Sentiment Classification of Crowdsourcing Participants’ Reviews Text Based on LDA Topic Model

YANRONG HUANG1, RUI WANG2,3, BIN HUANG1, BO WEI4, SHU LI ZHENG1, AND MIN CHEN5

1 College of Economics and Management, Zhejiang University of Water Resource and Electric Power, Hangzhou 310018, China
2 School of Economics and Management, Jiangxi University of Science and Technology, Ganzhou 341000, China
3 School of Economics, Fujian Agriculture and Forestry University, Fuzhou 35000, China
4 School of Informatics Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
5 State Key Laboratory of Software Engineering, School of Computer Science, Wuhan University, Wuhan 430072, China

Corresponding author: Yanrong Huang (hyanrong@whu.edu.cn)

This work was supported in part by the Scientific Research Projects of Zhejiang Education Department under Grant Y202045307, in part by the Jiangxi Social Science Fund Project under Grant 19GL25, in part by the National Natural Science Foundation of China under Grant 61806204, and in part by Humanities and Social Sciences of the Ministry of Education of China under Grant 19YJA630123.

ABSTRACT The review text received by crowdsourcing participants contains valuable knowledge, opinions, and preferences, which is an important basis for employers to make trading decisions, and crowdsourcing participants to improve service level and quality. However, there are two kinds of emotional polarity in the review text, the attention paid to sentiment classification of review text with fuzzy emotional boundaries is insufficient. This paper proposes a supervised text sentiment classification method with Latent Dirichlet Allocation (LDA) to improve the classification performance of review text with fuzzy sentiment boundaries. Taking the review text of crowdsourcing participants on the Zhubajie platform as the data set, using N-gram, Word2vec, and TF-IDF algorithms to extract text features. The LDA topic model is applied to expand the number of text features and extract eight topics that affect employers’ sentiment tendencies. Text classifiers are constructed based on Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) algorithms, and the effectiveness of the sentiment classification methods are verified by ten-fold cross-validation and confusion matrix. Experimental results show that using the LDA topic model to extend the features of review text can effectively alleviate the problem that the classifier is difficult to distinguish the sentiment categories of different emotion polarity words coexisting text, and enhance the ability of emotion boundary fuzzy text classification. Based on TF-IDF and LDA to extract and expand text features, the GBDT text sentiment classifier with the accuracy of 0.881; the F1-measure of the second, third, fourth, and fifth categories samples are 0.462, 0.571, 0.278, and 0.647 respectively, which is better than SVM, RF, and XGBoost classifiers and has the best classification performance.

INDEX TERMS Sentiment analysis, text classification, LDA topic model, crowdsourcing participants.

I. INTRODUCTION Crowdsourcing is an open innovation form, which can bring together talents from all fields to participate in technological innovation and value creation, stimulate the vitality of talent innovation and provide valuable results. With the development of the internet and the continuous update of computer and information technology, the transaction data of crowdsourcing platforms have been growing in geometric series, resulting in a large number of online customer review (OCR) information. The reviews text has reliable, rich, and useful emotions and information, including valuable knowledge, views, and preferences, reflecting the real transaction experience and feelings of traders [1].

Judging the emotional tendency of the review text received by crowdsourcing participants is crucial for employers, crowdsourcing participants, and crowdsourcing platforms. The review text contains the details of crowdsourcing activities, which can truly and comprehensively feedback the working ability and attitude of crowdsourcing participants.
It becomes an important information source for employers to understand the task completion of crowdsourcing participants and assist in trade decision-making. Crowdsourcing participants can understand the advantages and disadvantages of themselves and competitors through the review text, so as to improve their service level and quality [2]. The crowdsourcing platform can establish an effective reputation evaluation mechanism for crowdsourcing participants, optimize the task search and recommendation system of the crowdsourcing platform [3], [4], and improve the operation efficiency of the crowdsourcing platform.

However, in the review text of crowdsourcing participants, there are usually coexist two emotional polarities words, and insufficient attention has been paid to the classification of such emotion fuzzy texts. This paper intends to propose a supervised text emotion classification method based on Latent Dirichlet Allocation (LDA) to improve the classification performance of short texts with fuzzy emotion boundaries.

II. RELATED RESEARCH

At present, sentiment analysis of review text has become an important research field, which has attracted more and more attention in the academic domain. Text sentiment classification refers to the comprehensive use of machine learning technology and natural language processing technology to analyze, process, induce and infer the review text, analyze the emotion contained in the text, and judge the emotion, attitude, and viewpoint contained in the review text [5]. At present, text sentiment analysis is widely used in many fields, including social network users [2], [6], [7], product evaluation [8]–[10], etc. Generally speaking, text sentiment classification methods can be divided into fine-grained and coarse-grained.

The fine-grained analysis is based on the dictionary judgment feature view of emotion polarity and intensity [9], including two methods, the use of existing Chinese and English sentiment dictionary resources and self-built sentiment dictionary [9], [11], [12]. Lu et al. built a sentiment dictionary to quantify the sentiment tendency of crowdsourcing participants’ comments based on factors such as transaction price and transaction time [11]. Yang et al. established an online emotion dictionary for cigarettes, evaluated cigarettes in terms of packaging, taste, taste, and smoke, and calculated the sentiment scores of cigarettes brands of Yuxi and Furongwang [9]. The fine-grained sentiment analysis method highly depends on the completeness of the sentiment dictionary, which requires a lot of manpower and time to build a high-quality sentiment dictionary in the application, and its robustness is not ideal [6], [13].

The coarse-grained text sentiment analysis method has attracted much attention in recent years. It uses machine learning technology to classify the whole text sentiment, including three types: supervised [13], [14], unsupervised [15], and semi-supervised [16], among which the supervised learning method is the most commonly used. Wu et al. used Bayesian and Support Vector Machine (SVM) to analyze the emotion of hotel reviews; the results show that the classification effect of the SVM algorithm is better [2]. Yue et al. collected multi-domain product review data sets from e-commerce platforms Amazon and JD, and used attention Neural Network to conduct text sentiment classification research [10]. Yang et al. analyzed the sentiment of environmental public service microblog, and the XGBoost model had the highest prediction accuracy [17]. Neelakandan and Paulraj conducted sentiment analysis on the text of social network platform Twitter and found that the GBDT algorithm has the best classification effect [18].

To optimize the effect of text sentiment classification, researchers use improved algorithm or model fusion method to more accurately represent feature text. Zeng and Wang extracted the word level and sentence level features of microblog through Bi-LSTM to improve the accuracy of sentiment classification of public security event microblog [6]. Gao et al. used the N-gram algorithm to extract review text features and used a variety of ensemble learning methods for sentiment classification [19]. Chen et al. used a large amount of information related to knowledge graphs to increase the number of short text classification features and further designed a keyword semantic extension method based on knowledge graphs [20]. Sahu and Khandekar proposed using application computing language to preprocess the text, such as stem extraction, stop, and part of speech tagging, to solve the problems of negation, reinforcement, punctuation, and abbreviation in comments, and improve the accuracy of text sentiment classification [21].

As a mature technology, LDA topic model is valued by researchers. Gao et al. proposed a CO-LDA model based on the combination of word CO-occurrence analysis and Latent Dirichlet Allocation (LDA) [22]. Wu et al. constructed a short text emotion classification model based on LDA topic model and Long-short Term Memory (LSTM) to improve the accuracy of short text classification [23]. Ozyurt and Akcayo proposed SS-LDA from the product aspect of user reviews to solve the problem of text data sparsity [8].

At present, there are abundant research results in text sentiment analysis, but there are still some shortcomings. In the existing research on sentiment analysis of review text, whether the fine-grained text sentiment classification method is effective depends on whether a perfect text sentiment dictionary can be constructed. The coarse-grained text sentiment classification method can better make up for the defects of fine-grained method, but most of the default text words and sentences have a unified emotional polarity, and ignore the single review text may have two emotional polarities, lack of attention to different emotional polarity words coexist and emotional boundary fuzzy text. From the perspective of coarse granularity, natural language processing and machine learning technology are integrated to improve the classification accuracy by extracting and expanding the features of the review text, and to mine the key factors that affect employers’ emotional tendency in crowdsourcing activities. It is a new research field to accurately divide the sentiment categories of...
This paper proposes a sentiment classification method for crowdsourcing participants’ review text based on the LDA topic model. The steps of sentiment classification are shown in Figure 1.

**III. RESEARCH PROCEDURES AND METHODS**

**A. RESEARCH PROCEDURES**

This paper proposes a sentiment classification method for crowdsourcing participants’ review text based on the LDA topic model. The steps of sentiment classification are shown in Figure 1.

1. **DATA PREPROCESSING**
   - **Text Data Preprocessing**
     - De-duplication of review text
     - Stop word removal

2. **TEXT FEATURE EXTRACTION METHOD**
   - **Term Frequency-Inverse Document Frequency (TF-IDF)**
     - The TF-IDF algorithm is used to extract text features.

**FIGURE 1.** Sentiment classification steps of crowdsourcing participants based on LDA topic model.

First, raw data collection and preprocessing, this paper selects Zhubajie, the largest crowdsourcing platform in China, to collect the text of employers’ evaluations for crowdsourcing participants. Preprocess for review text, delete the text of the repetitive review, invalid irrelevant and fewer information reviews text. To segment Chinese words by using Jieba word segmentation tool, and to remove stop words by Harbin Institute of technology stop word list. Secondly, N-gram, Word2vec, and TF-IDF algorithms are used to extract text features. Thirdly, by calculating the perplexity degree of the LDA topic model, the optimal number of topics is determined, the review topic is divided, and the feature of review text is expanded. Fourthly, using machine learning technology, the supervised text sentiment classifiers based on Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost) are constructed; Finally, ten-fold cross-validation, confusion matrix, and statistical significance test are used to empirically analyze and evaluate the accuracy of the LDA topic model for emotion boundary fuzzy text classification and prediction, as well as the performance of different sentiment classification and prediction methods.

**B. RESEARCH METHODS**

1. **DATA PREPROCESSING**
   - Text data preprocessing includes de-duplication of review text, text segmentation, and de stop words. Word is the smallest semantic unit in text, and the accuracy of text segmentation directly affects the result of text sentiment classification. Chinese word segmentation includes two methods: Dictionary segmentation and statistical segmentation. Dictionary word segmentation is to match the text with the words in the dictionary. The completeness of the dictionary content affects the result of word segmentation. Statistical word segmentation is to divide the sentence into words, and then calculate the probability of the result to get the most probability.

   The statistical word segmentation method depends on the richness of the corpus. The commonly used word segmentation tools include the ICTCLAS Chinese word segmentation system of the Chinese Academy of Sciences, Fudan NLP word segmentation, Pangu word segmentation, Paoding word segmentation, HTTPCWS word segmentation, Jieba word segmentation, and so on. The common method to remove the stop words is the stop word list filtering method. By comparing the segmentation results with the stop word list, the stop words are removed, so as to reduce the complexity of subsequent calculation and improve the performance of classification prediction. This paper uses the statistical word segmentation method, uses Jieba for Chinese word segmentation, and uses the Harbin Institute of technology stop word list to remove stop words.

2. **TEXT FEATURE EXTRACTION METHOD**
   - **TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)**
     - Term Frequency-Inverse Document Frequency (TF-IDF) is a method of vectorization of word items, which can evaluate the importance of word items in documents by calculating their weight. Through TF-IDF, the word items in the review text are transformed into numerical data, which is convenient for the subsequent machine learning algorithm to construct the classification model. Term frequency (TF) reflects the number of times a word item appears in a document. The expression of word frequency of a word \( w \) in document \( d \) is as follows:

   \[
   TF(w, d) = \text{count}(w, d) \tag{1}
   \]
Inverse Document Frequency (IDF) is an important factor used in modern retrieval functions, which refer to the frequency of terms not appearing in many documents. \( M \) is the total number of documents in the document collection, \( df(\cdot) \) is the document frequency (i.e. the number of documents containing the word \( w \)). IDF can be defined as:

\[
IDF(w) = \log\left(\frac{M + 1}{df(w)}\right)
\]

For a specific term, the inverse document frequency can be introduced into the vector representation of the document by multiplying its IDF. The introduction of IDF can punish common words, which usually have a lower IDF, and reward words with information content usually have a higher IDF value. The calculation formula of TF-IDF is as follows:

\[
TF-IDF = TF(w, d) \times IDF(w)
\]

b: LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation (LDA) is a kind of document generation model, which can assign a probability to new documents and give the distribution of all possible documents. In the LDA topic model, the probability of documents is predicted by comparing the corresponding generation models of each category, and the documents are assigned to the category with the highest probability. Topic coverage distribution of each document is set as a prior distribution from Dirichlet, which defines the distribution of polynomial distribution in the whole parameter space, that is, a probability vector about the topic. Assuming that any term in each term distribution is unbiased and unbiased for any topic in each document, \( C \) is the whole set, and the Dirichlet distribution controlling topic coverage has \( k \) parameters \( \alpha_1, \alpha_2, \ldots, \alpha_k \). The Dirichlet distribution which controls the distribution of subject words has \( M \) parameters \( \beta_1, \beta_2, \ldots, \beta_k \). Each \( \alpha_i \) can be interpreted as a pseudo count of the subject \( \theta_i \), each \( \beta_i \) can be interpreted as a pseudo count of the corresponding term \( \omega_i \). LDA generation model component can be defined as a hybrid model of \( \theta_1, \theta_2, \ldots, \theta_k \) with \( k \) term distribution. The probability of the document \( \omega \) is observed, where the mixing coefficient is the topic coverage distribution \( \pi_{d,j} \) of document \( d \), the maximum likelihood estimation is used to obtain the parameters \( \alpha \) and \( \beta \) of LDA model.

\[
(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \log p(C|\alpha, \beta)
\]

After parameter estimation, in order to obtain the value of hidden variables in LDA, we need to use posterior inference to describe the \( k \) term distribution \( \{\theta_i\} \) of all topics in a collection and the topic coverage distribution \( \pi_{d,j} \) of each document. In other words, the Bayes rule is used to calculate \( p\{\theta_i\}, \{\pi_{d,j}\}|C, \alpha, \beta\):

\[
p\{\{\theta_i\}, \{\pi_{d,j}\}|C, \alpha, \beta\} = \frac{p(C|\{\theta_i\}, \{\pi_{d,j}\}) p\{\theta_i\} \{\pi_{d,j}\}|\alpha, \beta\}}{p(C|\alpha, \beta)}
\]

Equation (5) gives the posterior distribution of the possible values of variables, and further obtains the point estimation. LDA works by mapping documents into \( k \)-dimensional space through \( \pi_{d,j} \), which is an extension of probabilistic latent semantic analysis.

3) TEXT SENTIMENT CLASSIFICATION METHOD

(a) Gradient Boosting Decision Tree (GBDT) is a boosting algorithm proposed by Fridman (Fridman, 2001). GBDT takes CART as the base model, through the linear combination of the base classifiers to reduce the residual generated in the training process to achieve the data classification algorithm. A weak classifier is generated by iteration, and a weak classifier is generated in each iteration. The next iteration is trained on the basis of the classifier residual error of the previous round, and the classification results are obtained through multiple iterations. Weak classifiers generally require low variance and high deviation. In the process of training, the accuracy of weak classifiers is improved by continuously reducing the bias of weak classifiers. The goal of GBDT algorithm is to make the loss function decrease along the gradient direction and reduce the loss function as quickly as possible.

The training sample is \( T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \) the strong learner obtained by the previous iteration is \( f_{t-1}(x) \), the loss function is \( L(y, f_{t-1}(x)) \), the weak classifier obtained by iteration is \( h_t(x) \), minimize the round loss function \( L(y, f_t(x)) = L(y, f_{t-1}(x) + h_t(x)) \), to reduce sample loss. The GBDT algorithm uses the negative gradient of the loss function to fit the approximate value of the current round loss. The negative gradient of the loss function of the \( i \)-th sample in round \( t \) can be expressed as follows:

\[
r_{ti} = -\left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]_f(x)=f_{t-1}(x)
\]

Using \( r_{ti} \) fitted cart regression tree to get the \( t \) regression tree, and its corresponding leaf node region \( R_j \), where \( J \) is the number of leaf nodes, \( j = 1, 2, \ldots, J \). The best output value \( c_j \) of leaf node fitting is as follows:

\[
c_j = \arg \min_{c} \sum_{i \in R_j} L(y_i, f_{t-1}(x_i)) + c
\]

The fitting function of the decision tree is obtained as follows:

\[
h_t(x) = \sum_{j=1}^{J} c_j I(x \in R_j)
\]

Finally, the expression of this round of strong learning is obtained as follows:

\[
f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_j I(x \in R_j)
\]

Assuming that the number of classes is \( K \), the log likelihood loss function can be
expressed as follows:

\[ L(y,f_{t-1}(x)) = - \sum_{k=1}^{K} y_k \log p_k(x) \]  \hspace{1cm} (10)

If the sample output category is \( k \), then \( y_k = 1 \). The probability \( p_k(x) \) of class \( k \) is expressed as follows:

\[ p_k(x) = \frac{\exp(f_k(x))}{\sum_{i=1}^{K} \exp(f_i(x))} \]  \hspace{1cm} (11)

The negative gradient error of class \( l \) corresponding to the \( i^{th} \) sample of round \( t \) is calculated as follows:

\[ r_{il} = - \left[ \frac{\partial L(y_i,f(x_i))}{\partial f(x_i)} \right]_{f(x_i)=f_l(x_i)} = y_i \cdot p_t, t-1(x_i) \]  \hspace{1cm} (12)

In this case, for the generated decision tree, the best negative gradient fitting value of each leaf node is:

\[ c_{ijl} = \arg \min_{x_i} \sum_{i=0}^{m} \sum_{k=1}^{K} L(y_k,f_{t-1,l}(x)) + \sum_{j=0}^{J} c_j I(x_i \in R_{ijl}) \]  \hspace{1cm} (13)

GBDT algorithm is suitable for dealing with various types of data and can adapt to a variety of loss functions. This method is suitable for low dimensional and dense data, has good interpretability and wide application fields.

a: SUPPORT VECTOR MACHINE
Support Vector Machine (SVM), first proposed by Cortes and Vapnik in 1995, is a sparse kernel decision-making method based on statistics, optimization, and machine learning theory. The decision-making of this method is based on the linear combination of feature weights, that is, the point product between the non-existent document vector and the weight vector, and determines the document category by setting a certain threshold value. Support Vector Machine (SVM) algorithm tries to maximize the decision boundary and the boundary between the two classes, so that new instances leave more “space” for correct classification.

b: RANDOM FOREST
Random Forest (RF) is an ensemble machine learning algorithm that uses the decision tree as base classifier [24]. It uses a large number of unrelated decision trees to combine “weak learners” into “strong learners”. Random forest aggregates the output results of some shallow trees to form an additional layer and then bag them, instead of the results based on the output of a single deep tree. The bagging process uses independent follow-up trees to construct \( n \) predictors by bootstrapping the samples of the data set. These \( n \) predictors are combined to solve classification or estimation problems by averaging. Although a single classifier is a weak learner, if all the classifiers are combined, a strong learner can be formed. In view of the high variance and bias of decision trees, Random Forest improves its estimation performance by averaging multiple decision trees. Random Forest can show good accuracy and high efficiency in large-scale data sets. Even if part of the data is missing, this method is still an effective method in estimating missing data and ensuring accuracy.

c: EXTREME GRADIENT BOOSTING
Extreme Gradient Boosting (XGBoost) is a lifting tree model, which integrates many tree models into a strong classifier. XGBoost adds a regularization term to the cost function to control the complexity of the model, make the model simpler, and prevent over-fitting. XGBoost determines the structure of the tree by derivation, that is, how to find the best feature in each feature split, uses greedy algorithm to traverse all feature partition points of all features, takes the objective function value as the evaluation function, makes the split objective function value gain more than the objective function of the single leaf node, and increases the threshold value to prevent the tree from growing too deep. XGBoost optimizes the loss function through the objective function, and it can support parallelization, with fast training speed and wide application.

C. EVALUATION CRITERIA
1) PERPLEXITY
In this paper, perplexity is used to determine the optimal number of topics in the LDA topic model, and the number of topics with the minimum perplexity is used as the optimal number of topics in the LDA topic model. With the increase of the number of topics, the perplexity of the model decreases gradually. However, the more the number of topics, the more difficult it is to determine the content of each topic in the model, and the difference between topics will be smaller and smaller. In practical application, the number of topics whose perplexity degree is at the “inflection point” of the minimum value is generally selected. The calculation formula of perplexity is as follows:

\[ \text{perplexity} = \exp \left\{ - \left( \sum_{m-1}^{M} \sum_{n_m}^{N_m} \log \left( \sum_{k=1}^{K} p(w_n|z_k) p(z_k|d_m) \right) \right) \right\} \]  \hspace{1cm} (14)

where \( M \) is the number of texts in the data set, \( N_m \) is the total number of words in the \( M \)th text, and \( K \) is the number of topics, \( p(w_n|z_k) \) indicates the probability of the words \( w_n \) under the topic \( z_k \), and \( p(z_k|d_m) \) is the probability of the test text \( d_m \) under the topic \( z_k \).

2) K-FOLD CROSS VALIDATION
The original data are randomly divided into two groups, one is the training set, which is used to train the classifier; one is the validation set, which is used to verify the performance of the classifier. The corpus is divided into \( K \) parts. In \( K \) rounds, one partition is selected as the test set, and the remaining \( k-1 \) is
used for training. The accuracy of the classifier is obtained after the K rounds average. K-fold cross-validation can effectively avoid overfitting and underfitting, with relatively low bias and variance, and the final results are more convincing.

3) CONFUSION MATRIX
The confusion matrix is used to visualize the performance of the classification algorithm by using the data in the matrix. It is used to compare the predicted classification results with the actual classification results. It is an important evaluation method to evaluate the accuracy and robustness of the classifier. In the confusion matrix, each row adds up to the total number of real samples, and the diagonal line represents the correctly classified samples. Confusion matrix is a method to check the performance of classifiers at each label level. It can calculate the accuracy rate, recall rate, and F1-measure.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (15)
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \quad (16)
\]
\[
F1\text{-measure} = \frac{2P \times R}{P + R} \quad (17)
\]

True positive (TP) indicates that the actual sample category is consistent with the predicted result. False negative (FN) indicates that the sample is predicted to be another class; False positive (FP) indicates that other class samples are predicted as this class. F1-measure is the harmonic average of accuracy and recall.

4) STATISTICAL SIGNIFICANCE TEST
   a: FRIEDMAN TEST
Friedman test is a nonparametric test for significance test of multiple paired samples. Assuming that the feature extraction methods and algorithms are paired samples, the overall distribution of classifier accuracy under different combinations is compared. If there is no difference among the K feature extraction methods, the rank sum of the accuracy of each classifier under each feature extraction method is  \( \overline{R}_i \), and the average rank is \( \overline{R} \), \( n \) is the sample size, there is:

\[
\sum_{i=1}^{k} n(1 + 2 + 3 + \cdots + k) = \frac{n}{2}k(k + 1) \quad (18)
\]

Therefore, assuming that there is no difference between different feature extraction methods, then \( \overline{R}_i \) should be equal to \( \frac{k+1}{2} \), Friedman test statistics can be expressed as:

\[
Friedman = \frac{12n}{k(k + 1)} \sum_{k=1}^{k} (\overline{R}_i - \frac{k + 1}{2})^2 \quad (19)
\]

The Friedman test statistics and the corresponding probability \( p \) value are calculated. If the \( p \) value is less than the significance level, it indicates that there are significant differences between the text emotion classifiers constructed by different feature extraction methods. On the contrary, there is no significant difference.

b: KRUSKAL-WALLIS TEST
The Kruskal-Wallis is a multiple independent sample test. The processing flow of the Kruskal-Wallis test is: the data from different populations are mixed and arranged in ascending order, the rank of variable value is calculated, and then the rank mean is compared. Similar to the Friedman test, the Kruskal-Wallis test uses probability \( p \) value to judge whether there are significant differences among populations from multiple independent samples.

IV. DATA SOURCES AND PREPROCESSING
A. DATA SOURCE AND SAMPLE COLLECTION
This paper uses GooSeeker data collector to collect Zhubajie platform (https://www.zbj.com/) review text received by crowdsourcing participants. Zhubajie platform is the most active crowdsourcing platform in China. Its registered users are more than 16 million, and the market share is more than 50%. It is the leader of the crowdsourcing market in China. The crowdsourcing platform has rich reviews, and is also an important platform for many scholars to research [11], [25], [26], which is applicable to the research of this study.

Data collection is divided into two steps: the first step is to determine the transaction scope of the acquisition task, including brand design, IT software development, marketing promotion, e-commerce services, industrial design, film and television animation, cloud services, and game development. The collected data covers the data from eight major fields of the Zhubajie platform from December 1, 2006, to January 31, 2019. The evaluation indicators include store ID, store link, task completion quality, task speed, and task attitude. A total of 4357 crowdsourcing participant samples are obtained. After eliminating duplicate samples and information distortion samples, 3298 effective crowdsourcing participants are finally obtained. The second step is to collect detailed transaction evaluation of 3298 samples on the Zhubajie platform, including employer ID, single transaction amount, single transaction date, and single transaction review text. A total of 46782 reviews are acquired. After deleting the duplicate data and meaningless review text, 20633 valid information is finally obtained.

B. DATA PREPROCESSING
Because the text data has a lot of noise, it cannot be analyzed and mined directly, this study to remove duplicate reviews, delete invalid irrelevant reviews, and reviews with less information. In this paper, Python 3.7 is used to preprocess the review text data, Jieba word segmentation is used for Chinese word and part of speech tagging. Harbin Institute of technology stop word list is used to remove the stop words that have nothing to do with emotion and task, including conjunctions, exclamations, pronouns, such as “ah”, “but”, “so”, and so on, and then get the review text feature set.

The review text is annotated manually so that machine learning algorithms can be used for supervised learning. For
there are two kinds of emotion polarity coexisting in the reviews of crowdsourcing participants, to classify sentiment more accurately, the sentiment of review text is divided into five categories: “very satisfied”, “satisfied”, “general”, “dissatisfied” and “very dissatisfied”, which are marked as 1, 2, 3, 4 and 5. Some manually annotated review text are shown in Table 1.

### TABLE 1. Some manually annotated review text.

| Employer | Transaction amount (yuan) | Transaction date | Review text                                                                                                                                                                                                 | Tagging category |
|----------|---------------------------|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|
| **7420_59kempt** | 13000.00 | 2018-06-30 | Very satisfied, very good professionalism and service attitude. The work is innovative and the design concept is unconventional. There may be some misunderstanding in the early communication, but it is corrected in time. The final work delivered is beyond my expectation. | 1                |
| **0474ppppyk** | 3600.00 | 2018-06-31 |                                                                                                                                                                                                          | 1                |

aa66644999 | 699.00 | 2018-04-23 | I can only tell you the truth. Their attitude is very good, but their work really need to be improved.                                                                                                                                                           | 2                |
| m_9992_cgc5pm | 1120.00 | 2018-11-08 | In the beginning, it was very difficult to communicate with each other. The final modification was very patient. Out of conscience, I gave a good comment.                                                                                                                   | 2                |
| m_8913_1elirj | 580.00 | 2018-08-26 | Generally speaking, the design components are less, and they are basically set of templates. The photo editor is impatient.                                                                                           | 3                |
| m_1613_n063jx | 500.00 | 2018-03-05 | Score 60 points. It took two and a half days for the work that could be completed in a short period of time.                                                                                                                                               | 3                |

**Jianghu** | 100.00 | 2013-10-16 | The attitude is good, the level is really not good, it's completely rookie level. The construction period is too long, the quality and expectations are too poor, and communication is no problem.                   | 4                |
| m_6883_68xqy6 | 399.00 | 2017-01-13 |                                                                                                                                                                                                          | 4                |
| m_4406_4kcg50 | 950.00 | 2016-10-07 | Level rubbish, does not meet the requirements and has a bad attitude. The above praise is because the evaluation is forgotten, and the system defaults.                                                                                                        | 5                |
| dinggg1 | 5500.00 | 2016-07-06 | A real liar, I have encountered the worst programmer, anyone is better than him. Not to mention the programming ability, doing things is very irresponsible. The postponement for half a month has not been completed yet.                                                                 | 5                |

### V. EVALUATION OF EXPERIMENTAL RESULTS

Taking review text received by crowdsourcing participants on the Zhubajie platform as the case set, this paper uses a variety of feature extraction methods, including N-gram, Word2vec, and TF-IDF algorithm to extract text features, uses LDA topic model to expand the number of text features, and determines the optimal number of topics of LDA model by calculating the perplexity degree, and constructs models based on support vector machine (SVM), random forest (RF), XGBoost, and GBDT algorithm. The data set is divided into training set and test set, 70% of which are training set and 30% of which are test set. Using ten-fold cross-validation, confusion matrix validation, and statistical significance analysis to evaluate the influence of LDA model on the classification and prediction of sentiment fuzzy review text, and to compare the accuracy and stability of different text sentiment classifiers. In this paper, we use Python 3.6, Pycharm 2021, and SPSS 25.0 software.

### A. TEXT FEATURE EXTRACTION

Because the semantic features of review text can be reflected by the combination of various text feature data, the high-frequency text features in the comments are the crowdsourcing activity features concerned by employers and their corresponding emotional feedback. Therefore, the statistics of the content and frequency of the feature text in the review text can dig out the key factors that affect the employer’s transaction decision and improve the employer’s satisfaction, which is of great significance to improve the task service level of crowdsourcing participants and carry out customer relationship management.

Based on the statistics of word frequency, the high-frequency words in review text, including nouns, adjectives, and verbs, are obtained. It is found that employers pay more attention to the service, specialty, communication, speed, patience, modification, effect, attitude, quality, recommendation, carefulness, completion, and efficiency provided by crowdsourcing participants. The cloud picture of high-frequency words in the review corpus is shown in Figure 2.

This study uses TF-IDF method to calculate the weight of text features. In this method, word frequency and inverse document frequency are two factors. The weight of each keyword is calculated, that is, the weight of extracted features. The more times the extracted features appear in the review text, the more important they are, however, as the frequency of corpus increases, its importance decreases. If the keywords appear more in the review text and less in the corpus, they have a good ability to distinguish categories.

### B. LDA TOPIC MODEL EXTENDS TEXT FEATURES

To improve the prediction ability of the classifier, the LDA topic model is used to further expand the text features. The perplexity index evaluation method is used for fixing the topic
number, and the range of topic numbers is between 1 and 20, to calculate the confusion degree of LDA topic model. When the number of topics is 8, the decline rate of perplexity degree is significantly reduced, and the perplexity degree is nearly stable, that is, the “inflexion point” of perplexity degree tends to be gentle in the process of gradual reduction. When the number of topics is more than 10, the perplexity value gradually increases. With the increase of the number of topics, the computational cost of the LDA model increases correspondingly, and it is easy to over fit. Therefore, this study determines that the optimal number of LDA topics is 8, that is, selecting 8 topics can better cover the lexical information and reduce the lexical dimension. The perplexity curve of the LDA topic model is shown in Figure 3.

The LDA topic model is used to expand the features of review text and output the keywords of eight topics. Because the LDA topic model is an unsupervised distribution model, there are some words that are not obvious with topic content. In order to describe the content of the topic, this paper selects 10 keywords that are more valuable to describe the eight topics. The results of topic feature word extraction of crowdsourcing participants’ reviews are shown in Table 2.

The LDA topic model is used to determine the optimal number of topics, and the key factors affecting the employer’s transaction satisfaction are found out. According to the extracted keywords, the eight topics of employers’ attention in crowdsourcing activities are summarized as work speed, work attitude, work communication, professional level, task completion, trading platform, trading experience, and willingness to cooperate again.

Theme 1 “work speed” and theme 2 “work attitude” reflect the transaction characteristics of crowdsourcing activities. Crowdsourcing is different from traditional “e-commerce”, the subject matter of transaction is intangible goods. Through the Internet, crowdsourcing participants transform their wisdom, knowledge, ability, and experience into task results and obtain labor remuneration. The core indicators of crowdsourcing participants’ task completion are “work speed” and “work attitude” extracted from the LDA topic model.

Theme 3 “work communication” has not attracted the attention of crowdsourcing platform and has not been included in the reputation evaluation system of crowdsourcing participants. Crowdsourcing activities are different from physical commodity transactions. If the final crowdsourcing task results submitted by crowdsourcing participants can not meet the expectations and requirements of employers, the time, physical strength, energy, and other non-monetary costs invested by crowdsourcing participants cannot be redeemed, which will cause a waste of human resources and bring time and monetary losses to employers. Therefore, the successful implementation of crowdsourcing depends on repeated communication between employers and crowdsourcing participants before, during, and after the task to achieve the employer’s task goals. Effective work communication is the key to the successful completion of the task.

Theme 4 and theme 5 are “professional level” and “task completion” respectively. On the Zhubajie crowdsourcing platform, the quality of task completion is one of the important indicators to measure the reputation of crowdsourcing participants. “Task completion” and “professional level” further refine this indicator into whether the task can be successfully completed and whether the completed task has the professional level.

Theme 6 “trading platform” is an innovative conclusion drawn from the analysis of the LDA topic model. Through text analysis, it is found that employers are not satisfied with the crowdsourcing platform, and feedback is in the form of reviews text. At present, Zhubajie crowdsourcing platform has not yet provided a channel for feedback on the insufficient trading mechanism of the trading platform, such as whether the crowdsourcing platform can effectively restrain the behavior of crowdsourcing participants, whether the trading decisions are fair and just, whether the trading rules are perfect, whether the interface is friendly, and whether the process is standardized. Employers choose to feedback the evaluation of crowdsourcing platforms through the review text of crowdsourcing participants, which affects the overall evaluation of crowdsourcing participants.
Theme 7 “transaction experience” reflects the subjective feelings of employers in the transaction process, which is closely related to the employer’s task expectation. Whether the employer has a good trading experience depends on the difference between performance and expectation after receiving the submitted task results. If the perceived effect of services provided by crowdsourcing participants to meet the needs of employers exceeds the expectations of employers, the employers will be satisfied, otherwise, they will not be satisfied. The key to improving the performance of crowdsourcing is to improve the employer’s transaction experience and achieve the employer’s satisfaction.

Theme 8 is “willingness to cooperate”. Crowdsourcing participants to provide services to meet the needs of employers, satisfied employers will repeatedly purchase services, and crowdsourcing participants to cooperate again, resulting in customer loyalty. Since it costs more to attract new customers than to maintain old customers, it is of great significance for crowdsourcing participants to maintain old customers and cultivate customer loyalty among competitors. From the extracted keywords, “price” is an important factor for employers to decide whether to cooperate again.

C. TEXT SENTIMENT CLASSIFICATION

Based on 20633 review text obtained from the Zhubajie platform, sentiment classification prediction of review text is carried out. In this paper, N-gram, Word2vec, and TF-IDF algorithms are used to extract text features, and the LDA topic model is used to expand the text features. 100 topic words with the highest probability are selected under each topic to expand the text features. A text sentiment classifier of crowdsourcing participants based on SVM, RF, GBDT, and XGBoost is constructed. In this study, the kernel function of the SVM classifier is the Gaussian kernel function, which is widely used and flexible. The super parameters of the classifier are set manually. The kernel function parameter is set to 0.5, and the penalty parameter C is set to the default value of 1.0.

Because the data set is imbalanced, there are more samples in the first category and fewer samples in the second, third, fourth, and fifth categories, so the first category samples are under-sampled to form a data set with the other four categories of samples, and 70% of the text data is used as the training set to train the sentiment classifier; 30% text data is used as the test set to verify the performance of the model. Ten-fold cross-validation, confusion matrix, and statistical significance are used to verify the performance of the crowdsourcing participant sentiment classifier.

1) TEN-FOLD CROSS-VALIDATION

To test the improving effect of the LDA topic model for text classification with fuzzy emotion boundaries, this study uses N-gram, Word2vec, and TF-IDF to extract text features, and constructs text sentiment classifiers based on SVM, RF, GBDT, and XGBoost algorithms. By fusing N-gram, Word2vec, TF-IDF feature extraction, and LDA topic model to extend text features, text sentiment classifiers based on the above four machine learning algorithms are constructed and compared. The ten-fold cross-validation accuracy of the crowdsourcing participant text sentiment classifier is shown in Table 3.

|               | N-gram | Word2vec | TF-IDF | N-gram | Word2vec | TF-IDF | N-gram | Word2vec | TF-IDF |
|---------------|--------|----------|--------|--------|----------|--------|--------|----------|--------|
| SVM           | 0.843  | 0.831    | 0.844  | 0.835  | 0.838    | 0.856  |
| RF            | 0.823  | 0.822    | 0.821  | 0.829  | 0.834    | 0.845  |
| GBDT          | 0.857  | 0.854    | 0.858  | 0.876  | 0.874    | 0.881  |
| XGBoost       | 0.851  | 0.85    | 0.852  | 0.864  | 0.867    | 0.875  |
| Rank          | 2.13   | 1.63     | 3.5    | 3.75   | 4.00     | 6.00   |

From the perspective of the feature extraction method, after fusing TF-IDF text feature extraction and the LDA topic model to expand review text, the classification accuracy of SVM, RF, GBDT, and XGBoost reaches the maximum, which is 0.856, 0.845, 0.881, and 0.875 respectively. When using Word2vec to extract text features, SVM, GBDT, and XGBoost classifiers have the lowest classification accuracy. The accuracy of RF, GBDT, and XGBoost classifiers is improved by the LDA. The experimental results show that, except for the SVM classifier using N-gram to extract text features is better than the N-gram-LDA method, the performance of other classifiers can be improved in varying degrees through the LDA topic model. The prediction accuracy of the GBDT classifier is improved by 0.019, 0.02, and 0.013, respectively.

From the perspective of classifier accuracy, the accuracy of GBDT is the highest among the four classifiers, which are 0.857, 0.854, 0.868, 0.876, 0.874, and 0.881, respectively, when N-gram, Word2vec, TF-IDF, N-gram-LDA, Word2vec-LDA, and TF-IDF-LDA are used to extract text features. When using the TF-IDF-LDA method for feature extraction, the accuracy of XGBoost is 0.875, next to GBDT, followed by SVM classifier, which is 0.856, and RF classifier, which is 0.845.

The effect of different text feature extraction methods is verified by the Friedman test. The observed value of the Friedman test statistic is 13.77, and the asymptotic significance is 0.173, which is less than the significance level of 0.05. It shows that the performance of the six text feature extraction methods is significantly different. The average rank of the six text feature extraction methods N-gram, Word2vec, TF-IDF, N-gram-LDA, Word2vec-LDA, and TF-IDF-LDA are 2.13, 1.63, 3.5, 3.75, 4.00, and 6.00 respectively, as shown in Table 3. TF-IDF-LDA feature extraction method is the best, Word2vec-LDA can also achieve a good feature extraction effect. In this case, when using Word2vec for text feature extraction, the overall performance of the four classifiers is the worst. The results show that there are significant differences among different feature extraction methods, and the TF-IDF-LDA feature extraction method is the best.
2) CONFUSION MATRIX VERIFICATION

To further judge the classification effect of different feature extraction methods on the text emotion boundary fuzzy, there are two kinds of emotion polarity review text, this paper uses the confusion matrix to test the classification accuracy of the classifier for different categories of samples.

Because different classification errors will lead to different loss costs, the crowdsourcing participants who are “very dissatisfied” and “dissatisfied” are wrongly divided into “very satisfied” or “satisfied”, facing higher misclassification costs. The ratio of the number of correctly classified samples to the total number of predicted samples of this category is described precision; Recall describes the proportion of the number of correctly predicted samples in the real number of samples in the category, which are important factors to measure the stability of the classifier. The higher the precision and recall of the classifier, the better the classification performance is. F1-measure is a comprehensive index reflecting the classification precision and recall. This paper uses the precision, recall, and F1 measure value as the evaluation indexes to measure the classification effect of the classifier.

In this paper, the review text is divided into five sentiment categories: “very satisfied”, “satisfied”, “general”, “dissatisfied” and “very dissatisfied”. i represents the sample category, and the value range is 1 to 5. \( P_i \) is the precision rate of i category samples, that is, the ratio of the number of correctly classified samples to the total number of predicted category i samples. \( R_i \) refers to the recall rate of category i samples, that is, the ratio of correctly predicted category i samples to the total number of real samples of this category. \( F_i \) is the harmonic mean of the precision rate and recall rate of category i samples. The precision rate, recall rate, and F1-measure value of crowdsourcing participants’ sentiment classification prediction are calculated by the confusion matrix. The results are shown in Table 4.

In terms of precision, the RF classifier has the worst ability to distinguish the second, third, fourth, and fifth samples. When using N-gram, Word2vec, and N-gram-LDA to extract text features, the RF classifier can only recognize the first category of samples, but it is difficult to distinguish the second, third, fourth, and fifth kinds of samples. When TF-IDF and Word2vec-LDA are used to extract text features, the RF classifier can recognize the first and fifth category samples, but still cannot recognize the second, third, and fourth category samples. When the TF-IDF-LDA method is used to extract text features, RF can recognize all five different sentiment categories, and the precision rates of the second, third, and fourth categories are 0.356, 0.6, and 0.29 respectively. The results show that the TF-IDF method can significantly improve the ability of the RF classifier to distinguish the second, third, and fourth category samples after extracting text features and extending text features by LDA. When using N-gram, Word2vec, and TF-IDF to extract text features, SVM and XGBoost classifier cannot recognize the fourth kind of samples. The LDA topic model is used to extend the features of review text, which solves the problem that the SVM and XGBoost classifier cannot recognize the fourth category of samples.

The precision of different classifiers is compared when TF-IDF-LDA is used to extract and extend text features. The precision of the first category of GBDT is 0.927, followed by XGBoost, which is 0.923. The precision of the second, third and fifth category of the SVM classifier is the highest, which is 0.793, 0.887, and 0.667 respectively. The highest precision rate of the XGBoost classifier is 0.368, while the lowest precision rate of the SVM classifier is 0.125, with a difference of 0.243. It shows that the ability of the XGBoost classifier to distinguish the fourth class of samples is significantly improved after extending the text features by LDA.

In terms of recall rate, when using N-gram, Word2vec, and N-gram-LDA to extract text features, the recall rate of the first category of RF is the highest, which is 1, but the recall rates of the other four categories are 0, 0, 0, and 0 respectively, which indicates that the classifier wrongly divides the other four categories into the first category.

When TF-IDF-LDA is used to extract and extend text features, the recall rates of the second, third, fourth, and fifth categories of RF are 0.477, 0.062, 0.045, and 0.135, respectively. It can be seen that the feature expansion method combined with LDA effectively improves the classification ability of the RF classifier for the second, third, fourth, and fifth categories of samples.

In this case, the second and fourth category samples are important factors to measure the performance of the classifier. Compared with the other three types of review text, the second and fourth categories of samples have more indistinct emotional boundaries, and there are often different emotional polarity words coexisting, so it is also the focus of this paper. When TF-IDF-LDA is used to extract and extend text features, the second recall rate of the RF classifier is improved from 0 to 0.477. The fourth class recall rate of SVM and XGBoost classifiers increased from 0 to 0.01 and 0.178 respectively. Experiments show that the TF-IDF-LDA text feature extraction method can significantly improve the ability to distinguish the second and fourth types of review text.

From the perspective of the F1-measure, among the four classifiers constructed by TF-IDF feature extraction and LDA extended text feature, the highest F1-measure of the first category sample of the XGBoost classifier is 0.952. The F1-measures of the second, third, fourth, and fifth samples of the GBDT classifier are the highest, which are 0.462, 0.571, 0.278, and 0.647 respectively. Compared with the case of only using TF-IDF to extract text features, the GBDT classifier improves the F1-measures of the second, third, fourth, and fifth samples by 0.124, 0.065, 0.123, and 0.182, respectively. The experimental results show that GBDT has the best performance in precision, recall, and F1-measure. Through the method of LDA extended text features, it can significantly improve the ability to distinguish the review text.
TABLE 4. Evaluation results of crowdsourcing participants’ text emotion classifier confusion matrix.

| Feature extraction method | Classification algorithm | Accuracy | Precision | Recall | F1-measure |
|---------------------------|--------------------------|----------|-----------|--------|------------|
|                           |                          | ACC      | R1        | R2     | R3         | R4         | R5         | F1       |
|                           | SVM                      | 0.843    | 0.976     | 0.245  | 0.39       | 0.122      | 0.926      | 0.285    |
|                           | RF                       | 0.823    | 0.822     | 0.117  | 0.036      | 0.094      | 0.902      | 0.309    |
|                           | GBDT                     | 0.857    | 0.978     | 0.078  | 0.590      | 0.193      | 0.912      | 0.126    |
|                           | XGBoost                   | 0.851    | 0.997     | 0.548  | 0.019      | 0.938      | 0.918      | 0.664    |
| Word2vec                  | SVM                      | 0.831    | 0.174     | 0.067  | 0.036      | 0.916      | 0.244      | 0.061    |
|                           | RF                       | 0.822    | 0.000     | 0.117  | 0.094      | 0.902      | 0.000      | 0.000    |
|                           | GBDT                     | 0.854    | 0.982     | 0.252  | 0.345      | 0.933      | 0.293      | 0.041    |
|                           | XGBoost                   | 0.830    | 0.992     | 0.033  | 0.094      | 0.911      | 0.052      | 0.156    |
| TF-IDF                    | SVM                      | 0.844    | 0.994     | 0.128  | 0.336      | 0.21       | 0.918      | 0.226    |
|                           | RF                       | 0.821    | 0.000     | 0.000  | 0.000      | 0.903      | 0.000      | 0.000    |
|                           | GBDT                     | 0.868    | 0.986     | 0.248  | 0.11       | 0.34       | 0.935      | 0.506    |
|                           | XGBoost                   | 0.852    | 0.988     | 0.152  | 0.456      | 0.224      | 0.919      | 0.582    |
| N-gram -LDA               | SVM                      | 0.835    | 0.099     | 0.204  | 0.139      | 0.014      | 0.188      | 0.227    |
|                           | RF                       | 0.829    | 0.000     | 0.000  | 0.000      | 0.903      | 0.000      | 0.000    |
|                           | GBDT                     | 0.876    | 0.976     | 0.419  | 0.347      | 0.512      | 0.953      | 0.416    |
|                           | XGBoost                   | 0.864    | 0.981     | 0.361  | 0.242      | 0.386      | 0.947      | 0.423    |
| Word2vec -LDA             | SVM                      | 0.838    | 0.098     | 0.108  | 0.046      | 0.011      | 0.066      | 0.199    |
|                           | RF                       | 0.834    | 0.007     | 0.000  | 0.078      | 0.911      | 0.013      | 0.000    |
|                           | GBDT                     | 0.874    | 0.969     | 0.494  | 0.283      | 0.096      | 0.473      | 0.545    |
|                           | XGBoost                   | 0.867    | 0.972     | 0.422  | 0.193      | 0.078      | 0.392      | 0.427    |
| TF-IDF -LDA               | SVM                      | 0.856    | 0.993     | 0.105  | 0.36      | 0.01       | 0.273      | 0.295    |
|                           | RF                       | 0.845    | 0.988     | 0.477  | 0.062      | 0.045      | 0.135      | 0.365    |
|                           | GBDT                     | 0.881    | 0.982     | 0.313  | 0.422      | 0.165      | 0.511      | 0.462    |
|                           | XGBoost                   | 0.875    | 0.981     | 0.408  | 0.415      | 0.178      | 0.336      | 0.505    |

FIGURE 4. Kruskal-Wallis test results of text sentiment classification prediction accuracy of crowdsourcing participants.

with fuzzy emotional boundaries, and significantly improve the performance of the classifier.

3) STATISTICAL SIGNIFICANCE TEST
Kruskal Wallis test is used to compare the performance of crowdsourcing participants’ sentiment classifiers when using different methods to extract text features. When using different feature extraction methods, the accuracy of SVM, RF, GBDT, and XGBoost is arranged in mixed ascending order, and the average rank of each variable is calculated to judge whether there is a significant difference in the overall distribution of accuracy. The Kruskal-Wallis test results of four classifiers are shown in Figure 4.

Using six different methods of N-gram, Word2vec, TF-IDF, N-gram-LDA, Word2vec-LDA, and TF-IDF-LDA to extract text features, the average rank calculation results of SVM, RF, GBDT, and XGBoost are shown in Figure 4. Using six different feature extraction methods, the probability p-values of K-W test results of four classifiers are close to 0. If the significance level is 0.01, because the probability p-value is less than the significance level, it indicates that the average rank difference is significant, that is a significant difference in the overall distribution of four classifiers. From the Kruskal-Wallis test results, the average rank of GBDT and XGBoost classifiers is close and significantly higher than that of SVM and RF classifiers after using the LDA topic model.
to expand reviews text, which shows that the classification performance of GBDT and XGBoost classifiers is better than that of SVM and RF classifiers.

The average rank of the GBDT classifier is the highest when using six methods to extract text features. When the TF-IDF-LDA method is used, the lowest average rank of RF classifier is 6.8, the highest average rank of GBDT classifier is 22.4, and the average rank of XGBoost classifier is 20.9, which is slightly lower than that of the GBDT classifier, and the performance of GBDT classifier is the best. K-W test further verifies that the GBDT classifier has the best classification result and the most stable performance.

To sum up, after extracting and extending features by integrating the LDA topic model, the ability of sentiment classification and prediction of the classifier is significantly improved, especially the ability of the classifier to distinguish the second, third, and fourth categories of samples are improved, the problem that the classification effect is not obvious due to the fuzzy boundary of emotion category is alleviated, and the sentiment category of employer’s review text can be more accurately divided, and improve the classification accuracy. The GBDT text sentiment classification method is the best in classification accuracy, stability, and comprehensive performance.

VI. CONCLUSION AND PROSPECT

In this paper, a supervised text sentiment classification method based on Latent Dirichlet Allocation (LDA) is proposed to improve the classification performance of short text with fuzzy sentiment boundaries, aiming at more accurately mining the sentiment tendency of users when two kinds of emotional polarity appear in the review text. The work of this study is as follows: firstly, the text corpus is preprocessed, including deduplication, de stop words, and Jieba segmentation; then N-gram, Word2vec, and TF-IDF are used to extract features; the LDA topic model is used to divide the topics of the review text, determine the optimal number of topics, and expand the features of the review text. Taking 20633 reviews text received by crowdsourcing participants on Zhubajie platform as the data set, N-gram, Word2vec, TF-IDF method, and the method of fusing N-gram, Word2vec, TF-IDF, and LDA topic model were used to extract and expand text features, construct crowdsourcing participant text sentiment classifier based on Support Vector Machine (SVM), Random Forest (RF), Gradient boosting decision tree (GBDT), and Extreme Gradient Boosting (Xgboost) algorithms. The imbalanced data set is under-sampled, and 70% of the text data is used as the training set to train the classifier; 30% of text data is used as the test set to verify the performance of the classifier. Ten-fold cross-validation, confusion matrix, and statistical significance test are used to evaluate the performance of the text sentiment classifier.

The research shows that the method of combining TF-IDF and latent Dirichlet topic model (LDA) to expand and extract the features of review text can solve the problem that SVM and XGBoost classifiers cannot distinguish the fourth category samples, and RF classifiers cannot distinguish the second, third, and fourth category samples. It can significantly improve the performance of the GBDT classifier, improve the problem that the classification effect is not obvious due to the fuzzy boundary of emotion category, and improve the classification accuracy of the second, third, and fourth category samples. The proposed method combines with TF-IDF and LDA topic model, based on the GBDT algorithm, performs well in accuracy, recall, and F1-measure, and has the best classification performance and robustness.

For further directions, it can be studied from two aspects. Crowdsourcing involves many fields, the review text data in different task areas of crowdsourcing has its domain particularity, the effect of its classifier is affected by the quantity and quality of annotation data. To use the existing domain training model to deal with similar tasks in other fields, to reduce the time and energy of retraining the model, we can further explore cross-domain transfer learning to solve the problem of lack of annotation resources in the target domain.

Crowdsourcing participants have more positive comments and less negative comments in the review text, which is a typical class imbalanced data. Different classification errors will lead to different loss costs. If the “very satisfied” crowdsourcing participants are wrongly divided into “very dissatisfied” crowdsourcing participants, they will face higher misclassification costs. Therefore, we can try to use more complex or the latest classification technology to explore the emotion classification method of class imbalanced data sets.

Based on the research on the emotion classification of crowdsourcing participants, combined with the transaction behavior data of crowdsourcing participants, to establish effective crowdsourcing participant reputation evaluation system, task recommendation system, and price evaluation system, assisting employers to discover relevant knowledge, optimize decision-making, and encourage crowdsourcing participants to participate and complete tasks more efficiently, is the future research direction.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their insightful reviews and constructive suggestions.

REFERENCES

[1] W. Duan, B. Gu, and A. B. Whinston, “Do online reviews matter?— An empirical investigation of panel data,” Decis. Support Syst., vol. 45, pp. 1007–1016, Nov. 2008.
[2] W. F. Wu, “The impact of review text on hotel satisfaction: An emotional analysis method,” Data Anal. Knowl. Discovery, vol. 1, pp. 62–71, Mar. 2017.
[3] J. A. Chevalier and D. Mayzlin, “The effect of word of mouth on sales: Online book reviews,” J. Marketing Res., vol. 43, no. 3, pp. 345–354, 2006.
[4] C. X. Zhai and S. Massung, Text Data Management and Analysis. Beijing, China: Industrial Machinery Press, 2019, pp. 192–198.
[5] S. Dasgupta and K. Sengupta, “Analyzing consumer reviews with text mining approach: A case study on Samsung Galaxy S3,” Paradigm, vol. 20, no. 1, pp. 56–68, 2016.
[6] Z. M. Zeng and P. Y. Wang, “Micro blog sentiment analysis of public security incidents based on double attention and Bi-LSTM,” Inf. Sci., vol. 37, pp. 23–29, Mar. 2019.
YANRONG HUANG received the Ph.D. degree in management science and engineering from the School of Computer Science, Wuhan University, China, in 2020. She is currently a Lecturer with the School of Management, Zhejiang University of Water Resource and Electric Power, Zhejiang, China. Her research interests include the areas of crowdsourcing, behavior analysis, and customer behavior analysis.

RUI WANG received the Ph.D. degree in management science and engineering from Wuhan University, Wuhan, China, in 2015. Since 2015, he has been an Associate Professor with the E-commerce Department, Jiangxi University of Science and Technology. Since 2018, he has been holding a postdoctoral position with the Department of Agricultural Economic Management, Fujian Agriculture and Forestry University. His research interests include online comment analysis and customer behavior analysis.

BIN HUANG was born in Kunming, Yunnan, China, in 1977. He received the Ph.D. degree in economics and management of technology from Zhejiang University of Technology, Zhejiang, China, in 2018. He is currently an Associate Professor with the College of Economics and Management, Zhejiang University of Water Resource and Electric Power, Zhejiang. His research interests include crowdsourcing and electronic commerce.

BO WEI received the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, China, in 2013. In 2013, he was a Lecturer with the School of Software, East China Jiaotong University, Nanchang, China. In 2015, he worked as a Postdoctoral Researcher with the Institute of Automation, Chinese Academy of Science, Beijing, China. He is currently a Lecturer with the School of Informatics Science and Technology, Zhejiang Sci-tech University, Hangzhou, Zhejiang, China. His current research interests include the areas of computational intelligence techniques, machine learning, and data mining.

SHU LI ZHENG received the Ph.D. degree in business administration management from Zhejiang Gongshang University, Hangzhou, in 2018. She is currently an Associate Professor with the College of Economics and Management, Zhejiang University of Water Resource and Electric Power, Zhejiang, China. Her research interests include business model innovation and cross-border electricity.

MIN CHEN received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University. He is currently a Professor with Wuhan University. He is also the Director of Computer Society, in Hubei. His current research interests include data science, artificial intelligence, behavior informatics, and their enterprise applications. He was a recipient of second prize for scientific and technological progress of national defense science and technology commission, in 1992.