IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 2, 2019, accepted March 17, 2019, date of publication March 27, 2019, date of current version April 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907772

Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary

GUIXIAN XU, ZIHENG YU, HAISHEN YAO, FAN LI, YUETING MENG, AND XU WU
College of Information Engineering, Minzu University of China, Beijing 100081, China

Corresponding author: Guixian Xu (guixian_xu@muc.edu.cn)

This work was supported by the Project of Humanities and Social Science, Ministry of Education of China, under Grant 18YJA740059.

ABSTRACT The method of text sentiment analysis based on sentiment dictionary often has the problems that the sentiment dictionary doesn't contain enough sentiment words or omits some field sentiment words. In addition, due to the existence of some polysemic sentiment words with positivity, negativity, and neutrality, the words' polarity cannot be accurately expressed, so the accuracy of text sentiment analysis is reduced to some extent. In this paper, an extended sentiment dictionary is constructed. The extended sentiment dictionary contains the basic sentiment words, the field sentiment words, and the polysemic sentiment words, which improves the accuracy of sentiment analysis. The naive Bayesian classifier is used to determine the field of the text in which the polysemic sentiment word is. Thus, the sentiment value of the polysemic sentiment word in the field is obtained. By utilizing the extended sentiment dictionary and the designed sentiment score rules, the sentiment of the text is achieved. The experimental results prove that the proposed sentiment analysis method based on extended sentiment dictionary has certain feasibility and accuracy. The research is meaningful for the sentiment recognition of the comment texts.

INDEX TERMS Chinese text sentiment analysis, text classification, naive Bayesian, sentiment dictionary.

I. INTRODUCTION

In recent years, with the continuous development of the Internet, more and more users have expressed their views on the Internet. Therefore, a lot of user comments are generated on the Internet. For example, product comments are generated on E-commerce websites such as Jingdong and Taobao, and hotel comments are generated on travel websites such as Ctrip and ELong. These comments convey the views of Internet users about products, hot events, etc. Merchants can master the user satisfaction with the relevant product comments. Potential users can evaluate products by viewing these product comments. With the rapid growth of the comments number, it is difficult to analyze the comments manually. Thus information technology is utilized to mine the sentiment tendency contained in the texts, and text sentiment analysis technology comes into being. Text sentiment analysis refers to the tendency mining of text sentiment by information technology. According to the granularity of the text, text sentiment analysis can be divided into three levels of analysis: word level, sentence level and chapter level. The word level sentiment analysis is to analyze the sentiment of the words [1]. The sentiment analysis of word is the prerequisite of sentiment analysis of sentence and chapter. At present, the researches on the analysis of sentiment word mainly include the extraction of sentiment word, the class of sentiment word and the construction of sentiment dictionary [2]–[4]. With the generation of a large number of texts on the Internet, the researchers have gradually focused on the sentence-level and chapter-level sentiment analysis. Sentence-level sentiment analysis is the core of text sentiment analysis. On the one hand, sentence sentiment analysis results are determined by the sentiment words in the sentences. On the other hand, the text consists of sentences with the relationship of juxtaposition, transition, composition. The results of sentence-level analysis determine the chapter-level sentiment tendency. Chapter-level sentiment analysis is a relatively complex task. It is a comprehensive consideration of the sentence sentiment analysis results. Meanwhile, the chapter-level sentiment tendency is based on the semantic relation of the context. In terms of the research of text sentiment analysis, the key to determine sentiment tendency is sentiment words. Therefore, it is very meaningful

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

to construct a complete and effective sentiment dictionary for the task of sentiment analysis.

At present, Hownet sentiment dictionary, National Taiwan University Sentiment Dictionary (NTUSD) are most commonly used in Chinese text sentiment analysis. However, the above-mentioned sentiment dictionaries have problems such as lacking of the field sentiment words and the polysemic sentiment words. In this paper, the field sentiment words and the polysemic sentiment words in the fields of hotel, digital, fruit, clothing and shampoo are manually collected. Extended sentiment dictionary containing a wider range of sentiment words are constructed, which improves the effect of sentiment analysis in the above fields.

In this paper, taking the sentiment words and the polysemic sentiment words in specific fields into account, Chinese text sentiment analysis method based on the extended sentiment dictionary is proposed. The remainder of this article consists of four parts. Firstly, the research background and current situation of text sentiment analysis method are expounded. Secondly, the detail of Chinese text sentiment analysis method based on the extended sentiment dictionary is described. Thirdly, the experiments are carried out and the experimental results are analyzed and discussed. Finally, the proposed method is summarized and the next research direction is introduced.

II. RELATED WORK

At present, there are mainly two research ideas on text sentiment analysis: (1) The research method based on machine learning. It regards text sentiment analysis as text classification. (2) The research method based on the sentiment dictionary. It determines the sentiment tendency of the text by calculating text sentiment score with rules.

A. SENTIMENT ANALYSIS BASED ON MACHINE LEARNING

The sentiment analysis based on machine learning is essentially the task of text categorization. A large number of annotated corpus are used for training to get a sentiment classifier. The sentiment classifier can judge text sentiment tendency.

Pang *et al.* [5] were the earlier researchers engaged in text sentiment analysis based on machine learning. They applied naive Bayes algorithm, maximum entropy algorithm and SVM algorithm to analyze the sentiments of the film reviews. The experimental results showed that SVM algorithm had the better performance in the sentiment analysis of movie reviews. Kiritchenko *et al.* [6] used a variety of semantic and sentiment features to conduct supervised statistical text sentiment classification. The sentiment features were extracted from the high-coverage sentiment dictionary which was automatically generated from sentiment tweets. Considering the characteristics of short texts which were sparse, nonstandard and ambiguous in subject, Wang *et al.* [7] proposed a high-dimensional feature model based on SVM. The proposed model was used to compare with Recursive Auto Encoder, Doc2vec and so on, which showed that it was more effective for short text emotion classification. Huang *et al.* [8] believed that microblogging sentiment was closely related to its topic, but most of the current microblogging sentiment analysis methods could not achieve cooperative analysis of topic and sentiment in microblogging. They proposed a TSMMF model (Topic Sentiment Model based on Multi-feature Fusion), which integrated emoticons emotions and the personality of microbloggers into LDA inference framework, and used Gibbs sampling techniques to estimate parameters in the model. Finally, synchronized detection of sentiment and topic in microblogging was achieved. In the sentiment classification algorithm based on machine learning, each article is transformed into a corresponding eigenvector, which would directly affect the performance of the sentiment analysis task. Literature research [9]–[11] regarded the sentiment classification as a feature optimization task.

Sentiment analysis based on machine learning solves the problem of sparseness of sentiment words and transforms the text into structured data. However, it treats each feature as an isolated element and neglects the intrinsic connection between features. The classification effect extremely depends on artificial annotation quality of the corpus.

B. SENTIMENT ANALYSIS BASED ON SENTIMENT DICTIONARY

Text sentiment analysis based on sentiment dictionary is an unsupervised classification method. Firstly, sentiment words, negative words, adverbs, conjunctions and so on in the sentence should be find out and the weight of sentiment words should be assigned. Then the sentiment weights of the words are further calculated depending on the degree adverbs and negative words. Considering the influence of the conjunctions and the weights of sentiment phrases in the sentence, the sentiment scores of the sentences are obtained. Finally, combined with the sentiment results of the sentences, the sentiment tendency of the text is determined.

Many researchers have studied the sentiment analysis based on sentiment dictionary. Turney and Littman, [12] proposed a method for inferring the sentiment orientation of a word. They used point mutual information (PMI) and latent semantic analysis (LSA) to calculate the semantic correlation between a word and the set of positive and negative paradigm words. Then the word was classified as positive or negative, based on average semantic association. Taboada *et al.* [13] calculated the sentiment scores of words, phrases, sentences and passages by detailed rules, and used the method of threshold setting to judge the sentiment tendencies. Saif *et al.* [14] presented SentiCircles, a lexicon-based approach for sentiment analysis on Twitter. Different from typical lexicon-based approaches, SentiCircles offered a fixed and static prior sentiment polarities of words. It took into account the co-occurrence patterns of words in different contexts in tweets to capture their semantics and updated the pre-assigned strength. Li *et al.* [15] proposed a sentiment analysis method based on bilingual sentiment lexicon for microblogs. Experiments showed that the proposed method had a good classification effect on sentiment analysis of

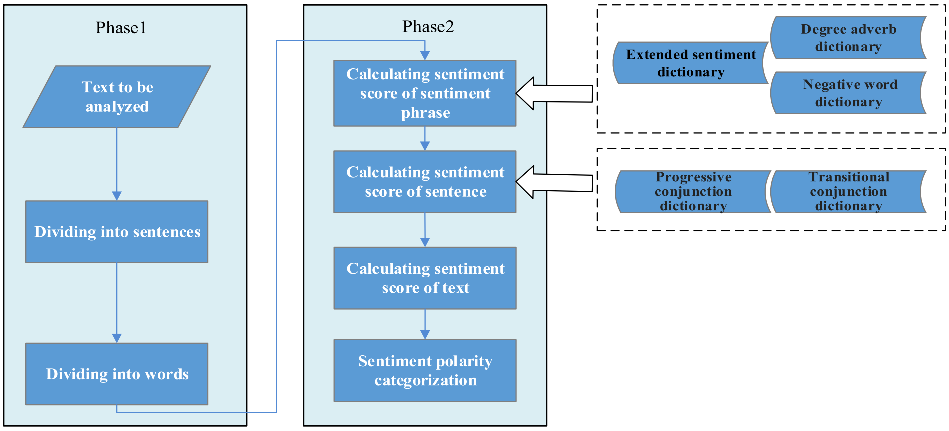

Phase1
Text to be analyzed
Dividing into sentences
Dividing into words
Phase2
Calculating sentiment score of sentiment phrase
Calculating sentiment score of sentence
Calculating sentiment score of text
Sentiment polarity categorization
Extended sentiment dictionary
Degree adverb dictionary
Negative word dictionary
Progressive conjunction dictionary
Transitional conjunction dictionary

FIGURE 1. The framework of text sentiment analysis method proposed in this paper.

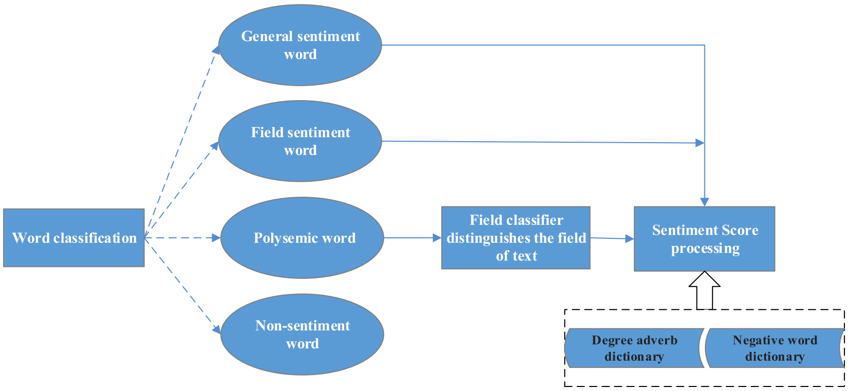


FIGURE 2. The score computing process of a sentiment phrase.

Chinese and English mixed microblogs. Wu *et al.* [16] constructed a sentiment dictionary suitable for shopping field. Nouns, adjectives, and some other parts of speech in the corpus were retained. By sorting the TF-IDF of words and setting threshold, the field sentiment words were obtained to construct the domain sentiment dictionary which had been applied to the sentiment classification of shopping reviews. Wu *et al.* [17] extracted features of financial texts based on Apriori algorithm. They constructed the financial sentiment dictionary and semantic rules to identify sentiment units and intensities. Then, the text sentiment tendency and intensity was obtained.

Sentiment analysis based on sentiment dictionary does not require manual labeling of samples and is easy to implement. However, the quality of sentiment analysis results depends on the sentiment dictionaries. At present, most of the Chinese sentiment dictionaries have not enough sentiment words and lack of field sentiment words.

In this paper, a text sentiment analysis method based on extended sentiment dictionary is proposed. The field sentiment words and the polysemic sentiment words in multiple fields are taken into account. To some extent, it extends the basic sentiment dictionary.

III. RESEARCH METHODS

The framework of text sentiment analysis method proposed in this paper is shown in Figure 1. The polysemic sentiment words which have multiple sentiment polarities are taken into account. The naive Bayes field classifier is used to distinguish the field of the text in which the polysemic sentiment word exists. The calculation process of a sentiment phrase's score is shown in Figure 2.

A. CONSTRUCTION OF DICTIONARIES

1) BASIC SENTIMENT DICTIONARY

Hownet sentiment dictionary and National Taiwan University Sentiment Dictionary (NTUSD) are the most commonly used dictionaries in Chinese text sentiment analysis. The sentiment words in the two dictionaries can be used in a wide range of areas of sentiment analysis. The words collected in these two dictionaries are relatively common sentiment words, such as "高兴(happy)", "伤心(sad)" and so on. In this paper, the initial sentiment dictionary is the union of Hownet sentiment dictionary and NTUSD. Meanwhile, 33 Internet popular sentiment words, such as "口耐(cute)", "给力(helpful)", "高富帅(perfect man)" and "鸭梨山大(high pressure)", were manually extracted from two Chinese popular social

TABLE 1. Examples of active words.

Weight	Examples of active words	Number
1	尊贵(honorable)、动听(enchanting)、诚恳(sincerely)、 栩栩如生(lifelike)、神圣(holy)、谦虚(humble)......	7192

TABLE 2. Examples of passive words.

Weight	Examples of passive words	Number
-1	无情无义(ruthless)、伤悲(sadly)、朝不保夕(precarious)、 焦虑(anxious)、怒斥(irritated)、指责(blamed).....	12079

TABLE 3. Examples of the field sentiment words.

Field Example	hotel	digital	fruit	clothing	shampoo
合身 (The size of the clothes is right)				+	
海边(Seaside)	+				
头屑(Dandruff)					-
黑屏(Black screen)		-			
水润(Moisturizing skin)			+		+

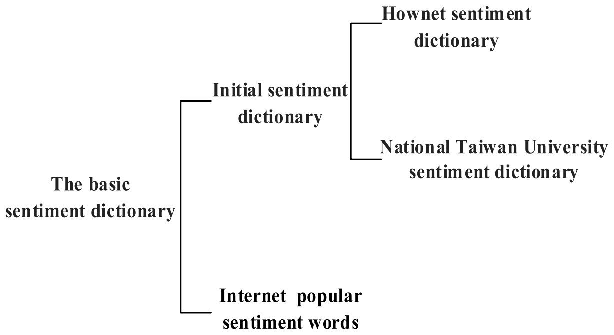

Hownet sentiment dictionary
Initial sentiment dictionary
National Taiwan University sentiment dictionary
The basic sentiment dictionary
Internet popular sentiment words

FIGURE 3. The composition of the basic dictionary.

network sites, Sina Weibo (https://www.weibo.com/) and Douban (https://www.douban.com/). The basic sentiment dictionary is made up of the initial sentiment dictionary and Internet popular sentiment words.

The composition of the basic dictionary is shown in Figure 3.

7191 active sentiment words and 12079 passive sentiment words were collected. The examples of active and passive sentiment words are shown in Table 1 and Table 2.

2) EXTENDED SENTIMENT DICTIONARY

Basic sentiment dictionary contains a large number of sentiment words, but omits some special words. Some sentiment words called the field sentiment words are usually used in one specific field and the sentiment tendency of the sentiment words is unique. For example, “死机((digital products) stop working)” and “掉线((digital products) dropped)” are mostly used in the text of the digital field. “定型(keep hair in shape)” and “干枯(dry hair)” are mostly used in the text of the shampoo field. Examples of the field sentiment words are shown in Table 3. “+” indicates active word and “−” indicates passive word. In addition, there are some sentiment words with different sentiment polarities in different fields. In this paper, such words are called the polysemic sentiment words. For example, “年代感(sense of period)” is often passive in text of the digital field, but is often active in the text of the hotel field. Examples of the polysemic sentiment words are shown in Table 4.

In this paper, the corpus of the sentiment analysis covers five fields: hotel, digital, fruit, clothing and shampoo. Based on the experience of the domain experts, 462 field sentiment words and 413 polysemic sentiment words were

TABLE 4. Examples of the polysemic sentiment words.

Example \ Field	hotel	digital	fruit	clothing	shampoo
老古董(antique)	+	-			
年代感(sense of period)	+	-			
热乎(warm)		-		+	
细腻(exquisite and small)		+	-		+

TABLE 5. Distribution of the field sentiment words and the polysemic sentiment words.

Field	The number of passive words	The number of active words	The number of field sentiment words	The number of the polysemic sentiment words
hotel	57	214	128	143
digital	70	182	106	146
fruit	34	146	94	86
clothing	31	255	158	128
shampoo	39	155	113	81

TABLE 6. Examples of degree adverbs.

Level	Weight	Examples of degree adverbs	Number
extreme	4	极度(extreme)、无比(Unparalleled)、最为(most)......	69
high	3	出头(too)、过于(too)、分外(extraordinary)......	71
medium	2	更(more)、很(very)、足足(very)......	37
low	0.5	略微(a little)、略为(a little)、稍微(a little)......	40

manually extracted from the comments of Jingdong (https://www.jd.com/) in five fields, as is shown in Table 5. Basic sentiment dictionary, the field sentiment words and the polysemic sentiment words form the extended sentiment dictionary.

3) DEGREE ADVERB DICTIONARY

Degree adverbs are a subcategory of adverbs, and they are often on the left of the sentiment word. They don't change the polarity of the sentiment word, but have a strengthening or weakening modification effect on the sentiment score. The degree adverbs are divided into four levels [18]: extreme, high, medium, and low, with weights of 4, 3, 2 and 0.5 respectively, as is shown in Table 6.

4) NEGATIVE WORD DICTIONARY

Negative words are usually on the left side of sentiment words. They reverse the sentiment tendency of a sentiment word. An active sentiment word and a negative word construct a negative sentiment phrase. On the contrary, a passive

TABLE 7. Examples of negative words.

Weights	Examples of negative words	Number
-1	没有(not)、没(not)、不是(not)、无(not)......	21

TABLE 8. Examples of conjunctions.

Type	Weights	Examples of conjunctions	Number
transitional conjunctions	1.5	但是(but)、但(but)、然而(however)、可是(but)......	10
progressive conjunctions	2	甚至(even)、况且(moreover)、尤其(especially)......	9

sentiment word and a negative word construct an active sentiment phrase. For example, "高兴(happy)" and "不高兴(not happy)", "丑(ugly)" and "不丑(not ugly)". In addition, the number of negative words on the left side of a sentiment word is also an important indicator. If the number is an even number, the sentiment tendency of the sentiment word does not change. In this paper, a negative word dictionary is constructed and the weights of negative words are set to -1, as is shown in Table 7.

5) CONJUNCTION DICTIONARY

There are many kinds of conjunction, such as progressive conjunctions, transitional conjunctions, causality conjunctions, and concession conjunctions. In this paper, the transition conjunctions and progressive conjunctions are studied. When there are progressive conjunctions such as "甚至(even)" and "尤 其(especially)" between the clauses, the sentimental polarity of clause after the progressive conjunction is the same as that before the progressive conjunction. Meanwhile, the sentiment score is obviously enhanced. For example, "这次买的水果都很好, 尤其是这个西瓜特别的甜(This time the fruits I bought are very good, especially the watermelon is particularly sweet)". When there are transitional conjunctions such as "但(but)" and "但是(but)" between the clauses, the sentiment polarity of the clause after the transitional conjunction is opposite to that before the transitional conjunction. The sentiment score of the clause after the transitional conjunction is enhanced. For example,"这个礼拜的天气都超级棒, 但是今天却特别糟糕(The weather this week is superb, but it's really bad today)". In this paper, a conjunctions dictionary containing transitional conjunctions and progressive conjunctions is constructed, with weights of 1.5 and 2, as is shown in Table 8.

B. NAIVE BAYESIAN FIELD CLASSIFIER

Due to the existence of the polysemic sentiment words, the accuracy of text sentiment analysis is reduced to an extent. In this paper, the polysemic sentiment words in clothing, fruit, shampoo, digital and hotel fields are collected. The field of a polysemic sentiment word is identified by naive Bayesian classification algorithm. Then, through the extended sentiment dictionary, the polarity of the polysemic sentiment word in the field is distinguished.

TABLE 9. Examples of stop words.

Examples of stop words	Number
的(of)、在(in)、了(the)、其(its)......	434

Naive Bayesian algorithm is a classification method based on Bayesian theorem. First, based on the independent assumption of features, the training set is trained, and the joint probability density of the input/output is learned. Then, based on this model, for a given input X, the Bayesian theorem is used to find the output with the maximum posterior probability Y [19]. The category for the text $X = \{x_1, x_2, \ldots, x_n\}$ is $Y = \{y_1, y_2, \ldots, y_n\}$, and y_i represents the text category label. The naive Bayes classifier can be expressed as formula (1).

$$y = \arg\max_{y_i} = \frac{P\{Y = y_i\} \prod_{j=1}^{n} P\{X^{(j)} = x_j | Y = y_i\}}{\sum_{i=1}^{k} P\{Y = y_i\} \prod_{j=1}^{n} P\{X^{(j)} = x_j | Y = y_i\}} \tag{1}$$

The first step in obtaining the classifier is to preprocess the dataset and de-duplicate the dataset to improve the reliability of the experiment. The jieba in Python is used to segment the Chinese text. After the text segmentation, the stop words should be removed according to the stop word list. The examples of the stop words are shown in Table 9.

In terms of the text representation, the vector space model for the document modeling is adopted. The value of each feature word adopts the Bool type weight method and the calculation method is as follows (2).

$$\text{Bool}: \begin{cases} 1 & freq(x_i, \mathrm{d}_j) > 0 \\ 0 & freq(x_i, \mathrm{d}_j) = 0 \end{cases} \tag{2}$$

Among them, $freq(x_i, d_j)$ represents the frequency that x_i appears in document d_j.

TABLE 10. Methods for calculating sentiment scores of sentiment phrases.

Serial number	Combination	Formula	Example	Score
1	S=SW	V(SW)	聪明(clever)	+1
2	S=NA+SW	$(-1)^{N}$ *V(SW)	不是不聪明(not not clever)	+1
3	S=DA+SW	V(DA)*V(SW)	极度聪明(extremely clever)	+4
4	S=DA+NA+SW	V(DA)* $(-1)^{N}$ *V(SW)	极度不聪明(extremely not clever)	-4
5	S=NA+DA+SW (NA is an even number)	V(DA)*V(SW)	不是不极度聪明 (not not extremely clever)	+4
6	S=NA+DA+SW (NA is an odd number)	0.5*V(DA)*V(SW)	不是极度聪明 (not extremely clever)	+2
7	S=SW+!	V(SW)*2	聪明!(clever!)	+2

After the text representation, the classifier is trained by the naive Bayes algorithm. The trained classifier is used to identify the text field.

C. CALCULATION OF THE TEXT SENTIMENT SCORE

The calculation process of the text sentiment score can be decomposed into the calculation of the sentiment phrase'score, the calculation of the sentence sentiment score, and the calculation of the text sentiment score.

1) CALCULATION OF THE SENTIMENT SCORE OF A SENTIMENT PHRASE

In the section of construction of dictionaries, the modifier words such as negative words and degree adverbs are introduced. Negative words play a key role in reversing the polarity of sentiment words, and degree adverbs play a key role in strengthening or weakening the score of sentiment word. At the same time, "!" is took into account. This article refers to the combination of sentiment words and these modifiers as sentiment phrases. The sentiment scores of sentiment phrases are calculated as shown in Table 10.

In Table 10, DA denotes degree adverbs, NA denotes negative words, SW denotes sentiment words, and S denotes sentiment phrases. V(SW) represents the sentiment score of the sentiment word, V(DA) represents the weight of the degree adverb, and NA represents the number of negative word.

The score computing method for sentiment phrases is shown in Table 11.

2) SCORE CALCULATION OF A SENTENCE AND TEXT SENTIMENT

The calculation of sentence sentiment score is based on sentiment phrases. Considering transitional conjunctions and progressive conjunctions, the sentiment scores of sentences can be calculated. The calculation method is shown in formula (3).

$$V(sentence) = V(lW)\sum_{i=1}^{n} V(s_i) \tag{3}$$

Among them, $V(s_i)$ is the sentiment value of the i-th sentiment phrase in the sentence; $V(lW)$ is the weight value of the conjunction. If there is a transition conjunction at the beginning of the sentence, $V(lW)$ is equal to 1.5; if there is a progressive conjunction at the beginning of the sentence, $V(lW)$ is equal to 2; if there is no conjunction at the beginning of the sentence, $V(lW)$ is equal to 1.

The sentence sentiment scores are summed to obtain the text sentiment score. According to the text sentiment score, the text sentiment tendency can be drawn. The rules are as follows:

(1) If the text sentiment score is more than 0, then the text sentiment tends to be active;
(2) If the text sentiment score is equal to 0, it is determined that the text sentiment tendency is neutral;
(3) If the text sentiment score is less than 0, it is determined that the text sentiment tends to be passive.

IV. EXPERIMENT AND ANALYSIS

Text sentiment analysis experiments are divided into three experiments. Experiment 1 is a study of the classification effect of the naive Bayesian field classifier. Experiment 2 is a study of the influence of the field sentiment words and the polysemic sentiment words on sentiment analysis. To further study the effect of the sentiment classifier proposed in this paper, experiment 3 compares it with SVM and the sentiment classification method in literature [20].

A. DATA SET CONSTRUCTION

In the experiment 1, 25000 reviews are used as the domain classification corpus (Data set 1) which are crawled from the websites of Jingdong and eLong. The domain classification corpus include hotel, clothing, fruit, digital and shampoo fields. There are 5000 reviews in each field.

In the experiment 2, 50000 reviews (Data set 2) are used as the sentiment classification corpus which are crawled from the review data on Ctrip and Jingdong websites. The sentiment classification corpus include hotel, clothes,

TABLE 11. Score computing method for sentiment phrases.

```
Input:   Sentence
Output:  The sentiment score of each sentiment phrase in the sentence
Process:
Words in the sentence are stored in the wordList, then traversing the wordList;
IF wordList[i] in the extended sentiment dictionary:
  IF wordList[i] is a polysemic word:
  Text category cls=NB_classifier.predict{document vector}
  Sc=1(active word in the field cls) or Sc=-1(passive word in the field cls)
  Elif wordList[i] is not a polysemic word:
  Sc=1(active word) or Sc=-1(passive word)
 End if
 Scaning the degree word and negative word before the sentiment word;
 Recording the position j and the weight dc of the degree adverb;
 Recording the position k of the negative word closest to the sentiment word;
 The number of negative word is neg_num.
 If there is only degree adverb before wordList[i]:
  Sc=Sc*dc
 Elif there is only privative before wordList[i]:
  Sc=Sc* $(-1)^{neg\_sum}$
 Elif there are privative and degree adverbs before wordList[i]:
  Update Sc according to the formulas of numbers 4, 5 and 6 in Table 10.
 End if
 If ！ exists behind wordList[i]
  Sc=Sc*2
 End if
End if
```

fruit, digital, and shampoo fields. 5000 active texts and 5000 passive texts are selected for sentiment analysis in each field. The corpus of experiment 1 and experiment 2 have no intersection.

In the experiment 3, the publicly available corpus is used. These are 6000 labeled hotel reviews (Data set 3) provided by Tan Songbo team [21] and used in literature [20]. Data set 3 are divided into three groups with equal number of active and passive reviews to conduct the experiment 3 [20].

In the experiment 2 and 3, for the sentiment analysis method proposed in this paper, 90% of Data set 1 are randomly selected as the training set and used to construct the naive bayesian field classifier.

B. EVALUATION CRITERIA

In this paper, precision, recall and F_1 are used to evaluate the classification results. Average_precision, average_recall, and average_F_1, which average the precision, recall and F_1 values for all fields, are used to evaluate the average classification performances for all fields.

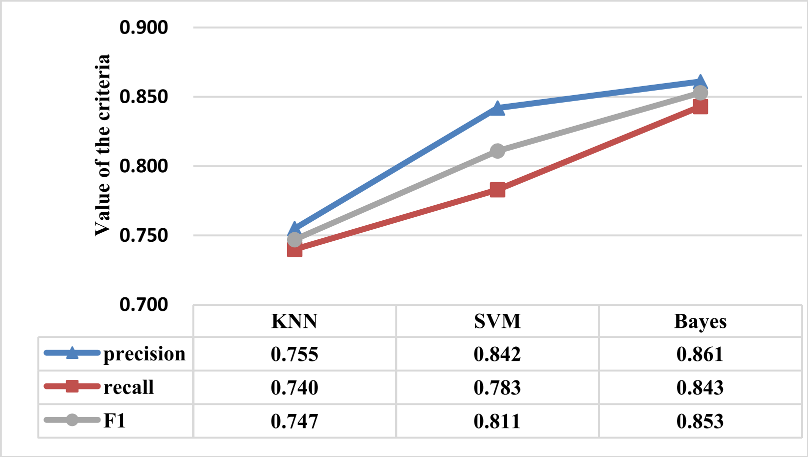
C. TEXT SENTIMENT ANALYSIS EXPERIMENT

1) EXPERIMENT 1: THE CLASSIFICATION EFFECT OF THE NAIVE BAYESIAN FIELD CLASSIFIER

The naive Bayesian field classifier is used to determine the field of the text in which the polysemic sentiment word is, and then the sentiment polarity of the word could be distinguished. Thus the field classification effect has an important

TABLE 12. The result of the text classification experiment.

Criteria Run	Precision	Recall	F_1
1	0.872	0.825	0.848
2	0.856	0.832	0.844
3	0.868	0.855	0.861
4	0.858	0.853	0.855
5	0.865	0.856	0.860
6	0.856	0.857	0.856
7	0.857	0.836	0.846
8	0.853	0.834	0.843
9	0.866	0.835	0.850
10	0.859	0.847	0.852
Average	0.861	0.843	0.853



	KNN	SVM	Bayes
precision	0.755	0.842	0.861
recall	0.740	0.783	0.843
F1	0.747	0.811	0.853

FIGURE 4. Relationship between three evaluation criteria and three field classifiers.

influence on the results of sentiment analysis. The experiment 1 is a study of the classification effect of naive Bayesian field classifier. Data set 1 is adopted to conduct the experiment 1. The experiment 1 performs 10 times of 10-fold cross-validation on Data set 1, and the experimental results are shown in Table 12.

It can be seen from the experiment results that the Bayesian field classifier works well for identifying the text field. The precision, recall and F_1 of each experiment are more than 80%. The average precision is 0.861. The average recall is 0.843. The average F_1 value is 0.853.

To further prove the feasibility of naive Bayesian field classifier, it is compared with KNN and SVM machine learning algorithms. The experimental results are shown in Figure 4.

As can be seen from Figure 4, compared with KNN and SVM algorithms, the precision, recall and F_1 of naive Bayesian field classifier are higher than those of KNN and SVM. Therefore, the feasibility of using the Bayesian algorithm to identify the text field is proved.

2) EXPERIMENT 2: COMPARISON OF SENTIMENT ANALYSIS EFFECTS BASED ON DIFFERENT SENTIMENT DICTIONARIES

HowNet sentiment dictionary and NTUSD are the initial sentiment dictionary. Meanwhile, some Internet popular sentiment words manually extracted from Chinese popular social network sites are added them as the basic sentiment dictionary. The extended sentiment dictionary contains the basic sentiment dictionary, the field sentiment words and the polysemic sentiment words in the hotel, digital, fruit, clothing and shampoo fields.

Basic sentiment dictionary is recorded as sentiment dictionary 1. Adding the field sentiment words to the basic

TABLE 13. Test results of using three sentiment dictionaries.

Field	Type	sentiment dictionary 1			sentiment dictionary 2			sentiment dictionary 3		
		P	R	F_1	P	R	F_1	P	R	F_1
hotel	Active	0.813	0.734	0.771	0.831	0.823	0.827	0.845	0.812	0.828
	Passive	0.778	0.697	0.735	0.845	0.735	0.786	0.850	0.776	0.811
	Average	0.795	0.716	0.753	0.838	0.779	0.807	0.848	0.794	0.820
clothing	Active	0.845	0.847	0.845	0.854	0.908	0.881	0.869	0.916	0.892
	Passive	0.904	0.654	0.759	0.924	0.656	0.767	0.935	0.756	0.836
	Average	0.874	0.751	0.802	0.889	0.782	0.824	0.902	0.836	0.864
fruit	Active	0.815	0.702	0.754	0.824	0.812	0.818	0.846	0.832	0.839
	Passive	0.809	0.656	0.725	0.854	0.713	0.777	0.866	0.762	0.811
	Average	0.812	0.679	0.739	0.839	0.763	0.798	0.856	0.797	0.825
digital	Active	0.814	0.658	0.728	0.865	0.722	0.787	0.874	0.805	0.838
	Passive	0.823	0.674	0.741	0.852	0.652	0.739	0.858	0.767	0.810
	Average	0.819	0.666	0.734	0.859	0.687	0.763	0.866	0.786	0.824
shampoo	Active	0.812	0.632	0.711	0.832	0.787	0.809	0.860	0.802	0.830
	Passive	0.821	0.670	0.738	0.872	0.698	0.775	0.878	0.753	0.811
	Average	0.817	0.651	0.724	0.852	0.743	0.792	0.869	0.778	0.820
Average	Active	0.820	0.715	0.762	0.841	0.810	0.824	0.859	0.833	0.845
	Passive	0.827	0.670	0.740	0.870	0.691	0.769	0.877	0.763	0.816
	Average	0.823	0.693	0.750	0.855	0.751	0.797	0.868	0.798	0.831

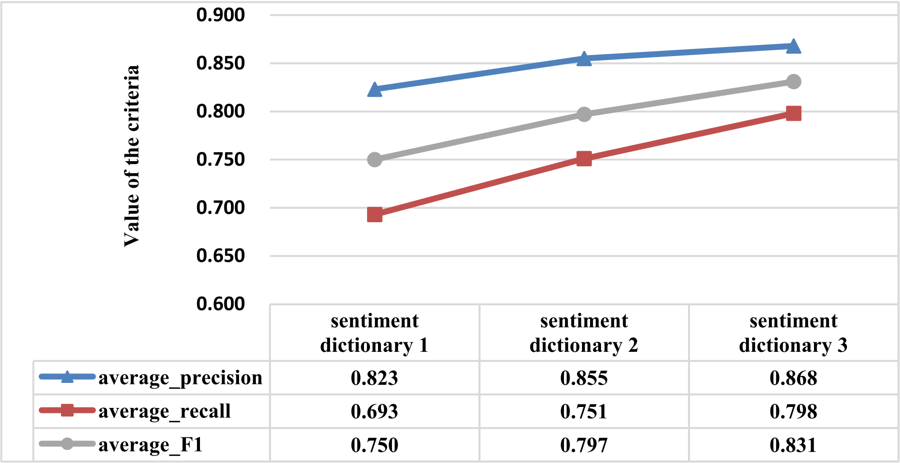
sentiment dictionary is recorded as sentiment dictionary 2. Adding the field sentiment words and the polysemic sentiment words to basic sentiment dictionary is recorded as sentiment dictionary 3. Sentiment dictionary 3 is namely the extended sentiment dictionary.

The experiment 2 is conducted based on Data set 2. Under the proposed sentiment semantic rules in this paper, sentiment analysis is conducted by using sentiment dictionary 1, sentiment dictionary 2 and sentiment dictionary 3, respectively. When using sentiment dictionary 3, the naive Bayesian field classifier is firstly used to determine the field of the text in which the polysemic sentiment word is. The naive Bayesian field classifier has the precision 0.862, the recall 0.835, and the F_1 0.848. The results of using three sentiment dictionaries are shown in Table 13.

In Table 13, P, R represent precision, recall respectively. Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9 are the graph analyses of the experimental results coming from Table 13.

Average_precision, average_recall and average_F_1 refer to the average of precision, recall and F_1 of five fields, respectively. Average_precision, average_recall, average_F_1 are shown in Figure 5.

As can be seen from Figure 5, the performance of the sentiment analysis based on the basic dictionary is not satisfactory, especially recall, which is only 0.693. After adding the field sentiment words, the effect of sentiment analysis improves significantly. The average_precision increases from 0.823 to 0.855, and the average_F_1 increases from 0.750 to 0.797. The average_recall increases most significantly, from 0.693 to 0.751. This is because the field sentiment words extend the coverage of sentiment dictionary. After considering the polysemic sentiment words with multiple polarities, the sentiment analysis effect is further improved. The average_precision,



	sentiment dictionary 1	sentiment dictionary 2	sentiment dictionary 3
▲ average_precision	0.823	0.855	0.868
■ average_recall	0.693	0.751	0.798
● average_F1	0.750	0.797	0.831

FIGURE 5. Relationship between threes criteria and threee sentiment dictionaries.

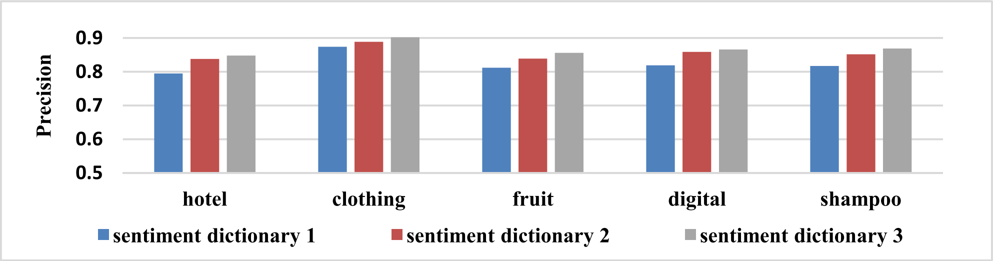


FIGURE 6. Comparison of precision of three sentiment dictionaries in each field.

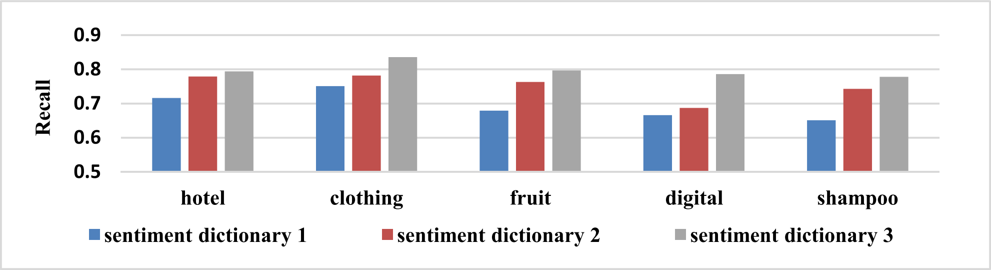


FIGURE 7. Comparison of recall of three sentiment dictionaries in each field.

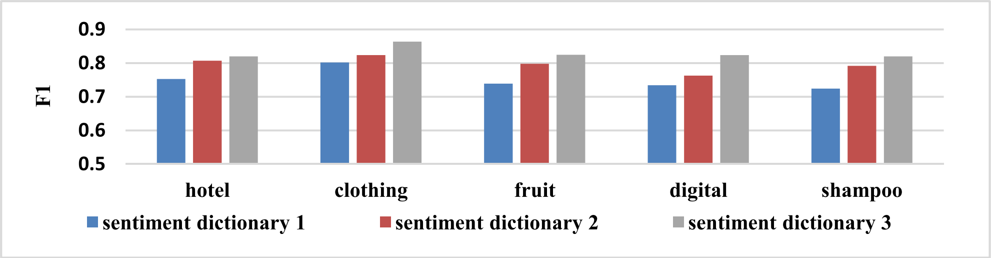
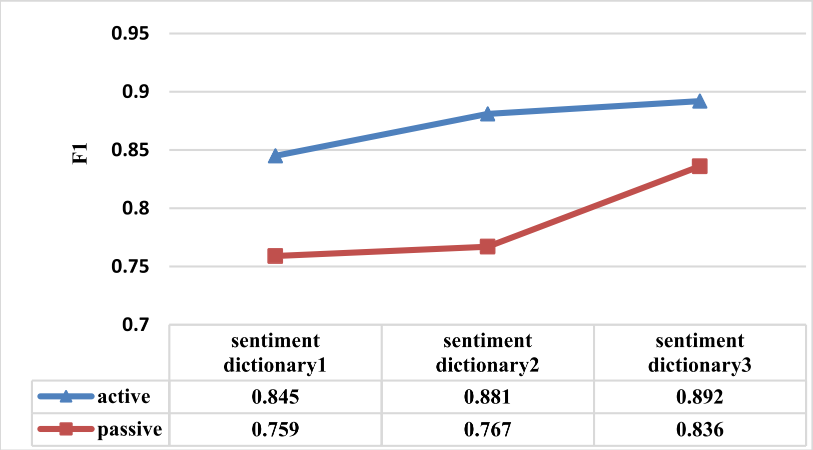


FIGURE 8. Comparison of F_1 of three sentiment dictionaries in each field.

average_recall, average_F_1 reaches 0.868, 0.798, and 0.831, respectively. In this paper, the domain classifier is introduced into the sentiment analysis method, and an extended sentiment dictionary containing the basic sentiment dictionary, the field sentiment words and the polysemic words is utilized for the sentiment computation. Therefore the ability of differentiating the text sentiment is stronger. Therefore the proposed sentiment analysis approach is better than the



	sentiment dictionary1	sentiment dictionary2	sentiment dictionary3
active	0.845	0.881	0.892
passive	0.759	0.767	0.836

FIGURE 9. Comparison of F_1 values of active texts and passive texts in the clothing field.

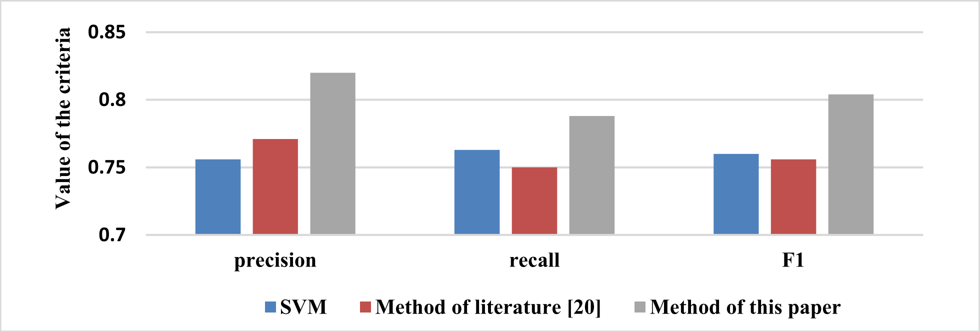


FIGURE 10. Comparison of precision, recall, and F_1 of three methods.

traditional sentiment analysis based on other two sentiment dictionaries.

Figure 6, Figure 7, Figure 8 show the comparison of precision, recall, and F_1 by using three sentiment dictionaries in each field. Compared with the basic sentiment dictionary, precision, recall, and F_1 in all fields have been improved after adding the field sentiment words. Furthermore, after adding the polysemic sentiment words, the precision, recall, and F_1 are further improved. When using an extended dictionary, precision, recall, and F_1 in the clothing field are the highest. As can be seen from the above, in the above-mentioned field, sentiment analysis based on extended sentiment dictionary is most suitable for the clothing field. Using the sentiment analysis method based on the extended sentiment dictionary proposed in this paper, the precisions and F_1 values in all fields are more than 0.80, and the recalls are all around 0.80. It can be proved that the sentiment analysis method based on extended dictionary proposed in this paper is applicable to the above five fields.

In addition, from the results of sentiment analysis in various fields, the sentiment classification effects of active and passive texts in some fields are quite different. This is due to the imbalance of sentiment words in some fields. For example, the result of sentiment analysis of the clothing field has a large difference between the F_1 values on the active texts and the passive texts, as is shown in the Figure 9. The F_1 of the acvitve texts is obviously higher than that of the passive texts. After analysis, it can be known that the reason for this phenomenon is that the number of active sentiment words in the clothing field is more than the passive sentiment words.

3) EXPERIMENT 3: COMPARISON WITH OTHER SENTIMENT CLASSIFIERS

To further verify the effectiveness of the sentiment analysis method proposed in this paper, based on Data set 3 coming from the literature [20], SVM and the method of the literature [20] are compared with the proposed method in this paper. Literature [20] is an improved method of the text sentiment classification based on semantic comprehension. Here the sentiment classification performances in literature [20] are utilized directly for comparison. The sentiment classification results in literature [20] are shown in Table 14. SVM is the classification method based on machine learning and is widely used in text sentiment analysis. The comment corpus of hotel field in Data set 2 is adopted as the training set of SVM method. First, the SVM sentiment classifier is trained. Then, the sentiment analysis of SVM and the proposed method is performed on three test sets which are obtained

TABLE 14. Sentiment classifier comparison experiment.

Method	Evaluation criteria	First group	Second Group	Third group	Average value
SVM	Precision	0.778	0.723	0.768	0.756
	Recall	0.764	0.745	0.781	0.763
	F_1	0.771	0.734	0.775	0.760
Method of literature [20]	Precision	0.794	0.726	0.794	0.771
	Recall	0.781	0.699	0.770	0.750
	F_1	0.790	0.702	0.776	0.756
Method of this paper	Precision	0.838	0.807	0.816	0.820
	Recall	0.792	0.766	0.807	0.788
	F_1	0.814	0.786	0.811	0.804

by dividing Data set 3 into three parts. In the experiment of the proposed sentiment analysis method, the naive Bayesian field classifier has the precision 0.855, the recall 0.843, and the F_1 value 0.849. Three group test results of SVM and the proposed method in this paper are also shown in Table 14. The average performance of the precision, recall, and F_1 of three methods are shown in Figure 10.

As can be seen from Figure 10 and Table 14, the precision of SVM is lower than that of the literature [20], but the recall and the F_1 are higher than those of the literature [20]. Compared with SVM and the method of literature [20], the proposed method significantly improves the precision, recall, and F_1 of sentiment analysis. In this paper, an extended sentiment dictionary which includes the basic sentiment words as well as covers some specific sentiment words in some detailed domains is constructed, which improves the effect of sentiment analysis.

The results of experiment 2 and experiment 3 show that the proposed sentiment analysis method in this paper is effective in the sentiment recognition of the comment texts.

V. CONCLUSION

In this paper, a Chinese sentiment analysis method based on extended dictionary is proposed. The main task of the research is to construct an extended sentiment dictionary covering five fields: hotel, digital, fruit, clothing and shampoo. The extended sentiment dictionary contains the basic sentiment dictionary, some field sentiment words and polysemic sentiment words in the fields. The naive Bayesian field classifier is used to classify the field of the text in which the polysemic sentiment word is, so the sentiment polarity of the word could be distinguished. The experimental results show that the sentiment classification method proposed in this paper has a good effect in hotel, clothing, fruit, digital and shampoo fields.

However, there are still many aspects which need to be considered. First, the weights of active words and passive words are "1" and "-1". The weights of sentiment words need to be further refined. Second, the generalization of sentiment classifier is limited because it's only used for a few specific fields. In future work, the above issues will be further researched.

REFERENCES

[1] V. Hatzivassiloglou and K. R. Mckeown, "Predicting the semantic orientation of adjectives," in *Proc. 35th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Valencia, Spain, 1997, pp. 174–181.

[2] A. Neviarouskaya and M. Aono, "Sentiment word relations with affect, judgment, and appreciation," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 425–438, Oct./Dec. 2014.

[3] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web-Internet Web Inf. Syst.*, vol. 17, no. 4, pp. 723–742, 2014.

[4] K. Matsumoto and F. Ren, "Construction of wakamono kotoba emotion dictionary and its application," in *Proc. Int. Conf. Comput. Linguistics Intell. Text Process.*, Tokyo, Japan: Springer-Verlag, 2011, pp. 405–416.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Language Process.*, Stroudsburg, PA, USA, 2002, pp. 79–86.

[6] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, no. 1, pp. 723–762, 2014.

[7] Y. Wang, X. Zheng, and X. Hu, "Short text sentiment classification of high dimensional hybrid feature based on SVM," *Comput. Technol. Develop.*, vol. 28, no. 2, pp. 88–93, 2018.

[8] F. Huang *et al.*, "Mining topic sentiment in microblogging based on multi-feature fusion," *J. Comput.*, vol. 40, no. 4, pp. 872–888, 2017.

[9] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, pp. 1–14, 2015.

[10] L. Wang and C. Cardie, "A piece of my mind: A sentiment analysis approach for online dispute detection," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Baltimore, MD, USA, 2014, pp. 693–699.

[11] A. Kumar and T. M. Sebastian, "Sentiment analysis on twitter," *Int. J. Comput. Sci. Issues*, vol. 9, no. 4, p. 372, 2012.

[12] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Trans. Inf. Syst.*, vol. 21, no. 4, pp. 315–346, 2003.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[14] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 5–19, 2016.

[15] Y. Li *et al.*, "A bilingual lexicon-based multi-class semantic orientation analysis for microblogs," *Chin. J. Electron.*, vol. 44, no. 9, pp. 2068–2073, 2016.

[16] X. Wu *et al.*, "Investigation on sentiment of reviews with shopping field dictionary construction," *Comput. Technol. Develop.*, vol. 27, no. 7, pp. 194–199, 2017.

[17] J. Wu *et al.*, "Sentiment analysis on Web financial text based on semantic rules," *J. Comput. Appl.*, vol. 34, no. 2, pp. 481–485 and 495, 2014.

[18] W. Zhong and L. Li, "A contrastive study on semantic prosodies of minimal degree adverbs in chinese and english," in *Proc. Joint Int. Semantic Technol. Conf.*, YiChang, China, 2015, pp. 11–15.

[19] Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in *Proc. Int. Conf. Comput. Inf. Sci. (ICIS)*, Wuhan, China, 2017, pp. 635–638.

[20] R. Wang *et al.*, "Research of text sentiment classification based on improved semantic comprehension," *Comput. Sci.*, vol. 44, no. 11A, pp. 92–97, 2017.

[21] S. Tan, "Chinese sentiment corpus," Inst. Comput. Technol. Chin. Acad. Sci., Zhongguancun Acad. Sci., Haidian District, Beijing, China. Accessed: Dec. 10, 2017. [Online]. Available: http://www.nlpir.org/wordpress/2017/09/04/

GUIXIAN XU was born in Changchun, Jilin, China, in 1974. She received the B.S. and M.S. degrees from the Changchun University of Technology, in 1998 and 2002, respectively, and the Ph.D. degree in computer software and theory from the Beijing Institute of Technology, in 2010. Since 2002, she has been a Teacher with the Information Engineering College, Minzu University of China, where she is currently an Associate Professor. Her research interests include data mining and machine learning.

ZIHENG YU was born in Taizhou, Zhejiang, China, in 1994. He received the B.S. degree in software engineering from Beijing Union University, in 2017. He is currently pursuing the master's degree in software engineering with the Minzu University of China. His research interests include data mining, natural language processing, and artificial intelligence.

HAISHEN YAO was born in Heze, Shandong, China, in 1989. He received the B.S. degree from the Shandong University of Science and Technology, in 2016. He is currently pursuing the master's degree in software engineering with the School of Information Engineering, Minzu University of China, Beijing, China. He is interested in scientific activities, such as data mining and natural language processing.

FAN LI was born in Jining, Shandong, China, in 1993. She received the B.S. degree in software engineering from Shandong Jianzhu University, in 2017. She is currently pursuing the master's degree in modern education technology with the Minzu University of China. Her research interests include artificial intelligence, natural language processing, and data mining.

YUETING MENG was born in Shijiazhuang, Hebei, China, in 1996. She received the B.S. degree in computer science and technology from the Hebei University of Science and Technology, in 2018. She is currently pursuing the master's degree in software engineering with the Minzu University of China. Her research interests include artificial intelligence, natural language processing, and data mining.

XU WU was born in Fenghuang, Hunan, China, in 1993. He received the B.S. degree in software engineering from the Chongqing University of Posts and Telecommunications, in 2017. He is currently pursuing the master's degree in modern education technology with the Minzu University of China. His research interests include data mining, natural language processing, and artificial intelligence.