Hindawi
Mobile Information Systems
Volume 2022, Article ID 1909312, 9 pages
https://doi.org/10.1155/2022/1909312

Research Article

Microblog Emotion Analysis Method Using Deep Learning in Spark Big Data Environment

Junya Yan and Xiaohui Ma

School of Information Engineering, Shanxi Vocational University of Engineering Science and Technology, Shanxi, Jinzhong 030619, China

Correspondence should be addressed to Junya Yan; yanjunya@sxgkd.edu.cn

Received 7 May 2022; Revised 13 June 2022; Accepted 19 July 2022; Published 24 August 2022

Academic Editor: Shadi Aljawarneh

Aiming at the problem that the existing methods in the big data environment cannot extract the emotional features of microblog sufficiently and the average accuracy of analysis results is low, a microblog emotion analysis method using deep learning in spark big data environment is proposed. First, the Jieba word segmentation method is used to process text comments, so as to reduce the interference of irregular grammar and nonstandard words on the emotion analysis task of microblog text. Then, features based on affective rules, unary word features, syntactic features, and dependent word collocation features are selected. In order to prevent the dimension disaster caused by excessive feature dimensions, the feature selection method of information gain is used to reduce the dimension of features. Finally, a microblog emotion analysis method based on deep belief network (DBN) is established, and the DBN is parallelized through spark cluster to shorten the training time. Experiments show that when the feature set is composed of TOP2000 features, the classification accuracy of the fusion of four features is 90.94%, which is higher than that of the comparison method. In addition, the training time of DBN algorithm parallelized by spark cluster is only 27.78% of that of single machine. Therefore, compared with the comparison method, the proposed method can significantly improve the performance of the microblog emotion analysis system.

1. Introduction

The development of network technology makes users communicate more and more frequently online, including blogs, forums, and e-commerce website comments [1]. Users express their feelings about certain events or things by publishing information. Analyzing the words in social networks can help the government and other management institutions understand the social mood fluctuations, conduct public opinion analysis, further judge the development of the situation, give reasonable guidance, and maintain social stability [2–5]. From a commercial perspective, with the rise and popularity of e-commerce platforms such as Taobao and Amazon, users can give product evaluation after purchase, making the information of purchasing products more transparent. The quality of product comments will greatly affect users' purchase desire. Therefore, businesses can analyze users' comments, improve goods or change sales strategies in time, further analyze users' consumption characteristics and hobbies, draw user portraits, and make decisions to maximize business profits. In addition, emotion analysis can also be used to predict the stock market, election support, and other fields [6–9]. It can be seen that emotional analysis has important value in the fields of society, business, politics, and management [10].

With the rapid development of domestic microblog, many netizens participate in the discussion of various events, from personal trivia to enterprise marketing, and then to global major events. Microblog has become a social platform for public opinion release. Through the analysis of user emotion in microblog, it is of far-reaching significance for the development of government, enterprises, and individuals [11–15].

According to the granularity of research, emotion analysis tasks can be divided into three categories: document level, sentence level, and aspect level [16]. Document level emotion analysis regards the whole document as a basic unit and believes that a document as a whole only expresses one polar emotion. However, the document contains multiple sentences, and different sentences may have different emotional polarity classifications [17–19]. Sentence level emotion analysis is more fine-grained than document level, which is used to classify the emotional polarity of a single sentence. Aspect level emotion analysis is different from document level and sentence level affective analysis. It will more finely consider the emotion polarity and the target of corresponding emotion. The target here is attribute words or aspects, which usually exist in the form of entity or entity characteristics [20].

Aiming at the problem that the existing methods in the big data environment cannot extract the emotional features of microblog sufficiently and the average accuracy of analysis results is low, a microblog emotion analysis method using deep learning in spark big data environment is proposed. The main innovations are as follows:

(1) Jieba word segmentation method is used to process text comments, which effectively reduces the interference of irregular grammar and nonstandard words on the emotion analysis task of microblog text

(2) The feature dimension reduction operation is carried out by using the feature selection method of information gain to prevent the dimension disaster caused by too large feature dimension

(3) A microblog emotion analysis method based on DBN is established, and the DBN is parallelized through spark cluster, which effectively shortens the training time of the model

The rest of the sections are arranged as follows: Section 1 is related work, which introduces the current research status of emotion analysis. In Section 2, the structure and principle of deep confidence network are described. Section 3 describes deep belief network. In Section 4, the proposed DBN microblog emotion classification model based on spark parallel optimization is introduced in detail. Section 5 is the experiment. Section 6 summarizes this study.

2. Related Works

Deep learning method can better capture the grammatical and semantic features of text, which is a research focus of emotion analysis. Jebbara et al. used the bidirectional gated recurrent unit (GRU) to extract attribute words and specific aspects of emotion and extract features from the text for prediction of sentence labels [21]. Considering the characteristics of part of speech and corpus, Liu et al. proposed a method to complete the task of attribute word extraction by using RNN, which achieved better performance than the traditional system based on conditional random field [22]. In order to overcome the limitation of fixed window size of convolutional neural network (CNN) model and better capture context information, Chen et al. combined with the named entity recognition (NER) task method, proposed a text emotion analysis method based on BiLSTM-CRF model to classify BIO labels of entities in sentences [23]. Yin et al. proposed a long short-term memory (LSTM) model for cross-domain attribute word extraction, which combined the rule-based method to generate the auxiliary label sequence of each sentence [24]. Li et al. incorporated attention into the task of attribute word extraction and aspect category recognition and constructed a truncated historical attention and selective conversion network on LSTM [25]. Wang et al. proposed a GRU-based coupled multilayer attention (CMLA) model to extract attribute words and opinion words [26]. In the learning process, it encoded and decoded the dual propagation of attribute words and opinion words, not just limited to syntactic relations. Zhang et al. proposed a text emotion classification model integrating content features and user features [27]. Jamal et al. proposed a Twitter emotion analysis framework based on the Internet of Things, which used the mixed model of term frequency inverse document frequency (TFIDF) and deep learning model for emotion analysis, filtered the original tweets with the tokenization method, so as to capture useful features without noise information, and used TFIDF statistical technology to estimate the importance of local and global features. The adaptive comprehensive class balance technology is used to solve the class balance problem between different emotions [28]. Jelodar et al. used the LSTM method to classify the comments of COVID-19. The research results have a certain impact on the guidance and decision-making of COVID-19-related issues [29]. Wei et al. proposed a BiLSTM model based on multipolarity orthogonal attention for implicit sentiment analysis. Compared with the traditional single attention mechanism model, this method can effectively identify the differences between words and emotional tendencies and has been verified in experiments [30].

3. Deep Belief Network

3.1. DBN Model Structure. DBN is a neural network model with multiple hidden layers. It is difficult to optimize the weight in deep structures such as deep confidence network, so a greedy unsupervised training method is proposed to solve this problem. Figure 1 shows a structure diagram of a deep confidence network with three hidden layers h_1, h_2, and h_3. x is the input data and y is the output label corresponding to the input data. In the first step, DBN pairs each two adjacent neural network layers, trains the parameters between the two layers with the parameters of the input layer, and constructs the output layer. Moreover, the propagation of input layer and hidden layer is bidirectional, which is divided into forward process and backward process to learn data distribution. This method of building networks between layers is realized by the restricted Boltzmann machine (RBM) model. RBM is a recurrent neural network with two layers. Each node in the same layer is not connected to each other, and the output and input layer nodes are connected symmetrically without direction, which is equivalent to the connection of an undirected graph. An

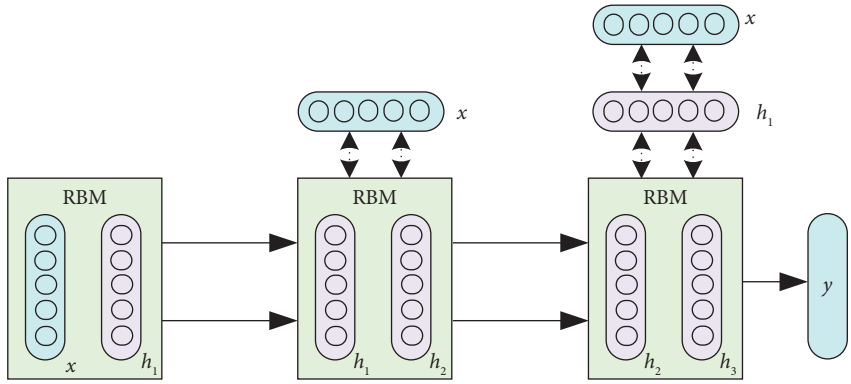

Figure 1: Deep belief network structure.

RBM consists of a hidden layer composed of random hidden units and a visible layer composed of random visible units.

Due to the special structure of RBM model, which has connection between layers and no connection within layers, it has the following important properties: when the visible unit state is given, the jth neuron in the hidden layer is calculated according to the neuron state of the visible layer, and the activation probability is as follows:

$$p\left(h_j = 1|\mathbf{v}\right) = \sigma\left(b_j + \sum_{i=1}^{m} w_{ij} v_i\right), \tag{1}$$

where σ is the sigmoid activation function, v_i represents the ith visible unit, h_j represents the jth hidden unit, w_{ij} is the weight between the ith visible unit and the jth hidden unit, and b_j is the offset threshold of the jth hidden unit.

Similarly, when the state of the hidden unit is given, the probability of the binary state v_i being 1 can be calculated, that is, the activation probability of the visible unit can be expressed as

$$p\left(v_i = 1|\mathbf{h}\right) = \sigma\left(a_i + \sum_{j=1}^{n} w_{ij} h_j\right), \tag{2}$$

where a_i is the offset threshold of the ith visible unit.

For the determination of the deep belief network model, the first thing is to know the number of nodes in the visible layer and the hidden layer. The number of nodes in the visible layer is the input data dimension. Second, the number of nodes in the hidden layer is related to the number of nodes in the visible layer in some research fields, such as processing image data with convolution restricted Boltzmann machine, which is not analyzed here. However, in most cases, the number of hidden layer nodes needs to be determined according to the use, or the number of hidden layer nodes that minimize the energy of the model under certain parameters.

3.2. DBN Model Training. The training of DBN model is divided into two parts: unsupervised pretraining process based on RBM and supervised parameter adjustment process.

The unsupervised pretraining process of DBN model adopts the layer-by-layer greedy learning strategy. The initial input layer is the visible layer, and the input data are the text feature vector. The data vector of the visible layer v combined with the weight w_1 is used to infer the data vector of the hidden layer h_1, which is the training process of RBM1. Then, the data vector of the hidden layer h_1 is combined with the weight w_2 to infer the data vector of the hidden layer h_2, which is the training process of RBM2, and so on. That is, multiple RBMs are stacked, the output of the previous RBM is the input of the next RBM, and the hidden layer of the previous RBM is the visible layer of the next RBM. By step-by-step training to the last layer, the pretraining process of DBN is completed. The specific steps are as follows:

Step 1. Randomly initialize the weight (W, a, b), in which W is the weight vector matrix, $a = [a_1, a_2, \ldots, a_n]$ is the offset coefficients of visible layer, and $b = [b_1, b_2, \ldots, b_n]$ is the offset coefficients of hidden layer. $v = [v_1, v_2, \ldots, v_n]$ is visible neurons, number is n; $h = [h_1, h_2, \ldots, h_n]$ is hidden neurons, number is m.

Step 2. Assign X value to the visible layer $v^{(0)}$ and calculate the probability that the hidden layer neurons can be activated:

$$p\left(h_j^{(0)} = 1|v^{(0)}\right) = \sigma\left(W_j * v^{(0)} + b_j\right). \tag{3}$$

Step 3. Perform a Gibbs sampling to obtain the value of each neuron in the hidden layer:

$$h^{(0)} \sim p\left(h^{(0)}|v^{(0)}\right). \tag{4}$$

Step 4. Reconstruct the visible layer v with the obtained $h^{(0)}$ in formula (4) and calculate the probability density:

$$p\left(v_i^{(0)} = 1|h^{(0)}\right) = \sigma\left(W_i * h^{(0)} + a_i\right). \tag{5}$$

Step 5. Perform Gibbs sampling again and reconstruct the value of each neuron in the visible layer. Let $r_i \in random[0, 1]$:

$$v_i = \begin{cases} 1, & p\left(v_i^{(0)} = 1|h^{(0)}\right) > r_i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Step 6. Calculate the activation probability of hidden layer neurons again with the reconstructed visible layer neurons:

$$p\left(h_j^{(1)} = 1|v^{(1)}\right) = \sigma\left(W_j * v^{(1)} + b_j\right), \tag{7}$$

where $\sigma(\cdot)$ adopts sigmoid activation function, and its function image is shown in Figure 2. Sigmoid is used to activate the function because its definition field is R and its value field is (0, 1). Therefore, no matter what range the input data of neurons in the visible layer is, the activation probability of nodes can be obtained by sigmoid function.

Step 7. Obtain the new weight vector matrix W, visible layer offset coefficient a, and hidden layer offset coefficient b:

$$\begin{aligned} a &= a + \varepsilon\left[v^{(0)} - v^{(1)}\right], \\ b &= b + \varepsilon\left[p\left(h^{(0)} = 1|v^{(0)}\right) - p\left(h^{(1)} = 1|v^{(1)}\right)\right], \\ W &= W + \varepsilon\left[p\left(h^{(0)} = 1|v^{(0)}\right)v^{(0)T} - p\left(h^{(1)} = 1|v^{(1)}\right)v^{(1)T}\right], \end{aligned} \tag{8}$$

where ε is the learning rate.

To sum up, pretraining only needs to iteratively calculate RBM1, RBM2, and RBM3 parameters in turn and finally get the best weight (W, a, b).

The supervised parameter optimization training of DBN model first uses the forward propagation algorithm to determine whether the hidden layer neurons are activated by using the parameters W and b obtained in the pretraining. Let l be the number of layers of the neural network and calculate the excitation value of each hidden layer neuron:

$$h^{(l)} = W^{(l)} * v + b^{(l)}. \tag{9}$$

Then, we propagate upward layer by layer, calculate the excitation values of neurons in all hidden layers using formula (9), standardize them with activation function, and finally calculate the excitation value $h^{(l)}$ and output vector $\widehat{X}$ of output layer:

$$\begin{aligned} h^{(l)} &= W^{(l)} * h^{(l-1)} + b^{(l)}, \\ \widehat{X} &= f\left(h^{(l)}\right). \end{aligned} \tag{10}$$

Then, the back propagation algorithm is used to update the parameters of the whole DBN network. The back propagation algorithm adopts the reconstruction error criterion, and the cost function is as follows:

$$E = \frac{1}{N}\left(\widehat{X}_l\left(W^{(l)}, b^{(l)}\right) - X_i\right)^2, \tag{11}$$

where E is the reconstruction error, $\widehat{X}_l$ is the actual output of the output layer, X_i is the theoretical output of the output layer, and $(W^{(l)}, b^l)$ represents the weight and offset coefficient of the layer l. The reconstruction error can reflect the likelihood of the training data to a certain extent. Finally, the gradient descent (GD) algorithm is used to update the weight and offset coefficient of the whole DBN network:

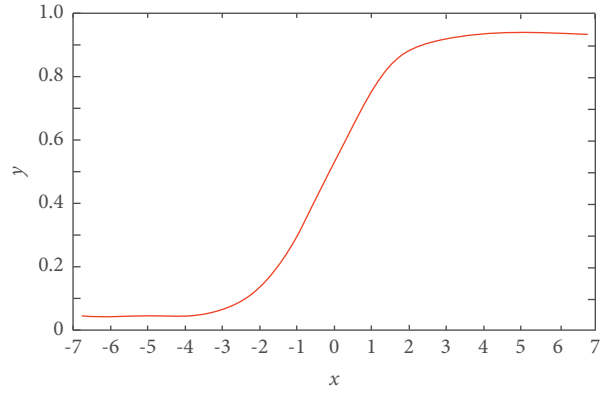

Figure 2: Sigmoid function image.

$$\left(W^{(l)}, b^{(l)}\right) = \left(W^{(l)}, b^{(l)}\right) - \varepsilon\frac{\partial E}{\partial\left(W^{(l)}, b^{(l)}\right)}. \tag{12}$$

To sum up, the training purpose of DBN model is to maximize the fitting of input data, and the output result is the reconstruction of training data. The visible layer neurons transfer their own features to the hidden layer neurons. The hidden layer neurons capture the higher-level features shown by the visible layer neurons through iterative training, so as to enhance the ability of feature extraction of the model.

4. DBN Microblog Emotion Classification Model Based on Spark Parallel Optimization

Figure 3 shows the work flowchart of microblog emotion analysis of the proposed method. Before classifying microblog emotion, it must be processed into a form that can be calculated by computer, that is, the representation model of data. Then, an emotional dictionary is built, the emotional features are extracted in the microblog text, the extracted features are taken as input, the whole spark parallel DBN model is trained, the classification results are obtained, and the emotional analysis of the microblog text is realized.

4.1. Microblog Preprocessing and Feature Vector Construction

4.1.1. Preprocessing. Text preprocessing is an indispensable part of the task of text emotion analysis. In text comments, due to the great differences in everyone's emotional thinking and speaking methods, it is often filled with strong personal emotional styles. All kinds of irregular grammar and nonstandard words will interfere with the task of text emotion analysis, so text preprocessing is very important. The text preprocessing part of this study includes as follows: filtering out repeated corpus, filtering out irregular words, removing stop words, emoticon processing, and Chinese word segmentation. The Chinese word segmentation part selects Jieba word segmentation. Jieba word segmentation can collect the dictionary established by users, and its Chinese word segmentation effect is good, which can well meet the needs of this study.

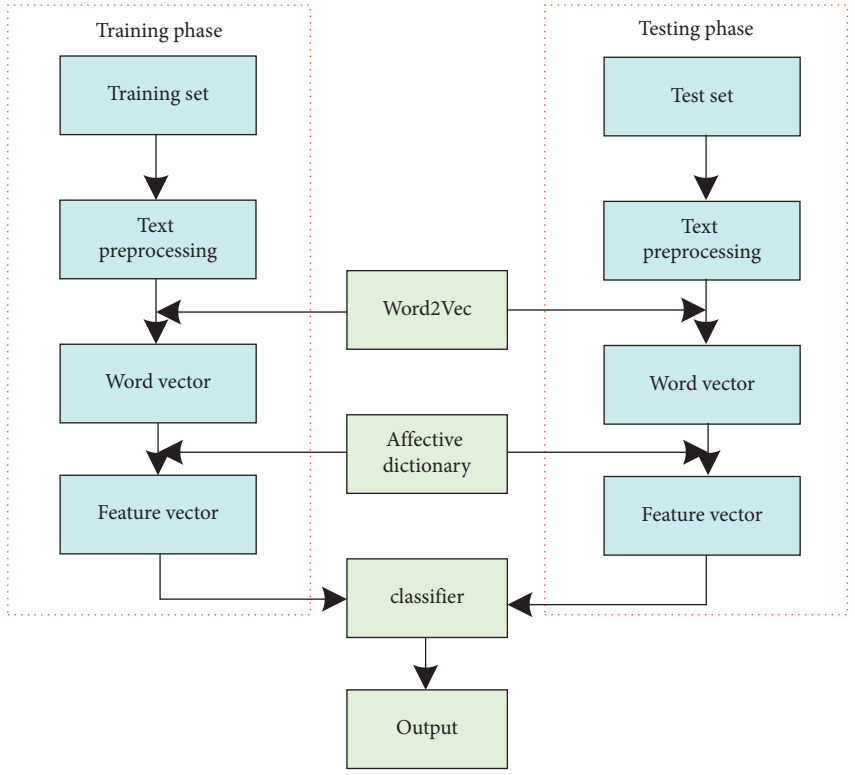

Figure 3: Flowchart of microblog emotion analysis.

4.1.2. Feature Construction. Text feature selection is a key step of machine learning, which determines the accuracy of emotion classification. This study selects four categories of features: features based on emotional rules, unigram features, syntactic features, and dependent word collocation features. The rule feature based on emotion is the feature obtained by extracting its effective information after improving the new rule method on the basis of predecessors. Considering that phrase structure can reduce sentence ambiguity, we add bigram and its combined part of speech tagging as features to the feature set. Dependency feature is the dependency identifier obtained from the dependency parsing tree. It plays an important role in the annotation of emotional category information and can save the information directly related to emotional words and other hidden information.

The method based on emotion dictionary plays an important role in the development history of text emotion analysis. Its core idea is to superimpose the polarity of emotion words and judge the emotional tendency of the text by numerical value. The formula of the classical method is as follows:

$$S = \sum_{i=1}^{n} Sw_i. \tag{13}$$

In the above formula, the parameter Sw_i represents the polarity of emotional word i. The parameter n represents the number of emotional words in the text. The method based on emotional dictionary can barely complete the task in some simple text tests, but considering the complex text grammar and the existence of various language structures in real use, the actual use is limited. Therefore, considering the defects of classical methods, a new emotion rule method is proposed. Considering that the length of the comment text is generally short and is basically a separated sentence, the method takes each clause as a meta unit. On the basis of considering the negative words, connectives, and other grammatical structures, the emotion calculation formula (equation (14)) is proposed to calculate the emotion tendency of each unit. The final text emotion tendency is judged by the value obtained by the superposition of the score values of each unit. If the score value is positive, the text emotion is classified as positive; if the score is negative, the text emotion is classified as negative:

$$S\,\text{unit} = \sum_{i=n}^{n} \left(K * Pw_i * \prod_{j=1}^{m} \text{mod}_j \right), \tag{14}$$

where the parameter n represents the number of emotional words in the text, the parameter Pw_i represents the emotional extremum of emotional word i, the parameter m represents the number of words modifying emotion word i, the parameter mod_j represents the weight of the corresponding modifier, and the parameter k represents the weakening or strengthening coefficient of rules. This parameter exists to solve a problem often ignored in emotion analysis tasks—the deviation of emotion analysis results caused by subject confusion.

Table 1 lists a brief description of the emotional rules designed by the proposed method. Generally speaking, the more complete the emotional rules are, the better the effect of the emotional rule method is. After combining the

Table 1: Brief emotion rules table.

Rule name	Rule content
Negative emotion rule	When negative affective words are detected, the affective polarity and intensity are reversed.
Expression emotion rule	When an expression word is detected, the corresponding score is given directly according to the expression dictionary.
Emotional rules of turning conjunctions	When the preinflection word is detected, it is weakened according to the strength value of the conjunction dictionary.
Borrowed emotion rules	Used to strengthen/weaken demonstrative pronouns.

emotional rules, the final score is calculated according to formula (14), and then, three parameters are extracted as emotional features: the score of emotional words, the number of positive/negative emotional words, and the ratio of strengthening/weakening times of rules.

For the other three emotional features, "the scenic spot service is really good, I like it very much!" is taken as an example sentence to show the feature extraction process and the corpus is input into Jieba word segmentation to get "scenic spot /*n* service /*n* really /ad good /*a* , /*wd* I /rr very /d like /*vi* ! /*wd*", where /*n* stands for noun, /ad stands for adverbial word, /*a* stands for adjective, /*wd* stands for punctuation mark, /rr stands for pronoun, /*d* stands for adverb, and /*vi* stands for verb.

Based on the above results of word segmentation and tagging, the syntactic features can be obtained: scenic spot service, service really, really good, good I, I very, like it very much, *n*, ad, *a*, rr, *d*, *vi*. The number of features is 12. After the result of word segmentation is obtained, the dependency and word collocation features of the input example sentences can be obtained by calling the StanfordNlp natural language processing toolkit. The specific relationship and collocation are listed in Table 2.

In practical use, in order to avoid various problems caused by excessive feature dimension, the feature selection method of information gain (IG) is adopted for feature dimension reduction. The formula is as follows:

$$IG(T) = -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i) + P(t)\sum_{i=1}^{n} P(C_i|t)\log_2 P(C_i|t), + P(\bar{t})\sum_{i=1}^{n} P(C_i|\bar{t})\log_2 P(C_i|\bar{t}), \tag{15}$$

where the parameter $P(C_i)$ is the probability of category C_i, the parameter $P(t)$ is the probability of feature t, the parameter $P(C_i|t)$ is the probability of simultaneous occurrence of feature t and category C_i, and the parameter $PP(C_i|\bar{t})$ is the probability that the category C_i appears when the feature t does not appear. The score of the feature is calculated according to the formula, and the feature of TOP N is selected according to the score, so as to select and reduce the dimension of the feature.

4.2. Parallel Optimization of Emotion Classifier Based on Spark Platform. The master node provides initialization parameters $\theta = \{W, b, c\}$ for training and distributes them to each worker node. Each worker node uses the training data on all split slices for parameter learning and uses minibatch as the criterion for training parameter update. When the worker node completes the training data of a batch, the generated parameter change $\Delta\theta$ is sent to the master management node for parameter update until all training is completed, and the feature data processed in each training are converted into RDD form for storage. The specific algorithm is shown in Algorithm 1.

Table 2: Dependency of example sentences.

Dependency	
Assmod (service-2, scenic spot-1)	Advmod (good-4,-5)
Punct (good-4, service-2)	Advmod (like-8,I-6)
Nsubj (good-4, good-3)	Root (like-8, very-7)
Nsubj (root-0, good-4)	Conj (good-4, like-8)

The parallelization structure of DBN network based on spark platform is shown in Figure 4.

5. Experiment and Analysis

5.1. Experimental Data and Evaluation Indices. The dataset of this experiment comes from COAE2015 Task 3. There are 133201 microblog sentences, including a large number of interfering sentences. Datasets are divided into four different areas to evaluate, including books (BOO), audio products (DVD), electronic products (ELE), and kitchenware (Kit). Each dataset contains 2000 positive and 2000 negative comments.

In this study, the accuracy is used as the evaluation index of the experiment, and the calculation formula is as follows:

$$\text{accuracy} = \frac{\text{Num}(\text{correct})}{\text{Num}(\text{all})}, \tag{16}$$

where Num (correct) is the number of samples correctly predicted by emotion classification and Num (all) is the total number of samples in the test corpus.

5.2. Relationship between Iteration Times and Prediction Accuracy. The advantage of deep neural network over shallow neural network is that it can iteratively learn, extract features, and constantly modify the model, but too high or too low iteration times will affect the overall performance. In a task, if the number of iterations is lower than a certain value, it will lead to incomplete learning of features and imperfect release of performance. If the number of iterations is higher than a certain value, it will take a too long time and be inefficient. Therefore, the selection of iteration times is very important in the task. In the experiment, with Ft1 as the

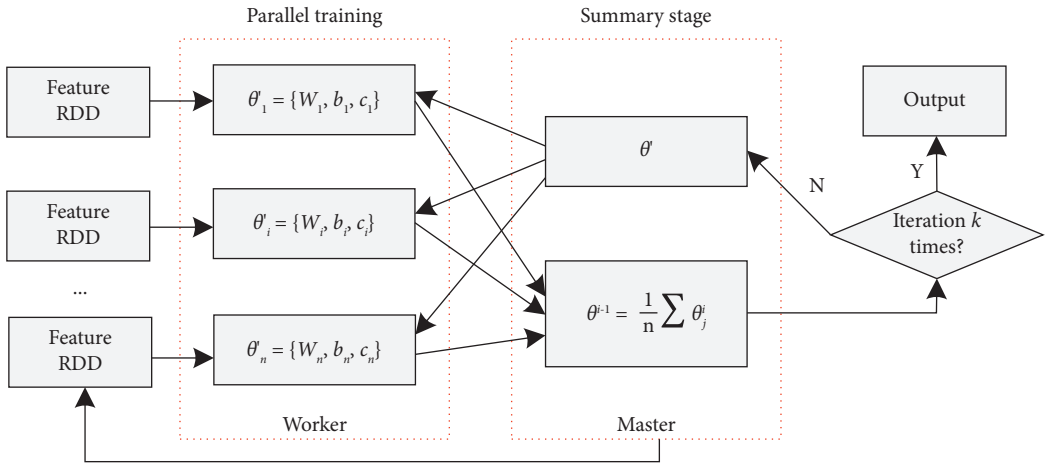

Figure 4: Parallelization structure of DBN network based on spark platform.

feature, the relationship between prediction accuracy and iteration times is shown in Figure 5. It can be seen from the figure that when the number of iterations is less than 60, the recognition rate increases significantly with the increase of the number of iterations. When the number of iterations is 65, the change range of accuracy is small and almost reaches a balanced state. Based on the above analysis, for the number of iterations, 65 iterations are selected to ensure the stability of the results.

5.3. Experimental Results and Analysis of Emotion Classification under Different Methods. Table 3 lists the experimental results of the text emotion classification method based on deep belief network designed in this study. In the network, the input is the vector composed of 1000-, 2000-, and 4000-dimensional features with the top information gain. The text abstract features are learned through hidden layer nonlinear mapping. The specific results are as follows: for the 1000-dimensional feature set, the training iteration of restricted Boltzmann machine is 100 times, and the node parameter corresponding to the network structure "input layer-hidden layer-output layer" is "1000-300-100." For the 2000-dimensional feature set, the training iteration of restricted Boltzmann machine is 100 times, and the node parameters corresponding to the network structure are "2000-600-300." For the 4000-dimensional feature set, the training iteration of restricted Boltzmann machine is 100 times per layer, and the node parameters corresponding to the network structure are "4000-600-300." It can be seen from Table 3 that the method based on depth belief network achieves the best classification accuracy of 90.94 when the structure is 2000-600-300 and the four features are combined.

In order to verify the learning and expression ability of the method in this study, the same features are used to compare the methods in reference [27], reference [28], and the proposed method. The recognition rates of reference [27] and reference [28] are 87.11% and 87.69%, respectively. When the structure of the proposed method is 2000-600-300,

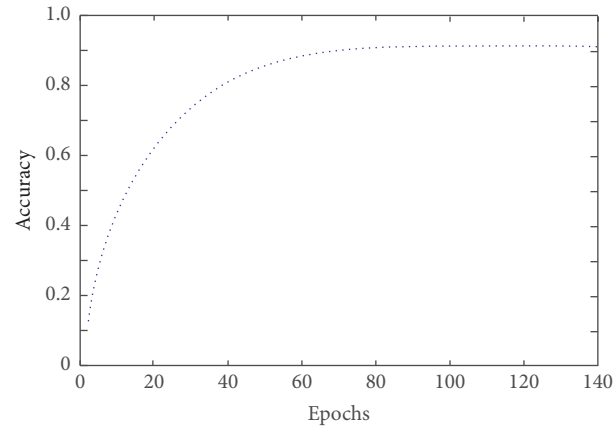

Figure 5: Relationship between iteration times and prediction accuracy.

the combination of four features achieves the best classification accuracy of 90.94%. Moreover, it can be found that the overall accuracy of the proposed method is higher than that of the methods in reference [27] and reference [28], because the proposed method will obtain more emotional knowledge than the comparison methods in the learning of features, so as to obtain better performance, as listed in Table 4.

5.4. Microblog Emotion Analysis Results under Spark Platform. The DBN network is optimized in parallel under the spark platform. The spark cluster used in the experiment is composed of 10 servers. One server is used as the management node of spark cluster, and the other nine servers are used as the computing nodes of spark cluster. The hardware configuration is CPUXeonE5520, 20 GB memory, and 1 TB hard disk. In Figure 6, the abscissa represents the size of the training data and the ordinate represents the time-consuming. It can be seen from the figure that when the amount of data increases to 60000, the spark training time is only 27.78% of the single machine training time. The Jieba word segmentation method is used to reduce the interference of irregular grammar and nonstandard words on the emotion

Input: Training data set S, set S as the feature vector set after microblog preprocessing
Output: Emotion classification result set
Determine the number of iterations K and the parameter θ^0 for initializing RBM
For $i = 0$ **to** K **do**
The Master node broadcasts θ^i to each Worker node;
The Worker node uses the data on Split to train the parameters of RBM network;
All Worker nodes send $\Delta\theta$ to the Master node;
The Master node calculates $\theta^{i+1} = 1/n \sum \theta^i_j$. The feedback mechanism of BP network is used to adjust and fine-tune the DBN network model.
End

Algorithm 1: Spark parallelized DBN network.

Table 3: Results based on the deep belief network method.

Method	Feature structure	Ft_1	$Ft_1 + Ft_2$	$Ft_1 + Ft_2 + Ft_3$	$Ft_1 + Ft_2 + Ft_3 + Ft_4$
Proposed method	1000-300-100	89.42	89.59	89.67	89.89
	2000-600-300	90.21	90.29	90.24	90.94
	4000-600-300	90.14	90.21	90.22	90.87

Table 4: Comparison of experimental results of different methods.

Method	Accuracy
Reference [27]	87.11
Reference [28]	87.69
Proposed method	90.94

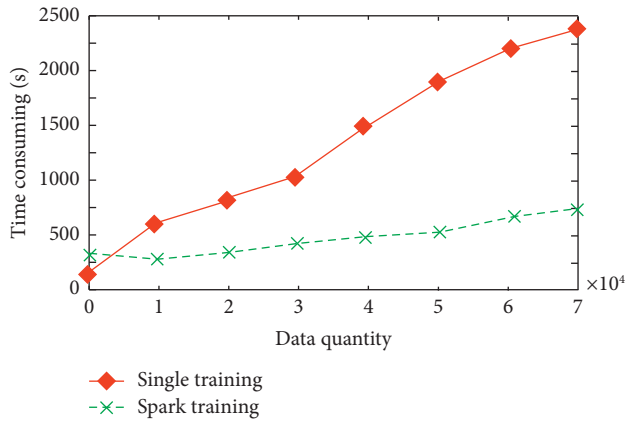

Figure 6: Comparison of training time changes when data increase between a single node and spark parallel computing.

analysis task of microblog text. The feature dimensionality reduction operation is carried out by using the feature selection method of information gain to avoid the problem of dimension disaster. It can be concluded that the parallel DBN algorithm based on spark platform can effectively improve the operation efficiency when processing massive data.

6. Conclusion

Aiming at the problem that the existing methods in the big data environment do not extract the emotional features of microblog sufficiently and the average accuracy of the results is low, a microblog emotion analysis method using deep learning in the spark big data environment is proposed. The DBN is parallelized through spark cluster, which greatly shortens the training time. Experimental results show that the proposed algorithm has good microblog emotion analysis ability.

In this study, the factors considered in the study of data parallel fragmentation strategy are not comprehensive enough. More data fragmentation strategies should be tried in the future. In the follow-up, other parallel optimization algorithms can be used for reference to improve the parallel speedup ratio of the algorithm. Moreover, in addition to word vector representation, researchers have developed new representation methods in recent years, such as Atlas and tree database, to represent text information. Therefore, the text emotion classification algorithm proposed in this study can be further improved. How to embed more and more effective text semantic information is still the focus of the next step.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the 2020 Horizontal Project (no. HX2020029).

References

[1] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech-A review," *Toward Robotic Socially Believable Behaving Systems-Volume I*, vol. 2, no. 14, pp. 205–238, 2016.

[2] A. M. Mohsen, A. M. Idrees, and H. A. Hassan, "Emotion analysis for opinion mining from text: a comparative study," *International Journal of e-Collaboration*, vol. 15, no. 1, pp. 38–58, 2019.

[3] D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W Shi, "Deep learning based emotion analysis of microblog texts," *Information Fusion*, vol. 64, no. 7, pp. 1–11, 2020.

[4] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, no. 3, pp. 309–317, 2019.

[5] F. Xia and Z. Zhang, "Study of text emotion analysis based on deep learning," in *Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 2716–2720, IEEE, Wuhan, China, May 2018.

[6] L. Cao, S. Peng, and P. Yin, "A survey of emotion analysis in text based on deep learning," in *Proceedings of the 2020 IEEE 8th International Conference on Smart City and Informatization (iSCI)*, pp. 81–88, IEEE, Guangzhou, China, February2020.

[7] L. Ran, L. Zheng, and L. Hailun, "Text emotion analysis: a survey," *Journal of Computer Research and Development*, vol. 55, no. 1, pp. 30–39, 2018.

[8] S. Peng, L. Cao, Y. Zhou et al., "A survey on deep learning for textual emotion analysis in social networks," *Digital Communications and Networks*, vol. 23, no. 5, pp. 12–21, 2021.

[9] X. Wang, L. Kou, V. Sugumaran, X. Luo, and H Zhang, "Emotion correlation mining through deep learning models on natural language text," *IEEE Transactions on Cybernetics*, vol. 51, no. 9, pp. 4400–4413, 2021.

[10] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: a survey," *Review: Data Mining and Knowledge Discovery*, Wiley Interdisciplinary, vol. 8, no. 4, pp. 1253–1260, 2018.

[11] J. Choudrie, S. Patil, K. Kotecha, N. Matta, and I Pappas, "Applying and understanding an advanced, novel deep learning approach: a covid 19, text based, emotions analysis study," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1431–1465, 2021.

[12] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H Prendinger, "Deep learning for affective computing: text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, no. 7, pp. 24–35, 2018.

[13] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and S Khan, "Classification of poetry text into the emotional states using deep learning technique," *IEEE Access*, vol. 8, no. 6, pp. 73865–73878, 2020.

[14] E. A. H. Khalil, E. M. F. E. Houby, and H. K. Mohamed, "Deep learning for emotion analysis in Arabic tweets," *Journal of Big Data*, vol. 8, no. 1, pp. 136–148, 2021.

[15] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, no. 2, pp. 111866–111878, 2019.

[16] N. Majumder, S. Poria, A. Gelbukh, and E Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[17] U. Rashid, M. W. Iqbal, and M. A. Skiandar, "Emotion detection of contextual text using deep learning," *IEEE*, in *Proceedings of the 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–5, Istanbul, Turkey, November 2020.

[18] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, pp. 12189–12198, 2020.

[19] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," vol. 1, no. 9, pp. 23–31, 2018, https://arxiv.org/abs/1804.05788.

[20] Y. Kumar, D. Mahata, and S. Aggarwal, "Bhaav-a text corpus for emotion analysis from Hindi stories," vol. 10, no. 4, pp. 1634–1642, 2019, https://arxiv.org/abs/1910.04073.

[21] S. Jebbara and P. Cimiano, "Aspect-based sentiment analysis using a two-step neural network architecture," *Semantic Web Evaluation Challenge*, pp. 153–167, Springer, Cham, 2016.

[22] P. Liu, S. Joty, and H. Meng, "Fine-grained opinion mining with recurrent neural networks and word embeddings," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1433–1443, Lisbon, Portugal, September 2015.

[23] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems with Applications*, vol. 72, no. 3, pp. 221–230, 2017.

[24] W. Yin, K. Kann, and M. Yu, "Comparative study of CNN and RNN for natural language processing," vol. 6, no. 1, pp. 200–207, 2017, arXiv preprint arXiv.

[25] X. Li, L. Bing, and P. Li, "Aspect term extraction with history attention and selective transformation," *IJCAI*, vol. 3, no. 11, pp. 4194–4200, 2018.

[26] W. Wang, S. J. Pan, and D. Dahlmeier, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3316–3322, CA, USA, February 2017.

[27] C. Zhang, L. Xie, Y. Aizezi, and X Gu, "User multi-modal emotional intelligence analysis method based on deep learning in social network big data environment," *IEEE Access*, vol. 7, no. 2, pp. 181758–181766, 2019.

[28] N. Jamal, C. Xianqiao, F. Al-Turjman, and F. Ullah, "A deep learning–based approach for emotions classification in big corpus of imbalanced tweets," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–16, 2020.

[29] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733–2742, 2020.

[30] J. Wei, J. Liao, Z. Yang, S. Wang, and Q. Zhao, "BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis," *Neurocomputing*, vol. 383, pp. 165–173, 2020.