IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2021, accepted April 28, 2021, date of publication May 3, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077059

Deep Learning Sentiment Classification Based on Weak Tagging Information

CHUANTAO WANG, XUEXIN YANG, AND LINKAI DING
School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
Beijing Engineering Research Center of Monitoring for Construction Safety, Beijing 102616, China

Corresponding author: Xuexin Yang (19801369012@163.com)

This work was supported in part by the Educational Science Research Funds for BUCEA under Grant Y1810, in part by the Open Research Fund Program of Beijing Engineering Research Center of Monitoring for Construction Safety under Grant 202103, and in part by the 2020 Graduate Education and Teaching Quality Improvement Project of BUCEA under Grant 2020013.

ABSTRACT The purpose of sentiment classification is to solve the problem of automatic judgment of text sentiment tendency. In the sentiment classification task of online reviews, traditional deep learning sentiment classification models focus on algorithm optimization to improve the classification performance of the model, but when the sample data for manually labeling sentiment tendencies is insufficient, the classification performance of the model will be poor. The deep learning sentiment classification model based on weak tagging information, on the one hand, introduces weak tagging information into the training process of the model to reduce the use of manually tagging data. On the other hand, weak tagging information can represent the sentiment tendency of reviews to a certain extent, but it also contains noise, the model reduces the negative impact of the noise in weak tagging information in order to improve the classification performance of the sentiment classification model. The experimental results show that in the sentiment classification task of hotel online reviews, the deep learning sentiment classification model based on weak tagging information has superior classification performance than the traditional deep model without increasing labor cost.

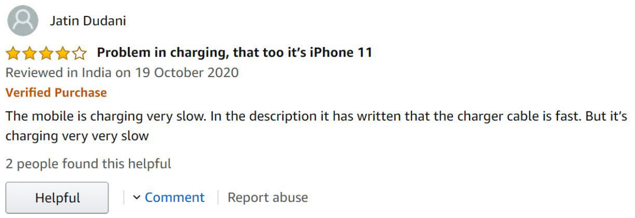
INDEX TERMS Deep learning, sentiment classification, weak tagging information, imbalanced classification.

I. INTRODUCTION

The development of the Internet has given users a platform for freely posting online reviews. Online reviews include product reviews, movie reviews, and service reviews. These reviews contain sentiment information that people want to express. Sentiment information in online reviews can help companies improve their products, help the government monitor public opinion, and provide references for other users. Mining sentiment information from review text requires sentiment analysis technology, and one of the basic tasks of sentiment analysis technology is sentiment classification. The process of automatically discriminating the sentiment tendency of the text is sentiment classification.

Traditional deep learning models focus on algorithm optimization to improve the classification performance of the model and perform well in sentiment classification tasks. However, the training process of traditional sentiment classification model based on deep learning requires large-scale manually tagging samples of sentiment tendency, which will consumes a huge amount of human resources. For reasons of clarity, the sample data of artificially tagging sentiment tendency is defined as tagging data below. In the face of massive review data, it is difficult to manually tagging the sentiment tendency of each review sample. The online reviews of many platforms not only contain the review text, but also some other information such as scores, emojis, labels, etc. These information can be regarded as the user's tagging of their own reviews text. Introducing this information into the model training of deep learning model can reduce the use of tagging data. However, there is no uniform standard for users to post reviews, as some reviews are arbitrary. The sentiment tendency of information tagging such as rating is inconsistent with the sentiment expressed in user's review text (such as negative review text with high score). As shown in Fig 1, a user posted a negative review text on an iPhone 11 purchased on an e-commerce platform, but gave a positive score of 4 stars, which is called noise. Because of the noise in scores, emoticons, labels and other information, we define this information as weak tagging information. The sample

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



Jatin Dudani

★★★★☆ **Problem in charging, that too it's iPhone 11**

Reviewed in India on 19 October 2020

Verified Purchase

The mobile is charging very slow. In the description it has written that the charger cable is fast. But it's charging very very slow

2 people found this helpful

FIGURE 1. Example of noise in weak tagging information.

data containing weak tagging information is defined as weak tagging data. The samples whose weak tagging information and the sentiment tendency expressed by the review text are inconsistent are defined as noise samples, and the samples whose weak tagging information and the sentiment tendency expressed by the review text are consistent are defined as the correct samples. The difference between weak tagging data and tagging data is that weak tagging data contains noise samples, while tagging data does not contain noise samples.

Traditional researchers usually equate the use of weak tagging data with the use of tagging data, that is, only using weak tagging data to train sentiment classification model. This solves the problem of insufficient tagging data, but due to the presence of noise samples in weak tagging data, noise samples will have a negative impact on the model during model training and reduce the classification performance of the model. Therefore, to reduce the negative impact of noise samples on the model while introducing weak tagging information into model training is difficult to accomplish.

A small number of samples are extracted from all samples to manually tag the sentiment tendency, and a small number of tagging data and massive weak tagging data composed of the remaining samples are obtained. The following two methods are proposed to train sentiment classification models using tagging data and weak tagging data to reduce the negative impact of noise samples in weak tagging data on the model, thereby improving the classification performance of the model.

(1) The training of sentiment classification model is divided into two stages. In the first stage of training massive weak tagging data is used to train the model, and then some of tagging data is used to continue training the model to fine-tune the model in the second stage.

(2) Firstly, the neural network noise reduction model is trained by using some of tagging data and the original weak tagging data corresponding to these tagging data, which is used to denoise the weak tagging data. Then, massive weak tagging data are denoised by noise reduction model, and the output of noise reduction model is used as the input of sentiment classification model to train sentiment classification model.

Experiments show that the two methods proposed in the process of introducing weak tagging information to participate in model training can effectively reduce the negative impact of the noise contained in weak tagging information on the sentiment classification model, thereby improving the sentiment classification performance of the model.

The structure of this article is as follows: Section 2 introduces some work related to this article in detail; Section 3 introduces the deep learning sentiment classification model based on weak tagging information proposed in this article; Section 4 gives the experimental results of the sentiment classification model comparative analysis with classification performance; Section 5 summarizes this article and proposes a prospect for the future.

II. RELATED WORK

A. SENTIMENT CLASSIFICATION

At present, sentiment classification is mainly divided into two research directions: sentiment classification based on dictionary and sentiment classification based on machine learning [1]–[3]. References [4]–[7] each proposed a new sentiment dictionary. Experiments show that the classification performance in sentiment classification tasks is superior to traditional sentiment dictionaries, and classification performance of sentiment classification method based on dictionary is excellent. However, sentiment dictionary is domain dependent. Once the application field of the sentiment classification task changes, the classification performance of the classification model based on the sentiment dictionary would decrease. Reference [8] proposed to use multi-domain data to create an sentiment dictionary to solve the domain dependency problem of sentiment dictionary. Although the domain dependence of sentiment dictionary was weakened, the largest problem of sentiment classification model based on dictionary is that the construction of sentiment dictionary needs much human participation, Moreover with the explosive growth of network data, it is difficult to solve the problem of unknown words in sentiment dictionary by manually.

Reference [9] proposed to use machine learning technology to complete the task of sentiment classification. Experiments show that the sentiment classification model based on machine learning has excellent classification performance. Reference [10]–[12] each proposed a new machine learning algorithm to complete sentiment classification tasks and achieved high classification performance. Compared with the sentiment classification model based on dictionary, sentiment classification model based on machine learning avoids the problem of unknown words, but the feature engineering of traditional machine learning algorithm still requires high labor costs. As a branch of machine learning, deep learning has developed rapidly in recent years, and has performed well in sentiment classification tasks of large-scale texts such as online reviews. Deep learning technology greatly reduces labor costs compared to machine learning technology. Word2vec, ELMO [13], BERT [14] promoted the application of deep learning in sentiment classification tasks. References [15]–[17] have proposed innovative deep learning models for sentiment classification. Experiments show that the classification performance is better than the traditional deep learning models. However, most of the data sets used by these models are standard data sets for artificially tagging sentiment tendency, and the lack of large-scale tagging data for supervised training of deep learning models is the application bottleneck of deep learning models.

B. WEAK TAGGING DATA

Introducing weak tagging data into the training process of deep learning models can solve the problem of insufficient training data. At present, there are relatively few researches on weak tagging data. Reference [18] in Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance, with weak tagging data is introduced in the task to complete the classification task. Reference [19] has tagged 1.6 million Twitter texts and built an emoji sentiment dictionary. The experiment verified the effectiveness of emoji, which was weak tagging information, in sentiment classification tasks. Reference [20] proposed two big data systems to use emoticons to complete the sentiment classification task of Twitter text. The experiment proved that the accuracy and robustness character of the two proposed systems were excellent. Reference [21] introduced weak tagging data in training of deep learning models, and achieved excellent classification performance in sentiment classification tasks. In sentiment classification tasks, these researchers all use weak tagging data as equivalent to tagging data, but ignore the characteristics of noise contained in weak tagging data. Therefore, while using weak tagging data, this paper proposes two methods to reduce the negative impact of the noise samples contained in weak tagging data on the sentiment classification model, thereby improving the classification performance of the sentiment classification model.

III. METHODS

A. DATA CLEANING

The sample data used in this article is the review data of Beijing Express hotels crawled from the hotel website. The weak tagging information used is the score given to the hotel by the user while publishing the hotel text review. According to the score system, the full score of each review is 5. The reviews with score more than 2.5 points are classified as positive sentiment tendency, and those with score less than or equal to 2.5 points are classified as negative sentiment tendency. A total of 983220 online reviews were crawled, and the sentiment tendency distribution after dividing all the data according to weak tagging information is shown in Fig 2:

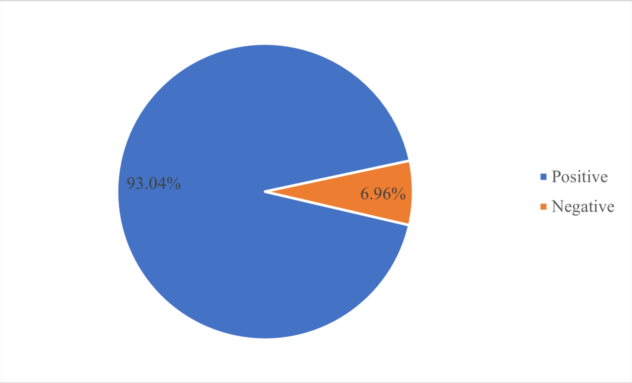


FIGURE 2. The number distribution of sentiment tendencies.

The review data often contains some information that is not helpful for subsequent sentiment classification tasks. The use of technologies related to data cleaning can improve the quality of data, thereby improving the classification performance of subsequent sentiment classification models. The specific data cleaning steps are shown in Fig 3:

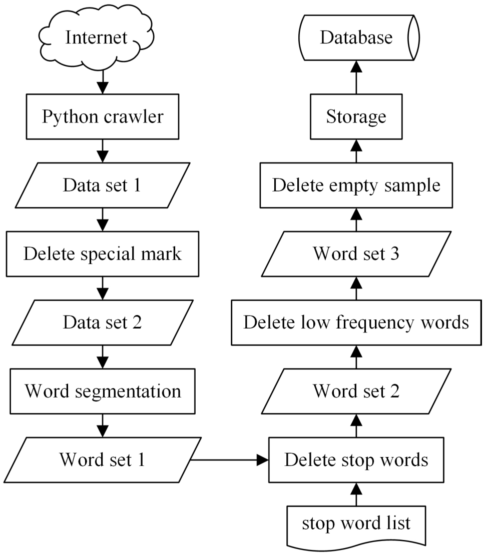


FIGURE 3. Data crawling and cleaning.

Delete special mark: Because there is no uniform rule when users post their own reviews on the website, they often include some special marks in the text reviews that have nothing to do with the sentiment tendency of the sample.

Word segmentation: Chinese online reviews posted by users are in the form of word sequences. Word segmentation refers to dividing the word sequence of each online review into multiple individual words.

Delete stop words: Stop words refer to words that have no semantic meaning for the entire online review, such as interjections, pronouns, etc. The removal of stop words and removal of special signs have the same effect, which can help the subsequent sentiment classification model to better capture the main semantics of the review sentence.

Delete low frequency words: due to the large number of Internet users, there are various words for each person to express their sentiments. Although some words can represent certain sentiment tendency, their frequency is too low. Learning the sentiment tendency of low frequency words has more disadvantages than advantages for the whole sentiment classification model.

After that, Delete the empty reviews. Finally, the data is stored in the database to facilitate the use of subsequent sentiment classification model.

In the experiment, a small number of samples are first extracted from all samples to manually tag the sentiment tendency. According to the definition of noise samples given in Section 1, Fig 4 shows the distribution of noise samples and correct samples before the sentiment tendency of the samples is manually tagging.

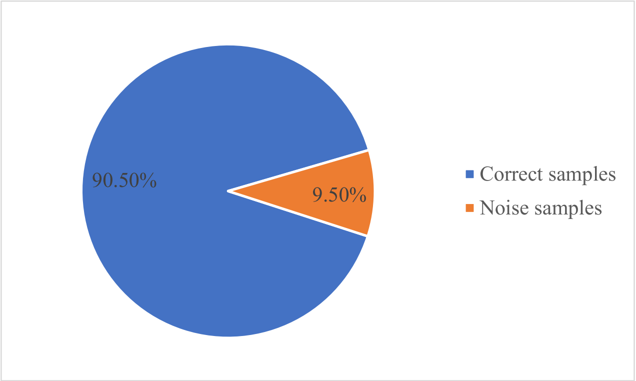


FIGURE 4. Data distribution of noise samples and correct samples.

A total of 19233 samples are tagged tagging in the experiment, including 1828 noise samples accounting for 9.50%. The proportion of noise samples in the weak tagging data is not very large, so the weak tagging data can represent the user's true sentiment tendency in most cases. However, this part of the noise samples would nonetheless have a great negative impact on the sentiment classification model. Therefore, if we want to further improve the classification performance of the sentiment classification model, we should not ignore the existence of noise samples.

B. SENTIMENT CLASSIFICATION MODEL BASED ON BiLSTM

Long Short-Term Memory (LSTM) [22] neural network is a recurrent neural network that introduces a "gate" mechanism. The LSTM neural network can capture the semantic dependence of a longer distance, and avoid the vanishing gradient problem of the traditional recurrent neural network due to the long sequence. Fig 5 shows the distribution of the number of words contained in all the review samples. Statistics show that most of the review samples contain fewer words, and the number of samples within 50 words accounts for 98.51% of the total sample number. Therefore, LSTM is suitable for the sentiment classification task of hotel reviews.

The calculation process of LSTM is given in (1)~(6). σ refers to the Sigmoid activation function, $\odot$ refers to the dot product operation between the weight matrices, $\{W_*, U_*, V_*\}_{*\in\{i,f,c,o\}}$ refers to the parameter set in the LSTM unit; $x_t, c_t, i_t, f_t, o_t, \tilde{c}_t, h_t$ They respectively represent the input of the LSTM unit at time t, the cell state, the value of the input gate, the value of the forget gate, the value of the output gate, the state of the candidate cell, and the output of the LSTM unit.

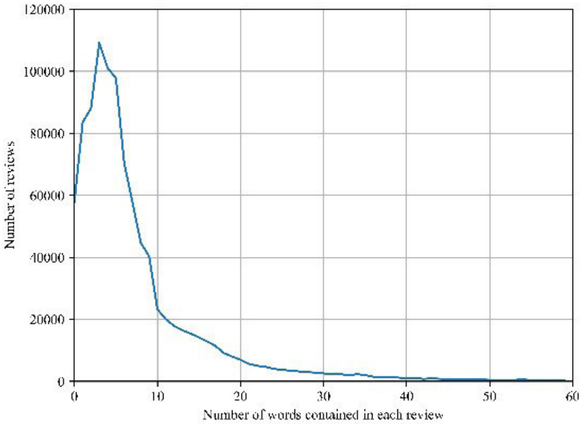


FIGURE 5. Distribution of the number of words contained in the review samples.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1}) \tag{1}$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1}) \tag{2}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1}) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t) \tag{5}$$

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

LSTM can understand text semantics as a whole and performs well in sentiment classification tasks [23]. However, LSTM can only capture the semantic dependence of a single direction. Bi-directional Long Short-Term Memory (BiLSTM) [24] neural network is an improved LSTM neural network that can capture bidirectional long-distance semantic dependencies. Therefore, in the experiment, the BiLSTM neural network is mainly used as the basic component in the deep learning sentiment classification model to obtain superior classification performance. Fig 6 shows the structure of the BiLSTM neural network.

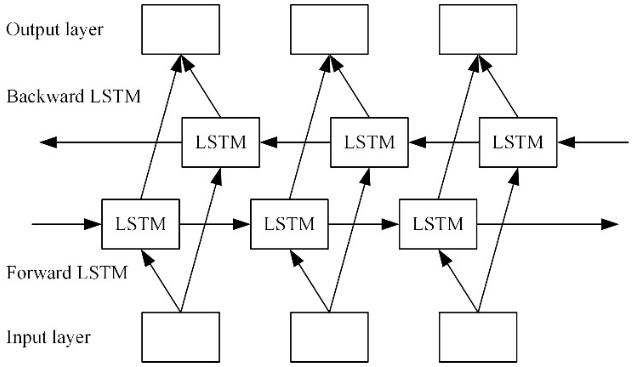


FIGURE 6. Structure diagram of BiLSTM.

Word2vec is used to convert the text information of the review data into a vector representation, which is then input

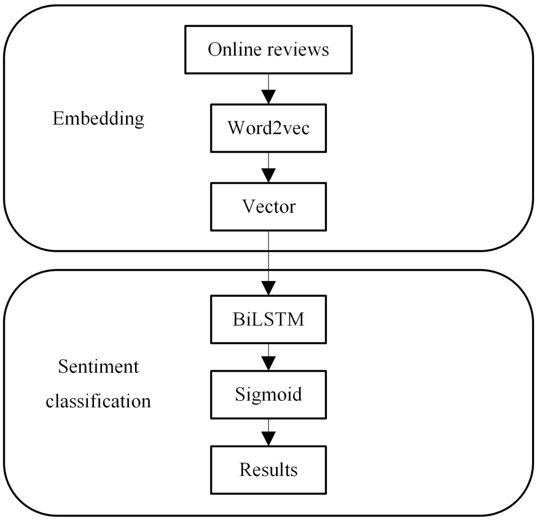


FIGURE 7. Structure diagram of sentiment classification model based on BiLSTM.

into BiLSTM. Final sample sentiment classification result is obtained by passing BiLSTM to the Sigmoid layer. The specific sentiment classification model is shown in Fig 7:

C. BILSTM SENTIMENT CLASSIFICATION MODEL BASED ON TWO-STAGE TRAINING

The BiLSTM sentiment classification model based on two-stage training trained in two stages is the first deep learning sentiment classification model based on weak tagging information. The structure of the sentiment classification model used in the experiment is consistent with the sentiment classification model based on BiLSTM (SC_BiLSTM), and only innovations are made in the training method of the model. The training of the model is divided into two stages, and a schematic diagram of the two-stage training method is shown in Fig 8.

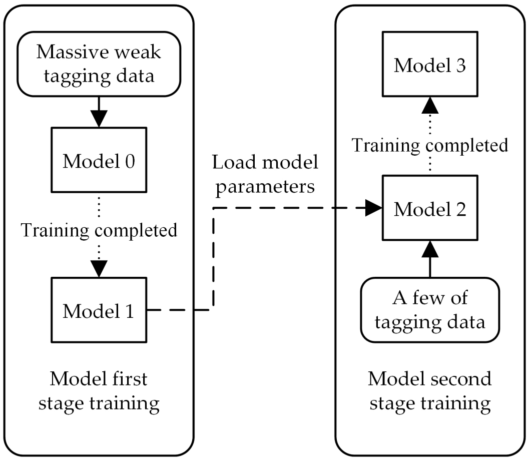


FIGURE 8. Schematic diagram of two-stage model training.

First, massive weak tagging data is used to train the SC_BiLSTM (model 0) in the first stage. In Fig 4 of Section 3.A, it can be seen that the proportion of noise samples in weak tagging data is small. Therefore, after training model 0 with massive weak tagging data, we can get a sentiment classification model (model 1) is gained which can capture the review semantics well. After that, the model 1 parameters are used as the initial parameters of the second stage model (model 2) training, and some of tagging data is input into the model for training, so as to fine-tune the model parameters, with the aim of reducing the negative impact of the noise samples input in the first stage of model training on the sentiment classification model, improving the classification performance of the model, and obtaining the final sentiment classification model (model 3). Table 1 shows the training process of the BiLSTM sentiment classification model based on two-stage training.

TABLE 1. BiLSTM sentiment classification model based on two-stage training.

```
Algorithm 1: BiLSTM sentiment classification model based on two-stage training
SC_BiLSTM of the first stage training.
Input: Massive weak tagging data after data cleaning.
Begin:
(1) Word2vec in SC_BiLSTM converts the text of the review into a vector representation.
(2) The parameters of the SC_BiLSTM model are optimized using equation (1-6).
(3) Save the model parameters after the SC_BiLSTM model converges.
Output:  Save the SC_BiLSTM model (Model 1) that has completed the first stage of training.
SC_BiLSTM of the second stage training.
Input: Some of tagging data after data cleaning.
Begin:
(1) Load the parameters of Model 1 to get Model 2.
(2) Word2vec in SC_BiLSTM converts the text of the review into a vector representation.
(3) The parameters of the SC_BiLSTM model are optimized using equation (1-6) again.
(4) Save the model parameters after the SC_BiLSTM model converges again.
Output:  Save the SC_BiLSTM model (Model 3) that has completed the two-stage training.
```

D. SENTIMENT CLASSIFICATION MODEL BASED ON DENOISING OF WEAK TAGGING DATA

The sentiment classification model based on denoising of weak tagging data is the second deep learning sentiment classification model based on weak tagging information. A deep learning model is constructed for denoising weak tagging data. The denoising model of weak tagging data based on deep learning uses the text information and tagging information of the tagging data as output, and uses the text information and weak tagging information of the corresponding weak tagging data before manually tagging sentiment tendency as the input for the denoising model of weak tagging data based on deep learning model training. Fig 9 shows the structure of the weak tagging data denoising model based on deep learning.

We use Word2vec to construct the review text into a vector representation (Vector_x), and convert the weak tagging

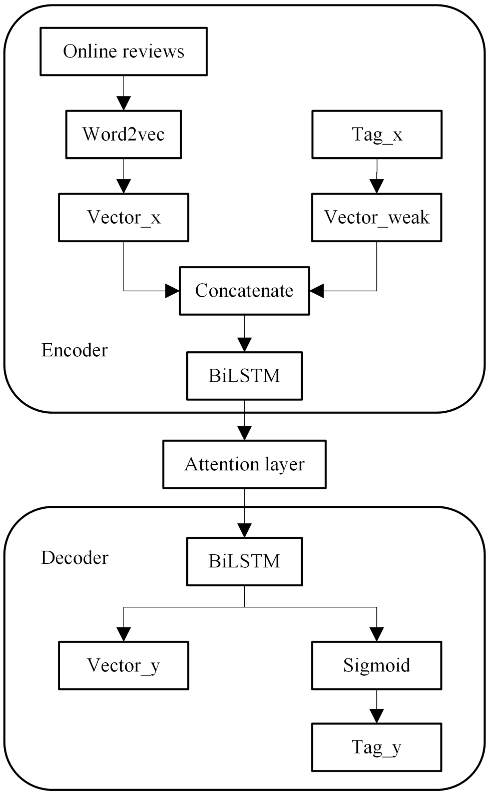


FIGURE 9. Structure diagram of weak tagging data denoising model based on deep learning.

information (Tag_x) of the corresponding review text into a vector (Vector_weak) with the same dimension as Vector_x. If the weak tagging information represents a positive sentiment tendency, the elements in Vector_weak are all 1. If the weak tagging information represents a negative sentiment tendency, the elements in Vector_weak are all 0. We join Vector_x and Vector_weak through the Concatenate layer and then input them into the BiLSTM layer to complete the Encode part of the weak tagging data denoising model based on deep learning.

Through the attention layer between Encode and Decode, the attention mechanism is used to improve the performance of the deep learning model. In (7)~(9), the specific calculation steps of the attention mechanism are given. Where T_x is the sequence length; c_t is the semantic vector of t; e_{ti} is the alignment model, representing the degree of influence of the hidden layer state h_i of the BiLSTM in the Encoder at time i on the hidden layer state s_t of the BiLSTM in the decoder at time t, calculated by h_i and s_{t-1}; α_{ti} is the attention weight normalized by e_{ti} through softmax.

$$c_t = \sum_{i=1}^{T_x} a_{ti} h_i \tag{7}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T_x} \exp(e_{tk})} \tag{8}$$

$$e_{ti} = a(s_{t-1}, h_i) \tag{9}$$

Entering the Decoder part of the deep learning model, the data is divided into two parts through the BiLSTM layer, one part is output directly as the vector representation of the review text (Vector_y), and the other part is obtained through the Sigmoid layer to get the denoised tagging information (Tag_y).

The sentiment tendency of the data is manually tagged, without changing the text information of the review sample. Since only the tag of the noise sample is modified, the input of the weak tagging data denoising model based on deep learning during the training process and the review text information in the output remain the same, that is, Vector_x same as Vector_y, both are word vectors converted from the review text by Word2vec. In the training process, the input and output of the weak tagging data denoising model based on deep learning have different tagging information. The output is the tagging information of the tagging data, and the output is the weak tagging information of the tagging data before the sentiment tendency of the review sample is manually tagging. Therefore, Tag_x and Tag_y is different.

After the weak tagging data denoising model based on deep learning training is completed, the model has the ability to reduce noise on weak tagging data. When using this model, once weak tagging data are input, output data with reduced noise are obtained. During the training process, the text information and tagging information interacted, and the text information is adjusted in the process of denoising weak tagging information. Therefore, when the model is used, Vector_x and Vector_y are different. Vector_x is still the vector representation of the original review text, and Vector_y is the vector representation of the review text after considering the denoising of weak tagging data.

After the training of the denoising model for weak tagging data based on deep learning is completed, the subsequent sentiment classification model based on BiLSTM (DN_BiLSTM) is constructed to complete the sentiment classification task. Fig 10 shows the structure of the entire sentiment classification model based on denoising of weak tagging data. The "DN_Model" in Fig 10 represents the trained denoising model of weak tagging data based on deep learning. Input the text information and weak tagging information of massive weak tagging data into the DN_Model,

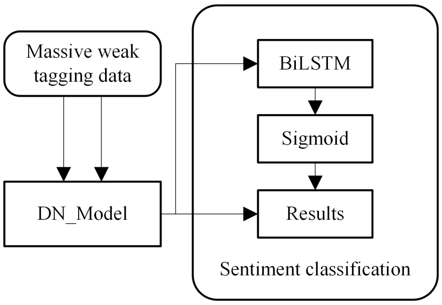


FIGURE 10. Structure diagram of the sentiment classification model based on denoising of weak tagging data.

and output the denoised text vector representation and tagging information. The denoised text vector representation is used as the input of the DN_BiLSTM, and the denoised tagging information is used as the output of the DN_BiLSTM to complete the training of the DN_BiLSTM. Thus, the training of the whole sentiment classification model based on denoising of weak tagging data is completed.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP

Two experiments were carried out in this paper. The purpose of the first experiment is to select the noise reduction model with the best noise reduction performance. The second experiment is to select the sentiment classification model with the best classification performance.

In the first experiment, in addition to the DN_Model model in Fig 9 in Section 3.D, two control groups were set up with this model as the template. Control group 1 changed all the BiLSTM layers in the DN_Model model of Fig 9 to BiGRU [25] layer, and control group 2 changed all the BiLSTM layers in the DN_Model model of Fig 9 to Text-CNN [26] layer. The parameters of BiLSTM and BiGRU used by DN_Model are the same. The parameters of BiLSTM and BiGRU are given in Table 2.

TABLE 2. Parameter table of BiLSTM and BiGRU in DN_Model.

Parameter	Parameter value
Input vector length	50
Output dimension	128
Dropout	0.2
Epochs	50

Table 3 shows the parameter settings of Text-CNN.

TABLE 3. Parameter table of Text-CNN in DN_Model.

Parameter	Parameter value
Input vector length	50
Output dimension	128
Number of convolutional layers	3
Convolution kernel size	3; 4; 5
Dropout	0.2
Epochs	50

Perform data cleaning is performed. The steps of data cleaning are shown in Fig 3 in 3.A. In data cleaning, use jieba segmentation to segment the review text, and use the HIT stop words list to remove the stop words. A total of 983,220 hotel online review samples have been crawled, and 19,233 samples are extracted from them to manually tagging the sentiment tendency. Thirty percent of the 19233 manually tagging data are selected as the test set of all experimental groups.

An experiment group (BiLSTM_Tag) is set up that uses only some of tagging data to train the sentiment classification model based on BiLSTM. An experiment group (BiLSTM_FullT) is set up that mixes some of tagging data and massive weak data to train the sentiment classification model based on BiLSTM. The first stage of the BiLSTM sentiment classification model based on two-stage training uses massive weak tagging data for training (BiLSTM_Weak), and the second stage uses some of tagging data to continue training on the basis of BiLSTM_Weak (BiLSTM_Continue). BiLSTM_Tag, BiLSTM_FullT, BiLSTM_Weak, BiLSTM_Continue use the same model structure, and they are different only in the model training data.

The training of the sentiment classification model based on denoising of weak tagging data includes the training of two models. First we train the denoising model of weak tagging data based on deep learning (DN_Model), and then train the sentiment classification model (DN_BiLSTM). The output of DN_Model contains the denoised tagging representation, which is also the result of sentiment classification. Therefore, it is also necessary to test the sentiment classification performance of DN_Model on the test set.

Table 4 shows the number distribution of each sentiment tendency of the samples in training set and test set of each model.

TABLE 4. The distribution of the number of sentiment tendencies in each data set.

Data set	Positive	Negative
Training set of BiLSTM_Tag	11357	2106
Training set of BiLSTM_FullT	909945	67505
Training set of BiLSTM_Weak	898588	65399
Training set of BiLSTM_Continue	11357	2106
Training set of DN_Model	11357	2106
Training set of DN_BiLSTM	898588	65399
Test set	4838	932

All models in the experiment are built using TensorFlow, an open source framework for deep learning. Since the number of samples within 50 words accounts for 98.51% of the total number of samples, the length of the vector output by Word2vec is set to 50. The hyperparameter settings of BiLSTM neural network used in sentiment classification model are shown in Table 5:

Use F1, geometric mean (G-mean) [27]–[29], and accuracy are used as the evaluation index of model classification performance. F1 and accuracy rate are commonly used classification evaluation indicators. From Fig 2 in Section 3.B, it can be seen that there is a large gap between the number of samples of positive and negative sentiment tendencies.

TABLE 5. Hyperparameter settings of BiLSTM neural network used in sentiment classification.

Parameter	Parameter value
Input vector length	50
Output dimension	128
Dropout	0.2
Epochs	20

Therefore, the sentiment classification of hotel online reviews is an imbalanced classification task, while in the imbalanced classification task, G-mean is Classical evaluation index. In (10) and (11), the calculation methods of F1 and G-mean are given respectively.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{10}$$

$$G - mean = \sqrt[n]{\prod_{i=1}^{n} Recall_i} \tag{11}$$

$recall_i$–in (12) represents the recall rate of category i. The experiment only completes the binary classification task of sentiment classification of hotel online reviews, so the value of n is 2.

B. EXPERIMENTAL RESULTS

Table 6 shows the performance comparison of the three noise reduction models, where the bold numbers are the optimal values of each evaluation index.

TABLE 6. Noise reduction performance comparison.

Method	F1	G-mean	Accuracy
BiGRU	0.456	0.0	0.838
Text-CNN	0.830	0.787	**0.915**
BiLSTM	**0.833**	**0.809**	0.913

It can be concluded that BiGRU has poor classification performance because it is greatly affected by the unbalanced data distribution. Compared with Text-CNN, BiLSTM has basically the same F1 and Accuracy, but BiLSTM is 0.022 higher than Text-CNN on the G-mean index. On the whole, DN_Model with BiLSTM as the component has the best noise reduction performance. Therefore, the DN_Model used in the subsequent experiments is constructed with BiLSTM.

Table 7 shows the sentiment classification performance of all experimental groups, where the bold numbers are the optimal values of each evaluation index.

Compare the classification performance of BiLSTM_Tag, BiLSTM_FullT, and BiLSTM_Weak, and make a histogram as shown in Fig 11:

TABLE 7. Comparison of classification performance of each experiment group.

Group	Method	F1	G-mean	Accuracy
1	*BiLSTM-Tag*	0.803	0.745	0.905
2	*BiLSTM_FullT*	0.708	0.568	0.886
3	*BiLSTM_Weak*	0.708	0.567	0.887
4	*BiLSTM_Continue*	0.841	0.816	0.918
5	*DN_Model*	0.833	0.809	0.913
6	*DN_BiLSTM*	**0.859**	**0.850**	**0.924**

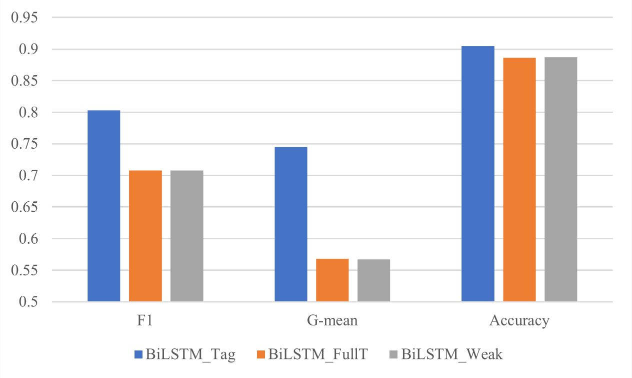


FIGURE 11. Comparison of classification performance of experiment groups (1), (2), (3).

It can be seen from Fig 11 that all the evaluation indicators BiLSTM-Tag in the three experiment groups are the highest. In the experiment, BiLSTM_Tag training only uses tagging data, while BiLSTM_FullT and BiLSTM_weak both use weak tagging data in training. The training set of BiLSTM_FullT not only contains the weak tagging data also contains the tagging data used for BiLSTM-Tag training. Experiments have proved that if the weak tagging data is directly used in deep learning model training without processing, it will cause the performance of the classification model to decrease. In addition, the evaluation index values of BiLSTM_FullT and BiLSTM_weak are basically the same, which proves that mixing some of tagging data with massive weak tagging data during training does not play the role of tagging data. The accuracy of BiLSTM_weak reached 0.887, and the F1 reached 0.708, which proves that the sentiment classification model trained on massive weak tagging data has been able to capture the sentiment tendency of the review text well.

Fig 12 shows the classification performance comparison of BiLSTM_Tag, BiLSTM_weak, and BiLSTM_Continue. The classification performance of BiLSTM_Continue is more satisfactory higher than BiLSTM_weak, which proves that the Bilstm sentiment classification model based on two-stage training can effectively use some of the tagging data. BiLSTM_Continue has the highest evaluation indicators in these three experiment groups, which proves that the Bilstm sentiment classification model based on two-stage training uses weak tagging data to participate in deep learning model training, while reducing the negative impact of noise samples in weak tagging data on the model, and achieved superior

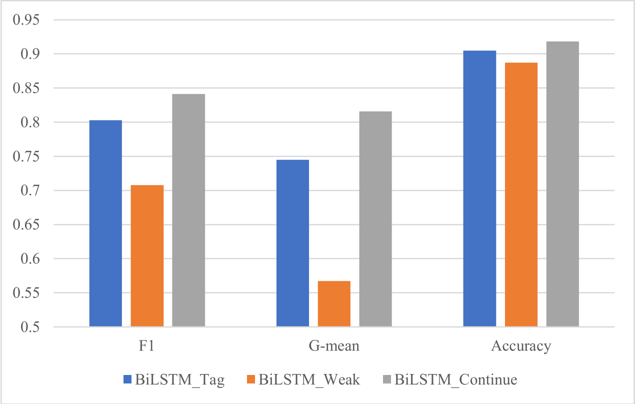


FIGURE 12. Comparison of classification performance of experiment groups (1), (3), (4).

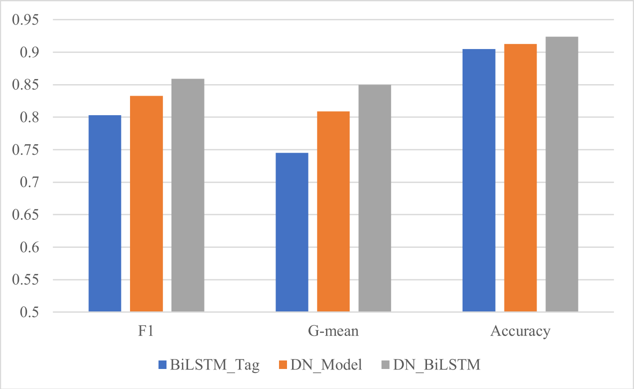


FIGURE 13. Comparison of classification performance of experiment groups (1), (5), (6).

classification performance. Compared with BiLSTM_Tag, BiLSTM_Continue has increased 0.038, 0.071, 0.013 in F1, G-mean, and Accuracy, respectively.

Fig 13 shows the classification performance comparison of BiLSTM_Tag, DN_Model, and DN_BiLSTM. The DN_Model, which also uses some annotation data for model training, has better classification performance than BiLSTM_Tag. The main reason for analyzing this phenomenon is that DN_Model has a more complex neural network structure. The simpler reason for the construction of the sentiment classification model based on BiLSTM is that complex network training takes more time. If using the same complex network structure as DN_Model, it will consume a large amount of time and cost when using massive weak tagging data to train the model. Therefore, in the experiment, only DN_Model with a more complex network structure will be trained with some tagging data. DN_BiLSTM has achieved more positive classification performance than DN_Model. The analysis reason is that on the one hand, massive weak tagging data has more comprehensive review sample characteristics, and on the other hand, after the weak tagging data is denoised by DN_Model, the negative impact of the noise samples in the weak tagging data on the classification performance of the model is reduced. Compared with BiLSTM_Tag, DN_BiLSTM increased by 0.056, 0.105, 0.019 in F1, G-mean, and Accuracy, respectively, and DN_BiLSTM achieved the best values of all experimental groups on all evaluation indicators.

BiLSTM_Continue uses tagging data to offset the negative impact of noise samples on the classification performance of sentiment classification model. This method is relatively simple, while DN_BiLSTM is a denoising method for learning weak tagging data to reduce the negative impact of noise samples on the classification performance of sentiment classification model. Therefore, DN_BiLSTM has superior classification performance than BiLSTM_Continue.

The above results prove that the two proposed deep learning sentiment classification models based on weak tagging information have superior classification performance than the traditional Sentiment classification model based on BiLSTM that only uses tagging data to train.

V. CONCLUSION

This paper proposes two deep learning sentiment classification models based on weak tagging information, which are a BiLSTM sentiment classification model based on two-stage training and a sentiment classification model based on denoising of weak tagging data. The deep learning sentiment classification model based on weak tagging information uses weak tagging data for model training while reducing the negative impact of noise samples in weak tagging data on the classification performance of the sentiment classification model, and improves the classification performance of the sentiment classification model. In the experiment, the deep learning sentiment classification models based on weak tagging information is compared with the traditional sentiment classification model that only uses tagging data for training, and the sentiment classification model that uses weak tagging data for training but equates the weak tagging data with tagging data. The experimental results show that the deep learning sentiment classification model based on weak tagging information compared with other sentiment classification models, in the sentiment classification task of hotel online review data, the classification performance is significantly improved without increasing the labor cost.

In the next step, the deep learning sentiment classification model based on weak tagging information will be applied to the sentiment classification task of more types of online reviews such as e-commerce reviews and Twitter reviews, and will try to use more types of weak tagging information such as emoticons, symbols, etc. to improve the emotion classification performance of the model. The sentiment classification model based on denoising of weak tagging data obtained the highest classification performance of all the experimental groups, but it is also the model with the highest time complexity. Therefore, in the future, we should try to reduce the time complexity of the model while maintaining high classification performance.

REFERENCES

[1] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, ‘‘Target-dependent Twitter sentiment classification,’’ in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol*, vol. 1, Jun. 2011, pp. 151–160.

[2] S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, ‘‘NRC-Canada-2014: Detecting aspects and sentiment in customer reviews,’’ in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 437–442.

[3] D. T. Vo and Y. Zhang, ‘‘Target-dependent twitter sentiment classification with rich automatic features,’’ in *Proc. IJCAI*, 2015, pp. 1347–1353.

[4] C. S. Khoo and S. B. Johnkhan, ‘‘Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons,’’ *J. Inf. Sci.*, vol. 44, no. 4, pp. 491–511, Aug. 2018.

[5] N. Rizun, Y. Taranenko, and W. Waloszek, ‘‘Improving the accuracy in sentiment classification in the light of modelling the latent semantic relations,’’ *Information*, vol. 9, no. 12, p. 307, Dec. 2018.

[6] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, ‘‘Lexicon-enhanced sentiment analysis framework using rule-based classification scheme,’’ *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171649.

[7] A. Al-Saffar, S. Awang, H. Tao, N. Omar, W. Al-Saiagh, and M. Al-bared, ‘‘Malay sentiment analysis based on combined classification approaches and senti-lexicon algorithm,’’ *PLoS ONE*, vol. 13, no. 4, Apr. 2018, Art. no. e0194852.

[8] V. Jha, S. R, P. D. Shenoy, V. K R, and A. K. Sangaiah, ‘‘A novel sentiment aware dictionary for multi-domain sentiment classification,’’ *Comput. Electr. Eng.*, vol. 69, pp. 585–597, Jul. 2018.

[9] B. Pang, L. Lee, and S. Vaithyanathan, ‘‘Thumbs up? Sentiment classification using machine learning techniques,’’ in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, vol. 10, 2002, pp. 79–86.

[10] Y. Wang, ‘‘Iteration-based naive Bayes sentiment classification of microblog multimedia posts considering emoticon attributes,’’ *Multimedia Tools Appl.*, vol. 79, pp. 19151–19166, Mar. 2020.

[11] S. N. Alyami and S. O. Olatunji, ‘‘Application of support vector machine for arabic sentiment classification using Twitter-based dataset,’’ *J. Inf. Knowl. Manage.*, vol. 19, no. 1, Mar. 2020, Art. no. 2040018.

[12] F. Xu, Z. Pan, and R. Xia, ‘‘E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework,’’ *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102221.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ‘‘Deep contextualized word representations,’’ 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘BERT: Pre-training of deep bidirectional transformers for language understanding,’’ 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] R. Fei, Q. Yao, Y. Zhu, Q. Xu, A. Li, H. Wu, and B. Hu, ‘‘Deep learning structure for cross-domain sentiment classification based on improved cross entropy and weight,’’ *Sci. Program.*, vol. 2020, pp. 1–20, Jun. 2020.

[16] M. Wang, Z.-H. Ning, T. Li, and C.-B. Xiao, ‘‘Information geometry enhanced fuzzy deep belief networks for sentiment classification,’’ *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 11, pp. 3031–3042, Nov. 2019.

[17] Li, Liu, Zhang, and Liu, ‘‘An improved approach for text sentiment classification based on a deep neural network via a sentiment attention mechanism,’’ *Future Internet*, vol. 11, no. 4, p. 96, Apr. 2019.

[18] O. Edo-Osagie, G. Smith, I. Lake, O. Edeghere, and B. De La Iglesia, ‘‘Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance,’’ *PLoS ONE*, vol. 14, no. 7, Jul. 2019, Art. no. e0210689.

[19] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, ‘‘Sentiment of emojis,’’ *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144296.

[20] A. Kanavos, N. Nodarakis, S. Sioutas, A. Tsakalidis, D. Tsolis, and G. Tzimas, ‘‘Large scale implementations for Twitter sentiment classification,’’ *Algorithms*, vol. 10, no. 1, p. 33, Mar. 2017.

[21] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, ‘‘Deep convolution neural networks for Twitter sentiment analysis,’’ *IEEE Access*, vol. 6, pp. 23253–23260, 2018.

[22] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] X. Zhu, P. Sobihani, and H. Guo, ‘‘Long short-term memory over recursive structures,’’ in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1604–1612.

[24] M. Schuster and K. K. Paliwal, ‘‘Bidirectional recurrent neural networks,’’ *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[25] K. Cho, B. Van Merrienboer, and C. Gulcehre, ‘‘Learning phrase representations using RNN encoder-decoder for statistical machine translation,’’ 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[26] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

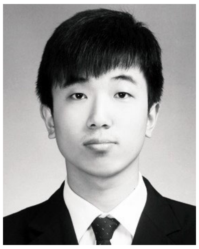
[27] B. Krawczyk, B. McInnes, and A. Cano, ‘‘Sentiment classification from multi-class imbalanced twitter data using binarization,’’ in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.* Cham, Switzerland: Springer, 2017, pp. 26–37.

[28] S. Li, G. Zhou, Z. Wang, S. Y. M. Lee, and R. Wang, ‘‘Imbalanced sentiment classification,’’ in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 2469–2472.

[29] R. Xu, T. Chen, Y. Xia, Q. Lu, B. Liu, and X. Wang, ‘‘Word embedding composition for data imbalances in sentiment and emotion classification,’’ *Cognit. Comput.*, vol. 7, no. 2, pp. 226–240, Apr. 2015.

CHUANTAO WANG was born in Hubei, China, in 1981. He received the B.S. degree in mathematics and applied mathematics from Qiqihaer University, in 2004, the M.S. degree in computational mathematics from Chongqing University, in 2007, and the Ph.D. degree in system engineering from Beijing Jiaotong University, in 2011.

From 2011 to 2016, he worked as an Assistant Professor with the Industrial Engineering Department, Beijing University of Civil Engineering and Architecture. Since 2017, he has been an Associate Professor with the Industrial Engineering Department, Beijing University of Civil Engineering and Architecture. He is the author of two books and more than 30 articles. His research interests include data mining, deep learning, and natural language processing.



XUEXIN YANG was born in Shandong, China, in 1997. He received the B.E. degree in industrial engineering from the Shandong University of Technology, Shandong, China, in 2018. He is currently pursuing the master's degree in industrial engineering with the Beijing University of Civil Engineering and Architecture. His current research interests include data mining, deep learning, and natural language processing.

LINKAI DING was born in Shanxi, China, in 1996. He received the B.E. degree in industrial engineering from the Beijing University of Civil Engineering and Architecture. He is currently pursuing the master's degree in industrial engineering with the Beijing University of Civil Engineering and Architecture. His current research interests include deep learning and natural language processing.