment lexicons," *Comput. Speech Lang.*, vol. 55, pp. 57–81, May 2019.

[37] H. Saif, M. Fernandez, L. Kastler, and H. Alani, "Sentiment lexicon adaptation with context and semantics for the social Web," *Semantic Web*, vol. 8, no. 5, pp. 643–665, Apr. 2017.

[38] H. Han, J. Zhang, J. Yang, Y. Shen, and Y. Zhang, "Generate domain-specific sentiment lexicon for review sentiment analysis," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 21265–21280, Aug. 2018.
[39] N. L. Johnson, S. Kotz, and N. Balakrishnan, "Chi-square distributions including Chi and Rayleigh," in *Continuous Univariate Distributions*, 2nd ed. Hoboken, NJ, USA: Wiley, 1994, pp. 415–493.
[40] A. L. Maas, R. E. Daly, and P. T. Pham, "Learning word vectors for sentiment analysis," in *Proc. ACL-HLT*, 2011, pp. 142–150.
[41] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. ACL*, 2005, pp. 115–124.

FULIAN YIN received the bachelor's, master's, and Ph.D. degrees from Harbin Engineering University, in 2005, 2007, and 2010, respectively. She is currently a Professor and a Master Tutor with the College of Information and Communication Engineering, Communication University of China. Her current research interests include natural language processing, data analysis, and data mining.

YANYAN WANG was born in Huainan, Anhui, China. She received the bachelor's degree in engineering from the Communication University of China, Beijing, China, in 2012, where she is currently pursuing the Ph.D. degree in communication and information system. Her research interests include natural language processing, sentimental computing, and text representation.

JIANBO LIU received the bachelor's degree from the Department of Radio Electronics, Tsinghua University, Beijing, China, in 1985, and the master's degree from the Communication University of China, Beijing, in 1988. He is currently a Professor with the College of Information and Communication Engineering, Communication University of China. His research interests include cable TV and broadband network technology.

LISHA LIN was born in Baoji, Shannxi, China. She received the B.S. degree in applied mathematics from Hunan University, China, in 2014, where she is currently pursuing the Ph.D. degree in applied mathematics. Her research interests include option pricing, Bayesian statistical inference, and stochastic differential equation. She did some studies on applying Bayesian methods to the problem of option pricing and make inference on general stochastic differential equations.