IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 18, 2022, accepted March 13, 2022, date of publication March 16, 2022, date of current version March 28, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160172

A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis

HAO LIU, XI CHEN, AND XIAOXIAO LIU
Collaborative Innovation Centre of Assessment for Basic Education Quality, BNU, Beijing 100875, China
Institute of Education, University of Alberta, Edmonton, AB T6G 2G5, Canada

Corresponding author: Hao Liu (liuhao@bnu.edu.cn)

This work was supported in part by the National Social Science Fund of China under Grant 20CTJ019.

ABSTRACT The most commonly used methods in text sentiment analysis are rule-based sentiment dictionary and machine learning, with the later referring to the use of vectors to represent text followed by the use of machine learning to classify the vectors. Both methods have their limitations, including inflexibility of rules, non-prominence of sentiment words. In this paper, we design a weight distributing method combining the two methods for text sentiment analysis, by which the sentence vectors obtained can both highlight words with sentiment meanings while retaining their text information. Empirical results show that based on this new method, the accuracy rate of text sentiment analysis can reach as high as 82.1%, which means 13.9% higher than rule-based sentiment dictionary method, and 7.7% higher than TF-IDF weighting method.

INDEX TERMS Text sentiment analysis, sentiment dictionary, sentence vector, weights distribution.

I. INTRODUCTION

Natural language processing (NLP) is a way to systematically analyze, understand and extract information from text data. Because the capacity for language is one of the central features of human intelligence, NLP is one of the most important branches of artificial intelligence, which integrates social sciences (linguistics, logic, etc.), natural sciences (computer science, statistics, etc.) and engineering (electrical engineering, etc.) [1], [2]. The goal of NLP is to provide new computational capabilities around human language, which typically involves information extraction, machine translation, text generation and so on [2].

Text sentiment analysis is an important application of NL, which refers to ‘‘analysis of subjective texts with sentiment overtones to ta and classify their sentiment connotations and attitudes’’ [3]. With the rapid development of the Internet and social media, people often log on to different types of websites to express their opinions on current affairs or products, so individuals and organizations are increasingly using the content in these media for decision making [4]. It has become one of the hot research topics today to effectively obtain and analyze the sentiment information contained in the massive digital texts. Sentiment analysis applications have spread to almost every possible domain, such as consumer products, services, healthcare, financial services and social events [4]. By analyzing sentiment information, governments can effectively grasp the trend of public opinion by analyzing the sentiment orientation in these comments and provide a basis for policy making; businesses can tap the feedback information of customers on various products for businesses to develop more precise marketing strategies; and ordinary consumers can know other usrs' opinions about a product before making better purchasing decisions [5]. In the era of big data when the analysis of massive text data can only be done by computers, the common challenge faced by researchers in this area is how to improve the accuracy and efficiency of text sentiment analysis.

II. RELATED WORK

The initial steps of text sentiment analysis include data collection and data pre-processing, the later includes such processes as removal o invalid characters, word tokenizing and stop words filtering. This section reviews the work done in the field of text sentiment analysis from three aspects: sentiment analysis based on sentiment dictionary, sentiment analysis methods based on machine learning and sentiment analysis methods based on integration of sentiment dictionary and machine learning.

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

A. SENTIMENT ANALYSIS BASED ON SENTIMENT DICTIONARY

The most important indicators of sentiments are sentiment words, also called opinion words, which are commonly used to express positive or negative sentiments [4]. The method of rule-based sentiment dictionary analysis refers to statistically calculate the sentiment weights of sentiment words in a text based on a sentiment dictionary [6]. The main strategy is rule-based, using a dictionary of words marked with sentiment to determine the sentiment of a sentence, which is an unsupervised method [7]. This method is the most intuitive and is also similar to theprocess of people' recognizing sentiments in a text. In sentiment analysis, the more widely used dictionaries are HowNet sentiment dictionary from CNK, affective lexicon ontology of Dalian University of Technology, and Chinese sentiment dictionary from Taiwan University [8]. Some scholars have conducted related sentiment tendency analysis exploration based on this method. Haodong and Wenqi [6], have improved the average accuracy of sentiment tendency analysis of Chinese microblogs to some extent by combining HowNet sentiment dictionary and lexical ontology database, and incorporating numerous features of emoticons, semantic rules, negation words, degree adverbs and Internet neologisms. Peiyu and Yan [8] constructed a Chinese novel sentiment dictionary based on lexicon and Word2Vec, including basic sentiment dictionary, expanded dictionary and sentiment-imagery dictionary, and obtained sentiment tendency of sentences after sentiment score accumulation and averaging. Xu *et al.* [9] constructed an extended sentiment dictionary, including basic sentiment words, the field sentiment words and polysemy sentiment words, and obtained the sentiment polarity of the text by using the extended sentiment dictionary and the designed sentiment scoring rules. Yu and Egger [10] used the VADER algorithm, a sentiment-scoring method for text based on sentiment lexicons and rules, to explore how tourists feel about safety, service, queues, etc. when visiting popular and crowded attractions.

B. SENTIMENT ANALYSIS BASED ON MACHINE LEARNING

Machine learning, a branch of artificial intelligence (AI) and computer science, generates empirical-data-based models that can make decisions and judgments in new situations [11]. The method is often used to process and analyze large amounts of data and is widely used in finance, healthcare, education, and other fields. In text sentiment analysis, the sentiment analysis task is usually modeled as a classification problem. The researchers convert the text into a feature vector, and then feed these vectors to the model to generate predicted labels for the corresponding vectors [7].

The key to using machine learning for sentiment classification is to select text features suitable for sentiment classification [12], represent the text data with a vector, and then train and classify the model using machine learning methods. Many studies have focused on the feature selection. Jianzhong *et al.* [13], For example, based on four features, i.e., sentiment word frequency, the polarity offirst occurrence of sentiment word, emoji polarity, and negation words, to analyzespace-related text data collected from Web, by using SVM (support vector machine) and Naive Bayes, respectively. The results is satisfactory in terms of accuracy and recall. Word2vec is one of the most commonly used methods for generating representation vectors of words, which is a Word Embedding tool open sourced by Google in 2013, because of a few initial successes that motivate early adopters to do more, and leaving plenty of room for early adopters to contribute and benefit [14]. Embedding is essentially the representation of words with a low-dimensional vector, in which word vectors with close distances to each other correspond to words with similar meanings [15].

After obtaining the vectors representing the sentence using word vectors, the sentence vectors are then classified via machine learning. TF-IDF weighting method is one of the commonly used methods for generating sentence vectors based on word vectors. For example, Thomas and Latha [16] used TF-IDF for feature selection of Kanada text and used decision tree classifier to classify text. Soumya and Pramod [17] used methods such as BOW and TF-IDF to form feature vectors. They then used different machine learning techniques, such as Naive Bayes Machines (NB), Support Vector Machines (SVM), and Random Forests (RF), to classify tweets into positive and negative ones. Mukwazvure and Supreethi [18] also used TF-IDF for information weighting to generate sentence vectors and then used SVM and K-nearest neighbors for classification, which have achieved good results for text sentiment classification. Gang and Fei [19] proposed to perform sentiment clustering through voting mechanism of multiple clustering, which is a kind of unsupervised machine learning. Specifically, each clustering uses the TF-IDF feature weighting method, which overcomes the low accuracy and instability of K-means.

As a new direction of machine learning, deep learning has gradually been widely used in sentiment analysis because it is more complex and accurate than traditional machine learning models. Jane [20] utilized deep learning techniques such as the Doc2Vec algorithm, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) to extract insights valuable to Electric Vehicles buyers, marketers and manufacturers. Wang *et al.* [21] used an unsupervised BERT (Bidirectional Encoder Representation for Transformer) model to classify texts on Sina Weibo into positive, neutral, and negative, and used a TF-IDF model to summarize the topics of posts. Aydin and Gungor [22] proposed a novel neural network framework that combines recurrent and recursive neural network models. Recurrent models propagate information about sentiment labels throughout word sequences; Recursive models extract syntactic structure from text. The neural network framework achieved state-of-the-art results and outperformed the baseline study by a significant margin.

C. SENTIMENT ANALYSIS INTEGRATING SENTIMENT DICTIONARY AND MACHINE LEARNING

Although both methods have good performance in sentiment analysis, their shortcomings are also obvious. The rules based on sentiment dictionary relies too much on sentiment words and the rules are not flexible enough, while the machine learning approach does not highlight the important role of emotional words. To address these drawbacks, some scholars have improved the method of sentence vector generation based on sentiment dictionaries: Yang [23] proposed to assign weights of 2 to sentiment words and 1 to neutral words, and then sum up the word vectors obtained from Word2Vec training based on the weights to obtain the corresponding sentence vectors; Hui *et al.* [12] adjusted the sentiment of the word vectors obtained from Word2Vec training to obtain word vectors considering both semantic and sentiment tendencies, and used TF-IDF values for the word vectors to weight and sum to obtain the text vector representation, and machine learning methods are used to classify the text for sentiment. Dashtipour *et al.* [24] firstly judged sentence sentiment polarity based on sentiment dictionary and rules. If the sentence could not be classified, the concatenated fastText embedding of the sentence is inputted to the DNN (Deep Neural Networks) to determine the polarity of the sentence. Yang *et al.* [25] used the BERT model to train word vectors, used a sentiment dictionary to enhance the sentiment features in the text, and then went through a convolutional layer and a pooling layer to classify the weighted sentiment features. Chiny *et al.* [26] proposed a hybrid sentiment analysis model based on Long Short-Term Memory network (LSTM), a rule based sentiment dictionary (VADER) and TF-IDF weighting method. The above three methods each get a sentiment score, and then treated the three scores as three inputs, ans used classification models such as Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) for sentiment polarity classification.

Currently, the major difficulty facing researchers in the area is how to combin sentiment dictionary and machine learning. To solve this problem, we proposes the Sentiment Dictionary Weighting method based on the TF-IDF, which combines sentiment dictionary and pre-trained word vectors, which enables obtained sentence vectors to retain text information while highlighting the words with sentiment tendencies.

III. RESEARCH METHODOLOGY

A. RULES BASED ON SENTIMENT DICTIONARY

The sentiment dictionary used in this study is a Chinese ontology resource compiled and labeled by the Information Retrieval Research Laboratory of Dalian University of Technology. The resource classifies sentiments into 7 major categories and 20 subcategories, which describe a Chinese word or phrase from different perspectives, including word lexical category, sentiment category, sentiment intensity and polarity, etc. [27], [28]. In this study, 2 sentiment major categories "乐 (happy)," "好 (good)" are combined into "positive sentiment" and 5 sentiment major categories "怒 (angry)," "哀 (sad)," "惧 (scared)," "恶 (disgusted)," "惊 (astonished)" are combined into "negative sentiment," and the sentiment score of the sentence is obtained by semantic rules and the intensity of the sentiment words, and if the score is greater than 0, the sentence is judged as positive sentiment; if the score is less than 0, it is considered as negative sentiment; if the score is equal to 0, it is considered as neutral sentiment.

The rules used in this study include the following three.

Rule 1: When a negative word, e.g., "不 (not)," "没有 (without)," etc., appears before a sentiment word, and there is no punctuation between the sentiment word and the negative word, the sentiment is reversed, and the weight of thenegative word is −1. For example, "我不开心 (I'm not happy). The word (开心) happy" is an positive sentiment word with an intensity of 7, so the sentence has an sentiment score of −7 and is finally judged as a negative sentiment.

Rule 2: When there is an adversative conjunction, such as "但是 (but)," "然而 (however)," etc.), the sentiment before the transitive word is reversed. For example, "虽然他废寝忘食, 但是成绩还是不令人满意 (Although he studied hard, his performance was still unsatisfactory)." The strengths of "废寝忘食 (someone study so hard that he forget to eat and sleep)" and "令人满意 (satisfactory)" are 7 and 5, respectively. The adversative conjunction "但是 (but)" makes the weight of "废寝忘食" be −1, and the negative word "不 (not)" makes the weight of "令人满意 (satisfactory)" be −1. The final sentiment score of the sentence is −12, which is finally judged as a negative sentiment. When there are multiple sentences in a paragraph, the adversative conjuctions only reverses the sentence that it is in, rather than the sentiment of all the sentences, so it needs to be calculated sentence by sentence.

Rule 3: When there are adverbs of degree, e.g., "一点 (a little)," "非常 (very)," etc., they are given different weights. In this study, the degree adverbs are divided into 5 categories: "极其 (extremely)," "很 (very)," "较 (more)," "稍 (slightly)," and "欠 (less)," are given weights of 1.5, 1.4, 1.2, 0.8, and 0.5, respectivel. When the degree adverb and the negative word appear at the samesentence, thenegative word is given a weight of −0.5 if it comes before the degree adverb, and a weight of −1.5 if it comes after the degree adverb. For example, "我很不高兴 (I'm very unhappy)." The strength of the word "开心 (happy)" is 7, the weight of the degree adverb "很 (very)" is 1.4, and the weight of thenegative word is −1. because it is after the degree adverb. Therefore, the final score is −14.7, which is a negative sentiment.

As the most widely used one, this method has the advantages of easy operation and highlighting the role of sentiment words, but it also has many shortcomings Only sentiment words, negative words and related words are considered with other information in the sentence being overlooked, and the rules of classification are rigid and inflexible, among others.

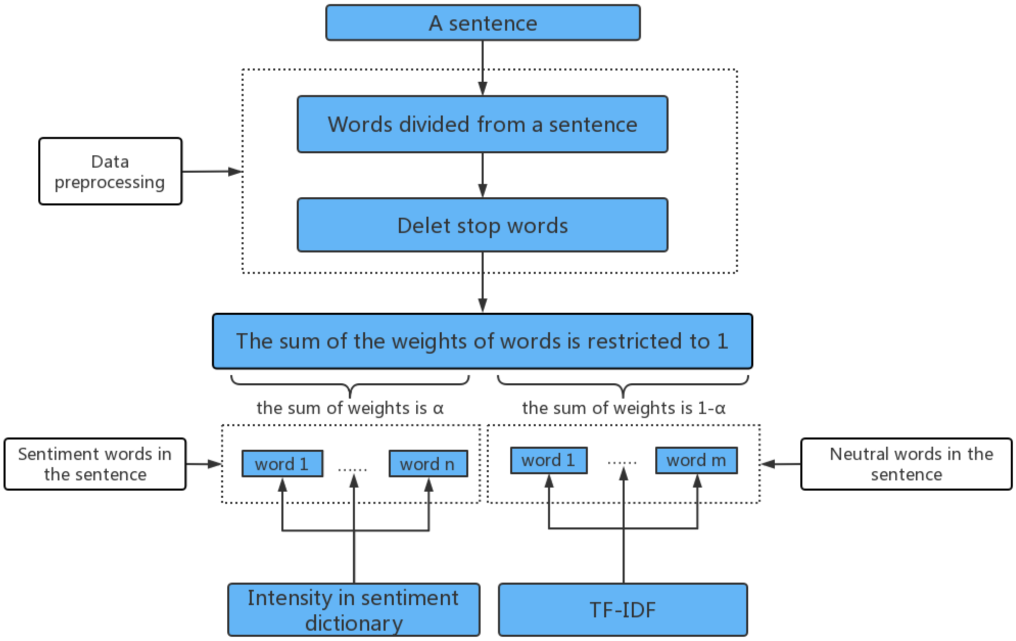
A sentence
Data preprocessing
Words divided from a sentence
Delet stop words
The sum of the weights of words is restricted to 1
the sum of weights is α
the sum of weights is 1-α
Sentiment words in the sentence
word 1
......
word n
word 1
......
word m
Neutral words in the sentence
Intensity in sentiment dictionary
TF-IDF

FIGURE 1. The framework of weight distribution method proposed in this paper.

B. TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) is a common weighting technique used in information retrieval and data mining. TF or term frequency refers to the frequency of word in a text) and IDF refers to the inverse text frequency index. The essential idea underlying TF-IDF is that words that appear more frequently in one document and less in others should be more important because they are more useful for classification [26]. Therefore, It is widely used in keyword extraction, text similarity comparison and topic classification. The TF-IDF algorithm is used to calculate every word and its weight value in every text of a document. It takes into account the probability TF of a word's occurrence in a single text and the weight IDF of the word in the entire document [29]. Assuming that there are N texts in one document, the weight of word w in text s is calculated as

$$t_{sw} = tf_w \times idf_w = tf_w \times log(\frac{N}{N_w}) \tag{1}$$

where N_w is the number of texts containing the word w. The Institute of Chinese Information Processing of Beijing Normal University provides 36 open source pre-trained word vector libraries on GitHub [30]. In this study, we use one of these pre-trained Chinese word vector libraries consisting of 8599 modern Chinese literary works, and each word is represented by a 300-dimensional vector of real numbers. The TF-IDF values of each word are used as weights, and the pre-trained word vectors are weighted and summed to obtain a vector of text s.

$$v_s = \sum\nolimits_{w\in s} t_{sw} * v_w \tag{2}$$

After the sentence vectors are generated, sentiment classification can be made by using tools such as logistic regression, support vector machines, Bayesian classifiers, and neural networks. This is a supervised learning process that manually mark the sentiment of the training set of texts. This method retains information of all words and is more flexible in application compared wit rule-based sentiment dictionary method. But there is, however a major deficiency in it, that is, the weight of each word in sentiment classification is only related to its frequency of occurrence and does not highlight the importance of sentiment words.

C. A NEW METHOD COMBINING SENTIMENT DICTIONARY AND TF-IDF

In order to combine the advantages of the above two methods, we designed a new sentence vector generation method. To both highlight the role of sentiment words in sentiment classification and maintain its flexibility, this study tries to improve the TF-IDF vector weighting method. Using the ontology library of sentiment words from Dalian University of Technology and the word vector library provided by the Institute of Chinese Information Processing of Beijing Normal University, the set of all sentiment words is noted as E. First, the weight sum of a text is constrained to 1. Then, the weight sum 1 is divided into two parts, the weight α is assigned to all sentiment words and 1-α is assigned to all neutral words in the text, and α is assigned to sentiment words in proportion to the strength of sentiment words, and the neutral words are assigned weights 1-α proportionally to their TF-IDF values. Thus we get the vector equation for text s:

$$v_s = \sum\nolimits_{w\in s} [\alpha \frac{d_w}{D_s} I_{w\in E} + (1-\alpha)\frac{t_w}{T_S} I_{w\notin E}]v_w \tag{3}$$

Among them, D_s is the sum of the intensities of all sentiment words and T_S is the sum of the TF-IDF values of

TABLE 1. The vector computing method for sentiment classification.

```
input:   a sentence
output:  the vector of the sentence
process:
Words is vectors set of words divided from sentence and Jieba tokenizer is used in Chinese text.
Delete stop words in the sentence.
sentiment = $\emptyset$, sentiment value = $\emptyset$, no sentiment = $\emptyset$, neutral value = $\emptyset$, V(sentence) = zeros()
for i in Words:
  if i in sentiment dictionary:
    sentiment = sentiment $\cup$ {i}, sentiment value = sentiment value $\cup$ {value[i] from sentiment dictionary}
  else: no sentiment = no sentiment $\cup$ {i}, neutral value = neutral value $\cup$ {tf-idf(i)}
sentiment value /= sum(sentiment value), neutral value /= sum(neutral value)
if sentiment = $\emptyset$:
  for i in length(Words):
    V(sentence) += neutral value[i]*Words[i]
else: a = 0, b = 0
  for i in length(Words):
    for j in length(sentiment):
      a += sentiment value[j]*dot(sentiment[j],Words[i] )
    for j in length(no sentiment):
      b += neutral value[j]*dot(no sentiment[j],Words[i] )
  If a>b: $\alpha$ = 0.5
else: $\alpha$ = 0.75
  for i in length(sentiment):
    V(sentence) += $\alpha$* sentiment value[i]*sentiment[i]
  for i in length(no sentiment):
    V(sentence) += (1-$\alpha$)* neutral value[i]*no sentiment[i]
return(V(sentence))
```

all neutral words in text s. I_C is a schematic function and takes 1 when condition C is satisfied, otherwise it takes 0. The framework of weight distribution method proposed in this paper is shown in Figure 1.

As the weight sum of sentiment words, α represents the importance of sentiment words. To highlight the importance of sentiment words, it cannot be too small, but if it is too large, the role of neutral words would be reduced. According to the latent variable generation model of Arora *et al.* [31], the generation of sentences is considered as a dynamic process, where the sentence vector does a slow random tour in order to generate similar words in the context at moment t. his means that the probability of the occurrence of words at moment t is related to the sentences that already exist at moment t. Assuming that the sentence vector does not change much when words appear in the sentence, simply replace all in the sentence with the sentence vector [32], i.e.

$$P(w \in s|v_s) \propto exp(v_s * v_w) \tag{4}$$

Thus we make a hypothesis that when the sentence vector of sentence s is known, the probability that word w appears in sentence s is proportional to the exponent of the inner product of the sentence vector and the word vector. To facilitate the calculation, we assume that the intercept term is 0, so

$$P\,(w \in s\,|\,v_s) = A \times exp\,(v_s * v_w)$$

where A is a constant, and we obtain a log-likelihood function for a sentence

$$\begin{aligned}\mathrm{lnL}(\mathrm{w}, v_{\mathrm{s}}) &= \ln[\prod_{\mathrm{w}\in\mathrm{s}} \mathrm{A} \times \exp(v_{\mathrm{s}} * v_{\mathrm{w}})] \propto \sum_{\mathrm{w}\in\mathrm{s}} v_{\mathrm{s}} * v_{\mathrm{w}} \\ &= \sum_{\mathrm{w}\in\mathrm{s}} \sum_{w'\in s} [\alpha \frac{d_{w'}}{D_s} I_{w'\in E} + (1-\alpha) \frac{t_{w'}}{T_S} I_{w'\notin E}] v_{w'} * v_{\mathrm{w}} \\ &= \alpha \sum_{\mathrm{w}\in\mathrm{s}} \sum_{w'\in s1} \frac{d_{w'}}{D_s} v_{w'} * v_{\mathrm{w}} \\ &\quad + (1-\alpha) \sum_{\mathrm{w}\in\mathrm{s}} \sum_{w'\in s2} \frac{t_{w'}}{T_S} v_{w'} * v_{\mathrm{w}}\end{aligned}$$

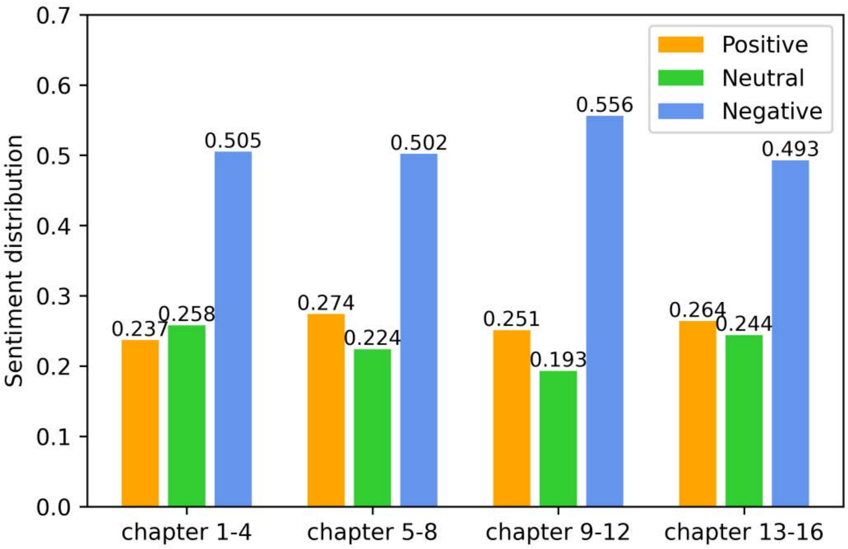

FIGURE 2. Sentiment distribution in the text data.

where s1 is the set of sentiment words in s and s2 is the set of neutral words in s. We define:

$$\mathrm{a} = \sum\nolimits_{\mathrm{w} \in \mathrm{s}} \sum\nolimits_{w' \in s1} \frac{d_{w'}}{D_s} v_{w'} * v_{\mathrm{w}},$$
$$\mathrm{b} = \sum\nolimits_{\mathrm{w} \in \mathrm{s}} \sum\nolimits_{w' \in s2} \frac{t_{w'}}{T_S} v_{w'} * v_{\mathrm{w}} \tag{5}$$

It can be seen that, if a $<$ b, the smaller the α, the largerthe likelihood function, which means that for the whole sentence structure, neutral words are more important than sentiment words; given the importance of sentiment words in sentiment classification, sentiment words and neutral words can be considered as equally important, so α is taken as 0.5. When a $>$ b, the larger α is, the larger the likelihood function, which means that for the whole sentence structure, sentiment words are more important than neutral words. Considering the importance of sentiment words in sentiment classification, the weight of sentiment words can be enlarged, but it cannot be taken as 1, otherwise the role of neutral words will be completely eliminated. So α takes 0.75, the middle value between 0.5 and 1. In summary, the following formula is used to determine the values taken in the text s.

$$\alpha_{\mathrm{s}} = \begin{cases} 0.75 & \text{if a} > \text{b} \\ 0.5 & \text{if a} < \text{b} \\ 0 & \text{if S1} = \emptyset \end{cases} \tag{6}$$

The vector computing method for sentiment classification is shown in Table 1.

IV. EMPIRICAL STUDY

A. DATA SOURCES

In the experiment, Yu Hua’s novel “"在细雨中呼喊 (Cries in the Drizzle)” was used as the domain classification corpus which includes 16 chapters. We utilized leave-four-chapters-out cross-validation to compare the predictive power of the three methods mentioned, each experiment uses four chapters as the test set and the remaining 12 chapters as the training set, repeating four times altogether. The sentiment word of novel text were manually marked. Because one assumption that researchers often make about sentence-level analysis is that a sentence expresses a single sentiment from a single opinion holder [4] long, emotionally-rich paragraphs are segmented so that each paragraph contained only one kind of sentiment as much as possible. A total of 2079 sentences were obtained, including 486 neutral sentences, 535 sentences with positive sentiment and 1058 sentences with negative sentiment. There are 503 sentences in chapter 1-4, 478 sentences in chapter 5-8, 363 sentences in chapter 9-12, 735 sentences in chapter 13-16. Figure 2 shows the distribution of sentiment in the novel. It can be found that the sentiment distribution of each part of the text is almost the same, and there are more negative sentiment sentences than positive sentiment and neutral sentences, which account for almost half of the total, which is in line with the tragic characteristics of loneliness, fear and sadness of the novel.

B. EVALUATION METRICS

All three methods determine the sentiment category (negative sentiment, neutral sentiment, andpositive sentiment) of each sentence. For the overall performance evaluation, accuracy is used to measure.

$$\text{accuracy} = \frac{\text{Number of texts correctly classified}}{\text{Total number of texts}}$$

In order to show the details of the three methods more clearly, we use precision, recall and F1 score as evaluation

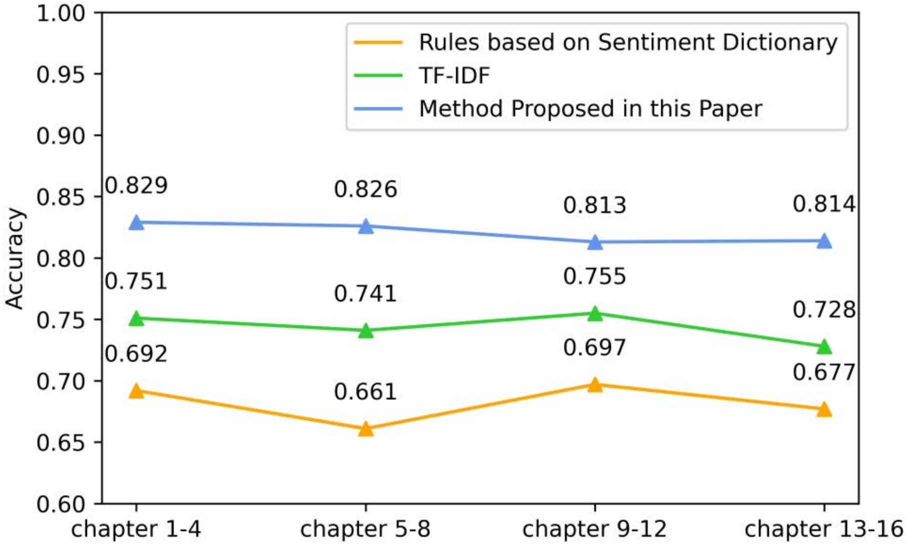

FIGURE 3. Comparison of accuracy of three sentiment classification methods.

TABLE 2. Comparison of the precision of three sentiment classification methods.

	method	Chapter1-4	Chapter5-8	Chapter9-12	Chapter13-16
Positive	Rules Based on Sentiment Dictionary	0.514	0.528	0.504	0.523
	TF-IDF	0.650	0.730	0.675	0.695
	Method Proposed in this Paper	**0.774**	**0.782**	**0.747**	**0.763**
Neutral	Rules Based on Sentiment Dictionary	0.748	0.728	**0.789**	0.765
	TF-IDF	0.802	0.696	0.631	0.719
	Method Proposed in this Paper	**0.884**	**0.804**	0.700	**0.860**
Negative	Rules Based on Sentiment Dictionary	0.813	0.765	0.813	0.775
	TF-IDF	0.776	0.765	0.842	0.745
	Method Proposed in this Paper	**0.830**	**0.863**	**0.898**	**0.82**

criteria, where:

$$\text{precision} = \frac{\text{the number of texts correctly classified into the category}}{\text{the number of texts classified into the category}}$$

$$\text{recall} = \frac{\text{the number of texts correctly classified into the category}}{\text{Total number of texts in the category}}$$

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

C. COMPARISON OF THE CLASSIFICATION PERFORMANCE OF THE THREE SENTIMENT ANALYSIS METHOD

In order to test the effectiveness of the above three methods for sentiment classification in novels, they were applied to the leave-four-chapters-out cross-validation for sentiment classification, and the classification performance of the three methods was measured using the above criteria. Support vector machine (SVM) is used to classify sentiment from the sentence vectors obtained by TF-IDF weighting method and the method of this paper. SVM is a linear classifier with maximum interval defined on the feature space [33], and the basic idea is to find a maximally spaced division hyperplane in the sample space based on the training set to separate different classes of samples [11]. Define a hyperplane by

$$\{x : f(x) = \mathrm{x}^{\mathrm{T}}\beta + \beta_0 = 0\}$$

The optimization problem is

$$\min_{\beta,\beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$\text{Subject to } \xi_{\mathrm{i}} \geq 0, \quad \mathrm{y_i}(\mathrm{x}^{\mathrm{T}}\beta + \beta_0) \geq 1 - \xi_{\mathrm{i}}\ \forall i$$

TABLE 3. Comparison of the recall of three sentiment classification methods.

	method	Chapter1-4	Chapter5-8	Chapter9-12	Chapter13-16
Positive	Rules Based on Sentiment Dictionary	**0.773**	**0.794**	**0.736**	**0.753**
	TF-IDF	0.655	0.641	0.615	0.624
	Method Proposed in this Paper	0.748	**0.794**	0.714	0.732
Neutral	Rules Based on Sentiment Dictionary	0.685	0.729	0.757	0.615
	TF-IDF	0.662	0.551	0.429	0.564
	Method Proposed in this Paper	**0.762**	**0.841**	**0.900**	**0.721**
Negative	Rules Based on Sentiment Dictionary	0.669	0.637	0.772	0.693
	TF-IDF	0.831	0.800	0.817	0.839
	Method Proposed in this Paper	**0.902**	**0.838**	**0.827**	**0.903**

TABLE 4. Comparison of the F1 score of three sentiment classification methods.

	method	Chapter1-4	Chapter5-8	Chapter9-12	Chapter13-16
Positive	Rules Based on Sentiment Dictionary	0.617	0.634	0.598	0.617
	TF-IDF	0.653	0.683	0.644	0.658
	Method Proposed in this Paper	**0.761**	**0.788**	**0.730**	**0.747**
Neutral	Rules Based on Sentiment Dictionary	0.702	0.628	0.555	0.649
	TF-IDF	0.738	0.712	0.688	0.663
	Method Proposed in this Paper	**0.818**	**0.822**	**0.788**	**0.784**
Negative	Rules Based on Sentiment Dictionary	0.734	0.695	0.792	0.732
	TF-IDF	0.802	0.782	0.829	0.789
	Method Proposed in this Paper	**0.864**	**0.850**	**0.861**	**0.859**

where the slack variable ξ_1 is the proportional amount by where the prediction $f(x) = \mathrm{x}^{\mathrm{T}}\beta + \beta_0$ is on the wrong side of its margin [34]. The cost parameter C represents the penalty for misclassification. Kernel function is used in nonlinear classification tasks. The basic idea is to use a transformation to map the sample points to a new space, making it linearly separable [30]. The most popular choice for kernel function in SVM literature is radial basis

$$\mathrm{K}(\mathrm{x}, \mathrm{x}') = \exp(-\gamma \left\|\mathrm{x} - \mathrm{x}'\right\|^2)$$

To maximize predictive ability of the model, parameter selection is first made. A grid search method was used for the parameters C and γ of the SVM, and a 5-fold cross-validation was made for each combination of the two parameters, respectively, The training set is randomly divided into five subsets, four subsets are used for model training, and the remaining one is used for testing. The above process is repeated five times, and the average of the five results is taken to represent the classification performance of the two parameter combinations. We found that the optimal parameter combination of both the TF-IDF and the method of this paper is C = 10 and γ = 0.00. The model was trained with the training set and the sentiment classification was done on the test set, and Figure 3 is a comparison of the accuracy of the three sentiment classification methods.

As can be seen from Figure3, the performance of the rule-based on sentiment dictionary method is not satisfactory, the accuracy is between 60% and 70%. The method proposed in this paper performed the best in four experiments, with the accuracy rate above 80%. The mean values of the accuracy of the three methods are 68.2%, 74.4% and 82.1%, respectively. Table 2, Table 3, and Table 4 show the comparison of precision, recall and F1 score of the experimental results of leave-four-chapters-out cross-validation.

From the result of our experiment, the rule-based sentiment dictionary method has a significant gap in classification effect compared with the other two methods; the TF-IDF, which does not consider sentiment tendency, also has significant shortcomings for sentiment analysis. The method proposed in the paper combines the advantages of the two, and the classification effect show an obvious improvement, in which the precision is improved by 13.9% compared with the rule-based sentiment dictionary method and 7.7% compared with the TF-IDF weighting method. The good results of the method proposed in the paper in classification illustrate that the method fully extracts the sentiment information in the text and verifies its effectiveness in sentiment analysis.

V. CONCLUSION

In this paper, a sentence vector generation method based on sentiment dictionaries and pre-trained word vectors is proposed for sentiment classification, which calculates the weights of sentiment words and neutral words in a sentence separately, and retains the overall information of the sentence while highlighting the sentiment words. A supervised machine learning method is used to achieve the sentiment polarity determination of the text (in this paper, SVM is used for classification), and the advantages and disadvantages of the classification effect are compared with the commonly used rule-based sentiment dictionary and TF-IDF weighting methods. Yu Hua's novel Cries in the Drizzle was used as the domain classification corpus. From the experimental result, the accuracy of text sentiment analysis using the method proposed in this paper reaches 2.1%, which is 13.9% and 7.7% higher respectively than the other two methods.

It must be admitted that there are still some limitations and shortcomings in the proposed method, which needs further improvemen. The method proposed in this paper cannot break the boundary of single sentence, which cuts the connection between text and context, which hinders the study of paragraph and discourse comprehension. Second, sentiment analysis at the sentence level is often insufficient for applications because they do not identify opinion targets or assign sentiments to such targets [4]. In addition, it can be seen from the classification results that the classification results of the three methods are significantly better in negative sentiment than neutral and positive sentiment. The reason for this phenomenon is that the number of positive sentiment sentences is more than the negative and neutral sentences. It indicates that the classification performance is interfered by the imbalance of text sentiment.

REFERENCES

[1] F. Zhi-wei, "Academic position of natural language processing," *J. PLA Univ. Foreign Languagesguage Process.*, vol. 3, no. 28, pp. 1–8, 2005.

[2] J. Eisenstein, *Natural Language Processing*. Cambridge, MA, USA: MIT Press, 2018.

[3] W. Ting and Y. Wenzhong, "Review of text sentiment analysis methods," *Comput. Eng. Appl.*, vol. 57, no. 12, pp. 11–24, 2021.

[4] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool, 2012.

[5] C. Long, G. Ziyu, H. Jianhong, and P. Jinye, "A survey on sentiment classification," *J. Comput. Res. Develop.*, vol. 54, no. 6, pp. 1150–1170, 2017.

[6] Z. Haodong and L. Wenqi, "Chinese micro-blog emotional analysis method based on semantic rules and expression weighting," *J. Light Ind.*, vol. 35, no. 2, pp. 74–82, 2020.

[7] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," *Global Transitions Proc.*, vol. 2, no. 2, pp. 205–211, Nov. 2021.

[8] S. Peiyu and X. Yan, "Research on multi–feature fusion method for sentiment analysis of Chinese microbiog," *Electron. World*, vol. 2, pp. 20–21, Feb. 2018.

[9] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access*, vol. 7, pp. 43749–43762, 2019.

[10] J. Yu and R. Egger, *Tourist Experiences at Overcrowded Attractions: A Text Analytics Approach*. Springer, 2021, pp. 231–243.

[11] Z. Zhihua, *Machine Learning*. Beijing, China: Tsinghua Univ., 2016.

[12] D. Hui, X. Xueke, W. Dayong, L. Yue, Y. Zhihua, and C. Xueqi, "A sentiment classification method based on sentiment-specific word embedding," *J. Chin. Inf. Process.*, vol. 31, no. 3, pp. 170–176, 2017.

[13] X. Jianzhong, Z. Jun, Z. Rui, Z. Liang, H. Liang, and L. Jiaojiao, "Sentiment analysis of aerospace microblog using SVM," *J. Inf. Secur. Res.*, vol. 3, no. 12, pp. 1129–1133, 2017.

[14] K. W. Church, "Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017, doi: 10.1017/S1351324916000334.

[15] C. Deguang, M. Jinlin, M. Ziping, and Z. Jie, "Review of pre-training techniques for natural language processing," *J. Frontiers Comput. Sci. Technol.*, vol. 15, no. 8, pp. 1359–1389, 2021.

[16] V. Rohini, M. Thomas, and C. A. Latha, "Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 503–507.

[17] S. Soumya and K. V. Pramod, "Sentiment analysis of Malayalam tweets using machine learning techniques," *ICT Exp.*, vol. 6, no. 4, pp. 300–305, Dec. 2020.

[18] A. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments," in *Proc. 4th Int. Conf. Rel., Infocom Technol. Optim. (ICRITO) (Trends Future Directions)*, Sep. 2015, pp. 1–6.

[19] G. Li and F. Liu, "A clustering-based approach on sentiment analysis," in *Proc. IEEE Int. Conf. Intell. Syst. Knowl. Eng.*, Nov. 2010, pp. 331–337.

[20] R. Jena, "An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach," *Ind. Marketing Manage.*, vol. 90, pp. 605–616, Oct. 2020.

[21] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 sensing: Negative sentiment analysis on social media in China via BERT model," *IEEE Access*, vol. 8, pp. 138162–138169, 2020, doi: 10.1109/ACCESS.2020.3012595.

[22] C. R. Aydin and T. Gungor, "Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations," *IEEE Access*, vol. 8, pp. 77820–77832, 2020.

[23] G. Yang, "Research and application of sentiment analysis based on Word2Vec method," Xiamen Univ., Xiamen, China, Tech. Rep., 2019.

[24] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks," *Neurocomputing*, vol. 380, pp. 1–10, Mar. 2020.

[25] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[26] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "LSTM, VADER and TF-IDF based hybrid sentiment analysis model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, p. 2021, 2021, doi: 10.14569/IJACSA.2021.0120730.

[27] X. Lin-Hong, L. Hong-Fei, and Z. Jing, "Construction and analysis of emotional corpus," *J. Chin. Inf. Process.*, vol. 22, no. 1, pp. 116–122, 2008.

[28] 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.

[29] T. MingZ Lei and Z. Xian-chun, "Document vector representation based on Word2CVec," *Comput. Sci.*, vol. 6, no. 43, pp. 214–217, 2016.

[30] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on Chinese morphological and semantic relations," 2018, *arXiv:1805.06504*.

[31] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "A latent variable model approach to PMI-based word embeddings," 2015, *arXiv:1502.03520*.

[32] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–16.

[33] L. Hang, *Statistical Learning Methods*. Beijing, China: Tsinghua Univ., 2019.

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Cham, Switzerland: Springer, 2009.

HAO LIU was born in Chifeng, Inner Mongolia, China, in 1985. He received the B.S. degree from Beijing Normal University, in 2008, the M.S. degree from Kent University, U.K., in 2010, and the Ph.D. degree in statistics from Beijing Normal University, in January 2017. He has been working with the Collaborative Innovation Center for Basic Education Quality Monitoring, Beijing Normal University, China, as a Lecturer and a Master's Supervisor. His research interests include basic education quality impact factors, educational data mining, and other areas.

XI CHEN was born in Linxia, Gansu, China, in 1997. He received the B.S. degree from the South China University of Technology, in 2016. He is currently pursuing the master's degree with the Beijing Normal University, Beijing, China. His current research interests include machine learning, sentiment analysis, and model interpretability.

XIAOXIAO LIU was born in Linyi, Shandong, China, in 1995. She received the B.S. degree in management from the University of Jinan, in 2018, and the master's degree from Beijing Normal University, in 2021. She is currently pursuing the master's degree in educational psychology with the University of Alberta, Edmonton, Canada. Her research interests include data mining, machine learning, and natural language processing.