IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 19, 2021, accepted March 27, 2021, date of publication April 1, 2021, date of current version April 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070375

An LSTM&Topic-CNN Model for Classification of Online Chinese Medical Questions

SONG MAO, LU-LU ZHANG, AND ZHEN-GUO GUAN
School of Economics and Management, Shanxi University, Taiyuan 030006, China

Corresponding author: Song Mao (maosong@sxu.edu.cn)

This work was supported in part by Humanities and Social Science Fund in Ministry of Education in China under Grant 18YJC910009, and in part by the Program for the Philosophy and Social Sciences Research of Higher Learning Institutions of Shanxi under Grant 201803050.

ABSTRACT In recent years, people's interest in health question and answer (Q&A) websites has been growing with the development of the internet technologies. How to seek appropriate professional medical information among the massive data has become the focus of all patients. Therefore, it is vital to obtain reasonable predictions and automatic recommendations on the basis of patients' keyword descriptions of their health status and question intention. The key to solving this problem is to achieve automatic text classification of health questions. This paper considered a feature fusion model for the classification of Chinese short texts on medical health Q&A websites by combining the text features and topic features. Firstly, we generated the text word vector by word embedding method and obtained the text features under Long Short-Term Memory (LSTM) model. Given the difficulty in determination of topic numbers, we conducted a sub-sample experiment to obtain the few optimal topic numbers under which the classification performances were good. Then we extracted the topic features and used the one-dimensional convolution idea of the Convolutional Neural Network (CNN) model for topic feature filtering. Finally, we combined the two features together subtly for text classification. Two experiments were conducted to illustrate our model in terms of recall rate, precision, and F1 value when the datasets were from different online medical Q&A websites. Results showed that the LSTM&Topic-CNN model could efficiently enhance the classification effect of Chinese medical health question texts.

INDEX TERMS Chinese medical Q&A text, long short-term memory, latent dirichlet allocation, convolutional neural network, feature fusion.

I. INTRODUCTION

With the rapid growth of intelligent medical field, many online medical question and answer (Q&A) websites have sprung up in recent years [1]. According to patients' intention and concerned illness type, they can search for information, inquire and obtain replying in the most economical and convenient manner. And the answers are provided by different experts who have experienced a sever selection and obtained certification [2]. These websites can not only break the time and geographical restrictions, maximize the integration of various medical information resources, but also has great significance for improving the health awareness and health level of users [3]. With people' continuous concern about health, there are massive Q&A data existing on these websites, how to effectively seek out the effective information is becoming a challenge especially for patients who are lack of professional medical knowledge. The key to solving this problem is to achieve the automatic question classification about information provided by experienced patients.

Numerous studies on medical Q&A websites mainly have considered text classification from the perspective of intent analysis or topic analysis. Intent classification explores the health demands for users based on intent theory, while topic classification concerns over topic features of health information communication. For example, suppose that "I suffer from heart disease, what should I pay attention to in the daily diet?" It is clear that this question focuses on suggestions about dietary among heart disease patients in the intent direction, and belongs to the topic of disease management based on the topic analysis. These researchers mainly have adopted feature extraction, feature enhancement, or latent semantic

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

analysis for text classification, but few have considered it in the terms of feature fusion. Besides this, it is a challenge for scholars to determine the optimal topic number during the process of LDA. And researchers often try the numbers for topic feature, which is somewhat subjective and local optimal.

Hence, we establish an LSTM&Topic-CNN model based on the fusion of text semantic features and topic features and apply it to the classification of Chinese medical text. We also determine the optimal five numbers for topic features from a sub-sample sensitivity analysis so as to guarantee the stability of the experimental results. The main contributions of this work are summarized below:

(1) Based on the perspective of feature fusion, we analyze the medical Q&A data by deeply integrating the text features and topic features to obtain rich semantic information and then to improve text feature representation ability in medical Q&A short text.

(2) Given the difficulty in determination of topic number during the procedure of LDA, we conduct a sub-sample experiment to obtain the few optimal topic numbers with best classification effect. In order to avoid information redundancy, we implement a further CNN-based feature filtering to achieve dimension reduction and improve the efficiency.

The rest of the paper is structured as follows. We review the development of medical Q&A classification and the corresponding deep learning methods briefly in section 2. Then we develop the LSTM&Topic-CNN model in section 3, and conduct two experiments for the analysis of question data from different medical Q&A websites in section 4. And finally, the concluding remarks are presented in section 5.

II. RELATED WORKS AND METHODS

A. RELATED WORKS

Researches on medical text classification have mainly focused on the intention of health behaviors and the content of health information. They either analyzed the purpose of users based on the theory of intention, or explored health demands for users by topic analysis.

For example, based on the technology of Apache Kafka and Apache Spark, Ahmed *et al.* [4] established a real-time system to identify heart disease according to patients' symptoms. Given attention mechanism, Wu *et al.* [5] considered a multi-task model for the analysis of Chinese online medical questions, and then set up a Q&A system about cardiovascular disease. Chen *et al.* [6] compared the performances of four intent classification models for Chinese medical question data. Aiming at improving the query intent classification, González-Caro and Baeza-Yates [7] considered an automatic classification model from the perspective of various facets. Tian *et al.* [8] extracted two datasets from large-scale high-quality Chinese medical Q&A corpus and then predicted the correlation between the question and answer and adopted answer among the given multiple answers.

In terms of topic classification, Schmidt *et al.* [9] adopted Transformer-based classification methods to solve the problem of ambiguity text during prediction process, and implemented optimal named entity recognition method to construct an automatic classification system, when the data were extracted from four topics in Clinical Trial Text. By analyzing Q&A data about diabetes from Yahoo Answers, Zhang and Zhao [10] utilized visualized method to present twelve concerned topics for diabetics. Li *et al.* [11] used LDA model and manual tagging method to determine rules of topic coding, and then discussed the topic features about cancer for questions from Baidu Knows.

Furthermore, many authors considered Chinese medical Q&A corpus by content analysis or deep matching neural networks. Chiu and Wu [12] explored consumers' expression about their health information demands in respect of action, cognition, and emotion. Through the use of text mining technology, Lu *et al.* [13] sought hotspot by social media and analyzed their sentiment expression from the perspective of health care stakeholder. He *et al.* [14] developed matching neural networks for the analysis of extensive Chinese medical Q&A corpus. In order to improve the results of Chinese word segmentation, they firstly adopted the semantic cluster method to generate a new representation.

B. RELATED METHODS

As an effective technology to organize and manage complex text information, text classification has been extensively used in natural language processing (NLP), machine learning, web search, spam filtering, emotional analysis and other fields [15].

Usually, traditional machine learning methods and single deep learning methods have been adopted to achieve the goal of text classification at present. Ge [16] utilized Word2Vec method and integrated traditional classifiers to improve text classification effect. Zhang *et al.* [17] adopted a new method which combined the Word2vec and SVMperf to train and classify the Chinese comment texts. Edara *et al.* [18] developed a LSTM model to analyze the sentiment and text categorization, when the data were collected from cancer medical records. Wang *et al.* [19] proposed a word2vec and LDA based hybrid approach to improve classification performance through generating word-context relationships and document-topic relationships. Liu *et al.* [20] adopted LSTM and factory-aware attention mechanism to construct an air pollutant predictive model. Based on the technique of deep neural network, Jiang *et al.* [21] considered a word representation model by combining topic information and word order information.

Due to the limitations of common model in text classification, more and more scholars have tried to combine multiple models to improve the text classification effect. Based on the theory of CNN, bidirectional gate control unit, highway network and full-connection layer, Liu *et al.* [22] extracted the global and the local textual semantics by a short text character-level model quickly and effectively. For the purpose of strengthening the relationship between words and texts, Luo [23] adopted a hybrid method by combing LDA, CNN with gated recurrent unit (GRU) to classify the texts. To further improve the important features' effect of Chinese texts, Xie *et al*. [24] proposed a fusion model to enhance texts by feature combining double-layer LSTM network with CNN. Based on the deep neural network, Tang *et al*. [25] conducted an entity recognition under attention-based CNN-LSTM-CRF model to obtain the local context information on interesting words of Chinese clinical texts.

Recently, medical Q&A and text classification methods have been greatly improved, however, recent studies rarely considered this issue from the perspective of feature fusion. It was noteworthy that Liu *et al*. [26] proposed a new method for Chinese question classification from the perspective of ERNIE and feature fusion. They firstly developed a generalized language representation model by integrating knowledge, then adopted the technology of Highway-CNN and Highway-DCU-BiLSTM to extract local features and sequence features separately, and finally integrated the two features by a linear formula. Obviously, Liu *et al*. focused on text feature fusion, however, they neglected the utilization of topic information.

In fact, topic information has always been crucial for analyzing medical Q&A text. Ignoring it would cause an inaccurate representation of the original text and even lead to an increase in misclassification. Considering this, we have established an enhanced feature fusion model by blending multi-granular topic features and text features together and have applied it to the classification of Chinese medical Q&A. And we hope our study enriches the research field of medical Q&A to a certain extent, hoping to provide convenience for patients to find precision accommodation, produce an increasing amount of traffic, and even a modest profit for a medical Q&A website.

III. LSTM&TOPIC-CNN MODEL

LSTM&Topic-CNN model is presented in this section. This section adopts the consecution of chief-part-chief in structure, which means to introduce a brief model framework firstly, then present basic methods involved respectively, and finally work out the concrete algorithm and visualized procedures for this model. Aiming at the short text classification task in medical field, the word embedding method is implemented to obtain the word vector representations, which are encoded to extract the text features through the LSTM neural network. Meanwhile, techniques of LDA and CNN are utilized in the extraction of latent topic features. Then the text feature and the topic feature are combined to obtain the fusion features, which further ensure the accuracy and integrity of text feature extraction. The structural framework is shown in Fig.1.

A. TEXT FEATURE EXTRACTION

For text data, the existing machine learning methods need to convert it into numerical data firstly, which leads to the concept of text feature representation. The accuracy and comprehensive expression of the text feature representations are closely related to the effect of text classification.

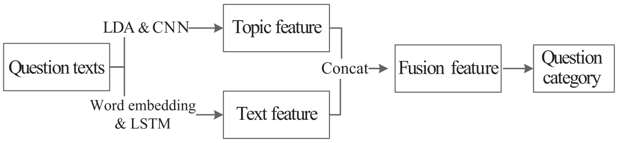

FIGURE 1. Question text classification framework.

Traditionally, the one-hot coding method [27] is used to deal with discrete features of text. However, due to the neglectful order between words, it may generate high-dimensional, sparse, and discrete features. Then some scholars implement a series of word embedding methods such as word2Vec, GloVe and FastText to map the words with the shorter word vector space [28], [29]. By learning the distributed representation of words, these methods can reduce the word space and represent the semantic relationship between words with the cosine similarity. Obviously, word embedding can have better expressive ability than one-hot coding method. Therefore, we utilize the idea of word embedding to represent the semantics of words, and then obtain the semantic feature vector of the text through independent training.

After obtaining the vector matrix of each short text, we need to train the word vector matrix to extract text features. The most frequently used method is LSTM, a variant of Recurrent Neural Network (RNN), which replaces hidden nodes in RNN with memory cell units [30]. It can resolve the problems of gradient disappearance and gradient explosion which may appear in long sequence text and longtime training. Based on the self-attention mechanism, the latest Transformer model [31] can capture the attention scores of multiple dimensions between words, represent the features by acquiring text information across distances superiorly. Significantly, Transformer model can perform well only when it is applied to massive data. Considering that the two datasets in our experiments are not adequate to support the training process of the Transformer model, as is evident in section 4, we implement LSTM to obtain short text features.

In summary, based on the idea of word embedding, we train the short queries text from online health Q&A website, obtain the word vector representation of the text, and then adopt LSTM to extract text features.

B. TOPIC FEATURE EXTRACTION

At present, LSTM, Transformer, CNN and their generalizations have been widely adopted to extract text sematic feature. However, their expression abilities are still confined by the external corpus, especially for specialized field, where the correlative contextual semantic clues are not enough to gain text features. Recently, there is a growing trend to deploy probabilistic topic model for the text feature extraction. Consequently, we consider the topic-based feature extraction to enrich and improve the feature expression ability.

Latent Dirichlet Allocation (LDA), a document topic-based generated model, can identify potential topic information from extensive texts collection and is widely applied in text modeling, topic mining and other fields [32]. Based on Bayesian theory, it is generally supposed that the prior distribution of topics, and the prior distribution of words in each topic are all Dirichlet distribution. Then Gibbs sampling method [33] is deployed to obtain the parameters of the LDA model and then to get the latent topic features of each text.

However, in the procedure of establishing LDA model, how to determine the optimal topic numbers of the topic model is a difficulty for scholars. To this end, we conduct a sub-sample experiment, select several topic numbers with best classification effect within a limited range, and obtain the topic feature vectors under these different topic number k. However, due to the unknown characteristics of this potential topic, it will inevitably generate the repeated expression of a certain meaning under different topic k, and then result in information redundancy. In order to avoid this phenomenon, our urgent task is to make a further feature filtering. Since there is no sequence relationship between topic feature vectors, it is not appropriate to use LSTM model for topic feature extraction.

So as to solve the above problems, we use Convolutional Neural Network (CNN) to extract topic features from the concatenated topic feature vectors. As a kind of feedforward neural networks, CNN has been extensively applied in the field of text classification and computer vision due to its powerful feature extraction ability in local space [34]. The following steps illustrate how to handle one-dimensional topic feature vector by CNN.

Input layer: Connect all topic features vectors under different number of k;

Convolutional layer: Extract different features of text by different convolution kernels using the following formula:

$$\alpha_i = f\left(\sum_x c_x \cdot w_{ix} + b_i\right), \tag{1}$$

where w_{ix} denotes the weight of the ith convolution check input in the current window, c_x denotes the window size, b_i represents the bias, f represents the activation function, α_i is the results of convolution and can be used as the input of pooling layer.

Pooling layer: Filter the topic features by the method of maximum pooling.

Full connection layer: Obtain the more representative text vector through integrating all the features of pooling layers.

Briefly, we utilize LDA to extract the text topic feature, then adopt one-dimensional CNN for a further filtering, and thus accomplish the purpose of dimension reduction.

C. MODEL ALGORITHM DESCRIPTION

Based on different feature extraction routes, we obtain the semantic features of the text and the latent topic features of the text. The LSTM&Topic-CNN model can be described as follows:

Preprocess the original texts to obtain the segment texts as the input layer of the LSTM&Topic-CNN model and extract text features and topic features respectively through two paths.

Extract text feature path: Generate the word vector matrix by word embedding method, and then implement LSTM to obtain the text feature; Extract topic feature path: Connect topic features of each text under the different topic number of k, then filter these to extract the topic feature under the idea of one-dimension CNN.

Finally, connect the text features and topic features and then use a classifier to classify the category of each text.

To represent our model more clearly and intuitively, the detailed procedures of our proposed LSTM&Topic-CNN model are given in visualized graph format and algorithm format as shown in Fig.2 and Table 1.

TABLE 1. Description of question text classification algorithm based on LSTM&Topic-CNN model.

```
Algorithm: LSTM&Topic-CNN model. The question text classification based on the feature fusion.
Input:$\{T_1, T_2, T_3, \ldots T_m\}$, a set of $m$ Chinese question texts;
Output:$\{y_1, y_2, y_3, \ldots y_m\}$, the category label of each text.

Methods:(The specific processing steps of the algorithm are introduced as follows)
  Begin
    1. $T_i^* = word\,embedding(T_i)$; /*Learn word embedding from the texts, process each text into $n_i \times d$ dimension vector matrix, where $n_i$ is the number of words and $d$ is the dimension of word vector*/
    2. $T_i{}^L = LSTM(T_i{}^*)$; /*Deploy the three-layers LSTM to output the text vector*/
    3. $k = \{k_1, k_2, \ldots, k_j\}$; /*Based on the LDA model, select the optimal $j$ topic number with the highest classification accuracy from the question text subset*/
    4. $T_i{}^k = concat(T_i^{k1}, T_i{}^{k2}, \ldots, T_i{}^{kj})$; /* Generate the text topic vectors for the given $k$ value, and then connect all vectors as the input layer of CNN */
    5. $T_i{}^C$=$CNN(T_i{}^k)$; /*Adopt the CNN to extract different features and enhance the expression of topic features*/
    6. $T_i{}^{full} = concat(T_i{}^L, T_i{}^C)$; /*Connect $T_i{}^L$ and $T_i{}^C$ as an input of full connection layer*/
    7. $y_i = classifier(T_i{}^{full})$; /*Classify the text and output the category label.*/
  End
```

IV. EXPERIMENTS

In this section, we evaluated the performances of the LSTM&Topic-CNN model based on two real datasets from different online medial Q&A websites. Considering the special problems of Chinese medical text data, data pre-processing technology was employed. After a series of pre-processing operations, all datasets were divided into training texts and testing texts at random. Meanwhile, for the purpose of improving the robustness of experimental results, a sub-sample sensitivity analysis was implemented to obtain the optimal topic numbers. Then the model parameters' setting was considered in the process of building our proposed model. And finally, we compared the performances

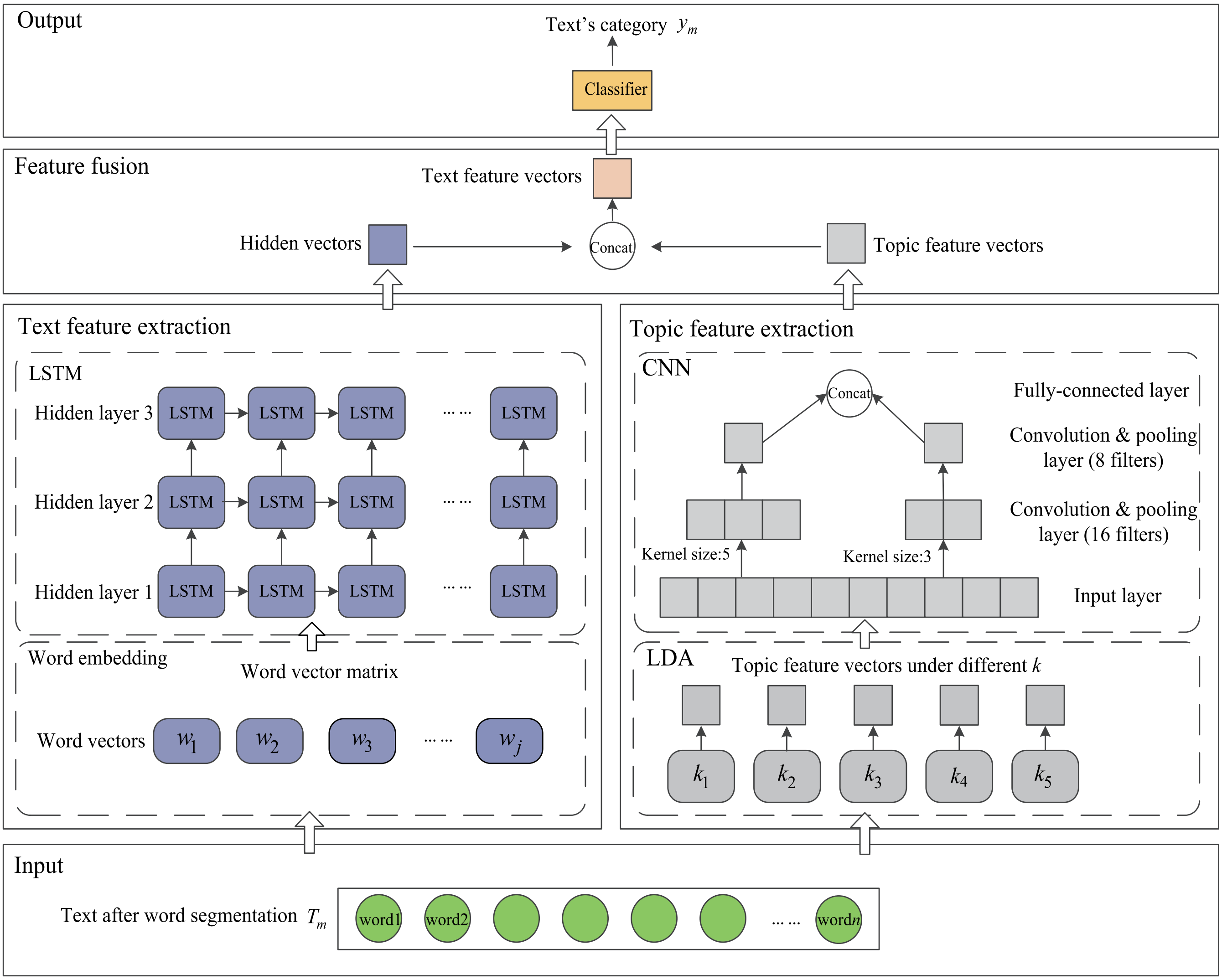

FIGURE 2. Details of the LSTM&Topic-CNN model.

of our model with the baseline model in terms of common evaluation indicators. The flow process was shown in Fig.3.

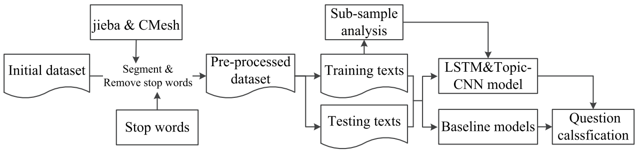

FIGURE 3. Question classification processing for real datasets.

A. DATA SOURCES AND PREPROCESSING

The cardiovascular disease (CVD) risk factors have been grown in popularity with the acceleration of the aging population and urbanization, leading to a continuous increase in the number of cardiovascular diseases. It was reported that about 290 million patients were suffering from CVD in China, and CVD remained the major cause of death in 2016 [35]. Considering continuous attention to this kind of disease, there were plenty of questions and answers about CVD in 39 health Q&A networks (https://ask.39.net). By using the Python software and Web Crawler, we selected the "heart disease" category and "hypertension disease" category in the "cardiovascular disease" module of this website to crawl the data (hereinafter referred to as Data1).

Since Data1 only involved two entity categories, we deployed a multi-classification problem to verify the general applicability of our model. We selected five categories of manually marked Q&A data from 120ask (https://www.120ask.com), namely ophthalmology and otorhinolaryngology (ENT), dermatology, gynecology, orthopedic and gastroenterology data (denoted as Data2). The detailed descriptions of these datasets were listed in Table 2.

It is known that English texts are word-based, with spaces separating words from each other, while Chinese texts are character-based (except for punctuation separating them),

TABLE 2. Detailed information of datasets.

Data sources	Sample size	Categories	Examples
Ask39(Data1)	13,865	Heart disease	"For two days a week, I woke up with severe pain on the right side of my chest."
	13,187	Hypertension	"Is it bad to have a blood pressure reading of 160/100?"
120ask(Data2)	2,675	ENT	"How to stop a nosebleed quickly?"
	2,152	Dermatology	"How to treat hives?"
	3,166	Gynecology	"What is up with the pain in the left abdomen during ovulation?"
	3,152	Orthopedic	"What are the causes of ankle injuries?"
	2,345	Gastroenterology	"Is bacterial enteritis contagious?"

with no spaces between characters. Furthermore, since the understanding of Chinese texts needs context, it is necessary to perform additional word segmentation before conducting classification studies on Chinese texts. Considering that the datasets applied in this paper belong to the medical field and involve medical professional terms, we intergrate the jieba word segmentation tool with CMest (an open-source Chinese medical subject word list) to conduct word segmentation processing for the purpose of increasing the accuracy of word segmentation. Then delete stop words according to the common stop words list, and further complete text segmentation.

B. MODEL EVALUATION INDICATORS

In this paper, we adopt the confusion matrix, precision, recall, F1 value, and receiver operating characteristic (ROC) curve as the evaluation index criteria of model classification effectiveness. Firstly, we introduce the confusion matrix as shown in Table 3, where the row of the matrix represent the actual category of the sample before classification, and the list of the matrix show the prediction of the sample category after classification.

TABLE 3. Confusion matrix.

	Positive	Negative
True	True Positive (TP)	True Negative (TN)
False	Positive (FP)	False Negative (FN)

The specific assessment methods are as follows:

Accuracy indicates the percentage of tuples with correct category identified by the classifier:

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}. \tag{2}$$

Precision represents the percentage of tuples with actual positive category in total tuples with positive category identified by the classifier:

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

Recall indicates the percentage of tuples marked with positive category in total actual positive tuples:

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

F1 value is the weighted harmonic average of precision and recall:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{5}$$

In addition, the ROC curve can be used to evaluate the classification effect of the model text by taking false positive rate (FP rate) and false negative rate (TP rate) as axis. The larger of the area under curve (AUC), the better performance of the classifier.

C. THE SETTING OF MODEL PARAMETERS

This subsection discussed the parameter settings in 4 links (word embedding, LSTM, LDA, and CNN) during the process of establishing the LSTM&Topic-CNN model. Taking Data1 as an example, we explained the detailed parameter configuration in each part below.

Word embedding part: Considering that the original data in the two datasets were almost 20,000, we initialized the word vector of 100 dimensions by the random initialization method, and then transformed segmented texts into word vector text matrixes.

LSTM part: It was known that the number of LSTM layers were generally set to 2, 3, 4. For example, Matthew *et al.* [36] proposed a two-layer and bidirectional LSTM model, achieving a good text feature representative ability in the pre-training of massive data. Since our paper focused on short texts (the average length of the text was 15.2 words in Data1), we utilized unidirectional LSTM instead of bidirectional LSTM to extract text features. Specifically, we chose the common 3-layers LSTM with the dimension sizes 64, 32 and 16 in sequences. The weights of the input and output layer were initialized randomly. To avoid model's over-fitting, neurons were dropped randomly with a probability of 0.2 after finishing the first LSTM layer. Batch Gradient Descent method was used to obtain the text word vector space matrix. After 30 iterations, 16-dimensional text feature vectors were obtained by 3-layers LSTM. The detailed process and related dimension changes were presented intuitively in Fig.4.

LDA part: We set the maximum iteration time in Gibbs sampling method as 10000 times and the prior parameters α as $50/k$ and β as 0.01 in Dirichlet function. In order to obtain optimal plan for k value, we randomly chose

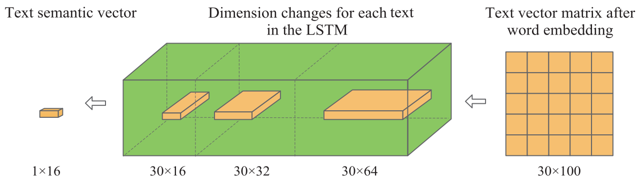
Text semantic vector
Dimension changes for each text in the LSTM
Text vector matrix after word embedding
1×16
30×16
30×32
30×64
30×100

FIGURE 4. Dimension changes in text feature extraction.

1000 sub-samples (500 hypertensive texts, 500 heart disease texts) with different k values from 5 to 30, and then computed the classification accuracy for the sub-samples. The results were presented in Fig.5. Then we selected the top five k values to extract topic features, which were 8, 10, 16, 18, and 22 exactly as evidenced in Fig.5.

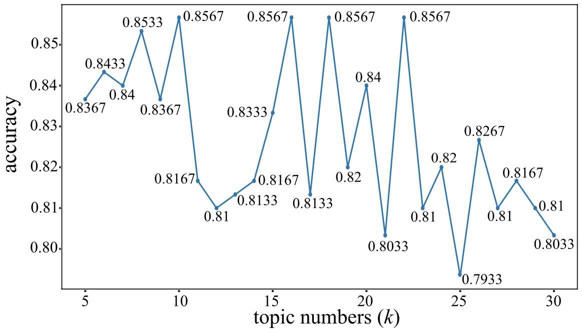

FIGURE 5. Classification accuracy of sub-samples under different *k* values.

CNN part: Concatenate the topic features vector under different k values as the input layer of the CNN model firstly. Then we set two convolution processes in the feature filter process.

Firstly, the 16-dimensional topic feature vectors were obtained by the convolution and pooling procedure, in which the convolution kernel number was 16 and kernel size was 5 respectively. Then further filtration with 8 convolution kernels was considered to achieve 8-dimensinal topic feature vector. Concerning the characteristics of weight sharing, we could only extract a kind of feature by a convolution kernel. In order to improve the feature extraction ability, we needed to deploy convolution kernels with different sizes urgently. Therefore, we also considered the same two convolution operation when the convolution kernel size was set as 3. And finally, the topic feature vectors with different convolution kernel sizes were spliced to get the 16-dimensional text topic feature vectors as the output of the CNN part.

Similarly, we could also set up parameters in each part for Data2. And the detailed change processes of the text vectors dimensions and the number of parameters of two datasets were presented in Table 4. Obviously, Table 4 manifests that we can accomplish the goal of reducing dimension and numbers of parameters with the proposed model.

TABLE 4. The dimensions of text vectors and the number of parameters.

Stage	Data1		Data2	
	Dimension	Parameters	Dimension	Parameters
Word embedding	30 × 100	2,200,000	30×100	2,200,000
LDA part	74		126	
LSTM layer 1	30×64	42,240	30×64	42,240
LSTM layer 2	30×32	12,416	30×32	12,416
LSTM layer 3	30×16	3,136	30×16	3,136
LSTM output	16		16	
CNN part	16	472	16	472
Full-connected layer	32	33	32	165

D. RESULTS & MODEL COMPARISON

To verify the performance of the LSTM&Topic-CNN model on text classification, we built the model environment and learned this model under the training texts. Then we evaluated the classification effects of the model based on the testing texts by the evaluation indicators above. We also conducted a comparative experiment to show the classification results for baseline models such as LSTM, RNN, Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR).

Traditional machine learning methods tend to be high-dimensional, sparse, and discrete, which leads to the weak text feature expressiveness. Among them, RF is an ensemble algorithm combining bagging algorithm and stochastic subspace identification (SSI), applying the Decision Tree as a base classifier [37]. LR is one of the generalized linear models that is classified by the membership degree (posterior probability) of each category [38]. SVM maps low dimensional nonlinear space into high dimensional linear space, which is more suitable for the classification of high dimensional and large sample sets [39]. All the above methods ignore the order and correlation between words. The deep learning methods such as RNN can train the texts into low-dimensional and dense vectors; therefore, they have stronger feature expression ability [40].

For Data1, the overall classification performances for different models were shown in Table 5 and Fig.6. Noticing from Table 5, the neural network models such as RNN and LSTM have moderately improved the classification effect compared with the traditional machine learning models of RF, SVM, and LR. Furthermore, compared with the LSTM model and RNN model, the precision of our LSTM&Topic-CNN model increases by 2.09% and 3.46%, the F1 value promotes about 1.53% and 2.36% respectively. Fig.6 reveals that the LSTM&Topic-CNN model outperforms the other baseline models in accordance with ROC, the area under the ROC of which reaches 0.998.

Since that Data1 contains two categories, we implement the common two-dimensional confusion matrixes to show the performances by category for different models based on testing data, where 0 represents heart disease and 1 represents hypertension in Fig.7. We also note from Fig.7 that

TABLE 5. Classification effects of different models based on testing texts (Data1).

Model	Recall	Precision	F1 value
LSTM&Topic-CNN	98.99%	99.03%	99.01%
LSTM	98.03%	96.94%	97.48%
RNN	97.76%	95.57%	96.65%
RF	95.70%	96.15%	95.92%
SVM	95.60%	95.74%	95.67%
LR	94.58%	94.25%	94.41%

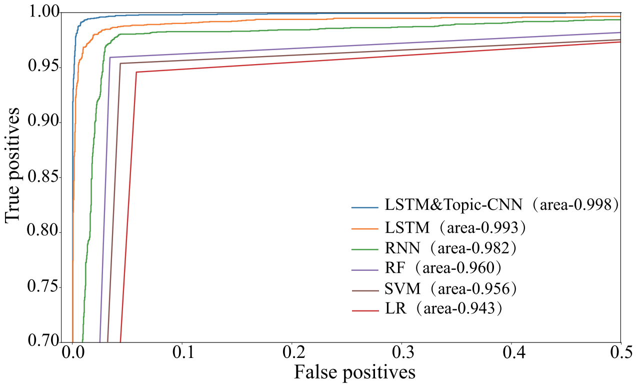

FIGURE 6. ROC curves for different models (Data1).

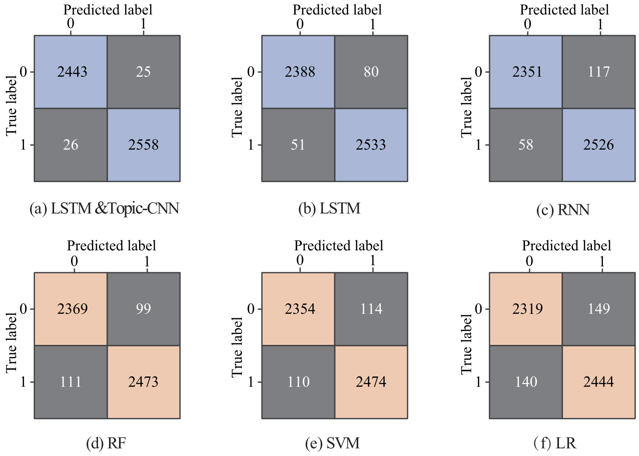

FIGURE 7. Confusion matrixes for different models (Data1).

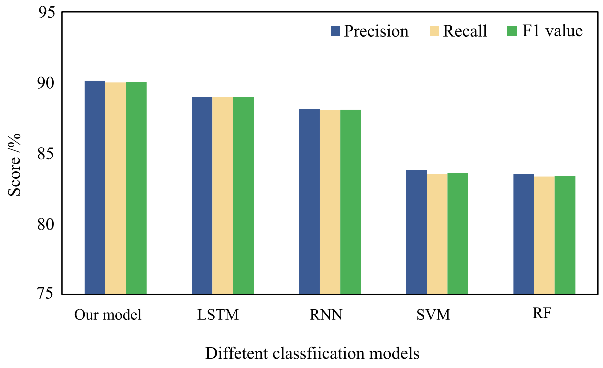

FIGURE 8. Classification effects of different models based on testing texts (Data2).

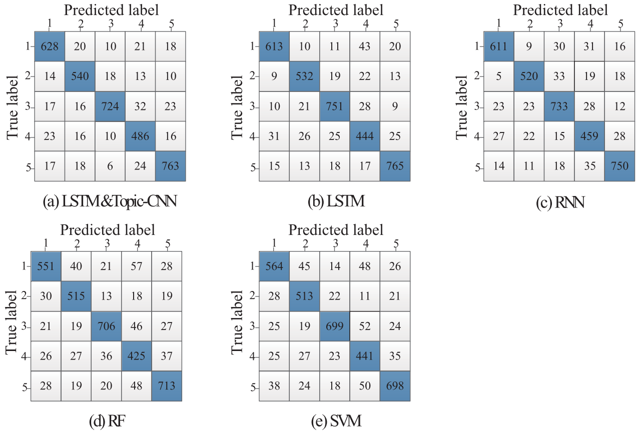

FIGURE 9. Confusion matrixes for different models (Data2).

classification performance of our model tops out, followed by the LSTM, RNN, RF, and SVM, and LR is the worst.

Similar to Data1, we first evaluated the overall performance of different models and then provided their assessments by category for Data2. It was noteworthy that adopting LR to solve the multi-class problem would cause an imbalance in the sample size of the training set or extend the algorithm time. Therefore, we presented the performances of our model, LSTM, RNN, RF, and SVM from the point view of the evaluation indexes and the confusion matrixes, as shown in Fig.8 and Fig.9 respectively. Furthermore, we also showed the classification performance of our model under different categories as shown in Fig.10. The numbers 1 to 5 in Fig.9 and Fig.10 represented the five categories of ENT, gastroenterology, gynecology, dermatology and orthopedics correspondingly.

It is obvious from Fig.8 and Fig.9 that our model achieves the best result, followed by LSTM and RNN, while RF and SVM get the worst result. Fig.10 demonstrates that our model performs well in different categories. The scores are mainly more than 88.2% in whatever evaluation indicators. Specifically, the performances of category 5 (orthopedics) are the best, while those of category 4 (dermatology) are the worst. The possible reason is that diseases in other classes may cause dermatology, resulting in a relatively poor performance in category 4. On the other hand, there exist fewer overlapping questions between category 5 and the others, creating high recognition rate for category 5.

E. DISCUSSION

Based on the experimental results, traditional machine learning methods (RF, LR, and SVM) achieve the worst results in our experiments, which may attribute to their weak ability of text feature expression and ignorance of the order and correlation between words. Moreover, the deep learning methods

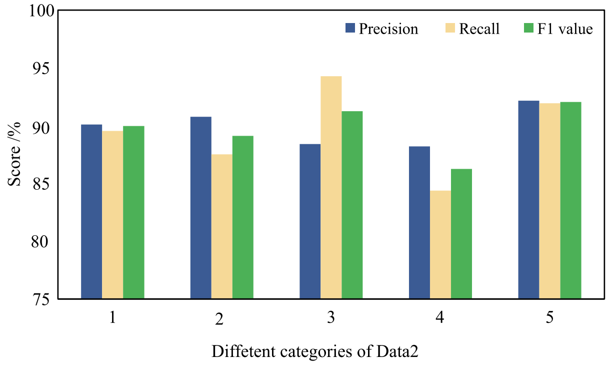

FIGURE 10. Classification effects of different categories under our model (Data2).

(RNN and LSTM) perform better than the above three methods. The possible reason is that they consider the dependence of sequence features and train the texts into low-dimensional and dense vectors, thus have a stronger feature expression ability. Last but not least, LSTM&Topic-CNN model outperform the other models either in binary classification or in multi-class classification. Specifically, based on the feature fusion, our model improves the information quality of feature extraction by advancing deep mining for latent topic information.

We can also note that the performances of Data1 are better than Data2 by whatever methods. It is not surprising because Data2 deals with multi-class problems and has more possibilities of error classification than Data1 with two-class problem. Besides, the sample size for Data1 is greater than Data2, as is evident from Table 2. On the other hand, due to patients' unprofessional background and negligence, it is unavoidable to result in marked errors, and these errors will be enlarged especially when the sample size is relatively small. Another possible reason is that there exist overlapping questions between different categories of Data2, which can increase the probability of classification errors. Notice that we loosen the restrictions of invalid information during the data preprocessing period for Data2 to overcome the problem of inadequate data and improve the algorithm efficiency.

V. CONCLUSION

In the context of poor performances of medical Q&A classification, we proposed an LSTM&Topic-CNN model for the analysis of two datasets from medical Q&A websites. Based on the feature fusion, our paper combined the text features and the topic features to improve the information quality of feature extraction. Through word embedding method and LSTM technology, we managed to quantify the question texts by which solved the problem of semantic absence and curse of dimensionality in traditional processing method. Furthermore, in view of the repeated expression of a certain feature under different topics, we creatively adopted the CNN model for topic feature filtering, by which we could hold topic features to the most extent during the dimension reduction process. We also compared the performances of our model with baseline models including LSTM, RNN, RF, SVM, and LR. The experimental results demonstrated the superiority of our model in terms of classification precision. We hope our study can provide helpful information for researchers in the field of medical Q&A classification, provide patients with more timely feedback on their questions, and increase web traffic for the medical Q&A websites.

ACKNOWLEDGMENT

The authors would like to thank the editor and the referees for their insightful comments that have led to a substantial improvement to an earlier version of the paper.

REFERENCES

[1] P. Jacquemart and P. Zweigenbaum, "Towards a medical question-answering system: A feasibility study," *Stud. Health Technol. Inform.*, vol. 95, pp. 463–468, Jan. 2003.

[2] Z. Hong, Z. Deng, R. Evans, and H. Wu, "Patient questions and physician responses in a Chinese health Q&A website: Content analysis," *J. Med. Internet Res.*, vol. 22, no. 4, Apr. 2020, Art. no. e13071, doi: 10.2196/13071.

[3] M. N. Hajli, "Developing online health communities through digital media," *Int. J. Inf. Manage.*, vol. 34, no. 2, pp. 311–314, Apr. 2014.

[4] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on spark," *Future Gener. Comput. Syst.*, vol. 111, pp. 714–722, Oct. 2020.

[5] C. Wu, G. Luo, C. Guo, Y. Ren, A. Zheng, and C. Yang, "An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions," *J. Biomed. Informat.*, vol. 108, Aug. 2020, Art. no. 103511, doi: 10.1016/j.jbi.2020.103511.

[6] N. Chen, X. Su, T. Liu, Q. Hao, and M. Wei, "A benchmark dataset and case study for Chinese medical question intent classification," *BMC Med. Informat. Decis. Making*, vol. 20, no. S3, pp. 1–7, Jul. 2020, doi: 10.1186/s12911-020-1122-3.

[7] C. González-Caro and R. Baeza-Yates, "A multi-faceted approach to query intent classification," in *Proc. SPIRE*, in Lecture Notes in Computer Science, vol. 7024, Pisa, Italy, 2011, pp. 368–379.

[8] Y. Tian, W. Ma, F. Xia, and Y. Song, "ChiMed: A Chinese medical corpus for question answering," in *Proc. BioNLP*, Florence, Italy, 2019, pp. 250–260.

[9] L. Schmidt, J. Weeds, and J. Higgins, "Data mining in clinical trial text: Transformers for classification and question answering tasks," in *Proc. IFMBE*, Valletta, Malta, Feb. 2020, pp. 24–26.

[10] J. Zhang and Y. Zhao, "A user term visualization analysis based on a social question and answer log," *Inf. Process. Manage.*, vol. 49, no. 5, pp. 1019–1048, Sep. 2013.

[11] C.-Y. Li, S.-S. Zhai, and L. Zheng, "Measurement of information demand characteristics in online health community," *Digit. Library Forum*, vol. 9, no. 148, pp. 34–42, 2016.

[12] M.-H. P. Chiu and C.-C. Wu, "Integrated ACE model for consumer health information needs: A content analysis of questions in Yahoo! Answers," in *Proc. ASIST*, Baltimore, MD, USA, 2012, pp. 28–31.

[13] Y. Lu, Y. Wu, J. Liu, J. Li, and P. Zhang, "Understanding health care social media use from different stakeholder perspectives: A content analysis of an online health community," *J. Med. Internet Res.*, vol. 19, no. 4, p. e109, Apr. 2017, doi: 10.2196/jmir.7087.

[14] J. He, M. Fu, and M. Tu, "Applying deep matching networks to Chinese medical question answering: A study and a dataset," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, pp. 91–100, Apr. 2019.

[15] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: A survey," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 1–66, Jan. 2017.

[16] L. Ge and T.-S. Moh, "Improving text classification with word embedding," in *Proc. IEEE Big Data*, Boston, MA, USA, Dec. 2017, pp. 1796–1805.

[17] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, Mar. 2015.

[18] D. C. Edara *et al.*, ‘‘Sentiment analysis and text categorization of cancer medical records with LSTM,’’ *J. Ambient Intell. Hum. Comput.*, 2019, doi: 10.1007/s12652-019-01399-8.

[19] Z. Wang, L. Ma, and Y. Zhang, ‘‘A hybrid document feature extraction method using latent Dirichlet allocation and word2vec,’’ in *Proc. IEEE DSC*, Changsha, China, Jun. 2016, pp. 98–103.

[20] D.-R. Liu, Y.-K. Hsu, H.-Y. Chen, and H.-J. Jau, ‘‘Air pollution prediction based on factory-aware attentional LSTM neural network,’’ *Computing*, vol. 103, no. 1, pp. 75–98, Jan. 2021, doi: 10.1007/s00607-020-00849-y.

[21] Z. Jiang, S. Gao, and L. Chen, ‘‘Study on text representation method based on deep learning and topic information,’’ *Computing*, vol. 102, no. 3, pp. 623–642, Sep. 2019, doi: 10.1007/s00607-019-00755-y.

[22] B. Liu, Y. Zhou, and W. Sun, ‘‘Character-level text classification via convolutional neural network and gated recurrent unit,’’ *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 8, pp. 1939–1949, Mar. 2020.

[23] L.-X. Luo, ‘‘Network text sentiment analysis method combining LDA text representation and GRU-CNN,’’ *Pers. Ubiquitous Comput.*, vol. 23, nos. 3–4, pp. 405–412, Jul. 2019.

[24] J. Xie, Y. Hou, Y. Wang, Q. Wang, B. Li, S. Lv, and Y. I. Vorotnitsky, ‘‘Chinese text classification based on attention mechanism and feature-enhanced fusion neural network,’’ *Computing*, vol. 102, no. 3, pp. 683–700, Mar. 2020, doi: 10.1007/s00607-019-00766-9.

[25] B. Tang, X. Wang, J. Yan, and Q. Chen, ‘‘Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF,’’ *BMC Med. Inform. Decis. Making*, vol. 19, no. S3, pp. 97–114, Apr. 2019.

[26] G. Liu, Q. Yuan, J. Duan, J. Kou, and H. Wang, ‘‘Chinese question classification based on ERNIE and feature fusion,’’ in *Proc. NLPCC*, in Lecture Notes in Computer Science, vol. 12431, Oct. 2020, pp. 343–354.

[27] L. Yuanqing, Y. Suying, X. Jiangtao, and G. Jing, ‘‘A self-checking approach for SEU/MBUs-hardened FSMs design based on the replication of one-hot code,’’ *IEEE Trans. Nucl. Sci.*, vol. 59, no. 5, pp. 2572–2579, Oct. 2012.

[28] A. Nikfarjam, A. Sarker, K. O’Connor, R. Ginn, and G. Gonzalez, ‘‘Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,’’ *J. Amer. Med. Inform. Assoc.*, vol. 22, no. 3, pp. 671–681, Mar. 2015.

[29] Y. Wang, S Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, ‘‘A comparison of word embeddings for the biomedical natural language processing,’’ *J. Biomed. Inf.*, vol. 87, pp. 12–20, Nov. 2018.

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

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 1735–1780.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, ‘‘Latent Dirichlet allocation,’’ *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jul. 2003.

[33] A. E. Gelfand, ‘‘Gibbs sampling,’’ *J. Amer. Stat. Assoc.*, vol. 95, no. 452, pp. 1300–1304, Dec. 2000.

[34] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1746–1751.

[35] *Report on Cardiovascular Diseases in China 2018*. Accessed: Mar. 2, 2021. [Online]. Available: https://www.nccd.org.cn/News/Information/Index/1089

[36] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ‘‘Deep contextualized word representations,’’ in *Proc. NAACL*, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[37] L. Breiman, ‘‘Random forests,’’ *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[38] S. Menard, ‘‘Logistic regression,’’ *Amer. Stat.*, vol. 58, no. 4, p. 364, 2004.

[39] T. Joachims, ‘‘Text categorization with support vector machines: Learning with many relevant features,’’ in *Proc. ECML*, Berlin, Germany, 1998, pp. 137–142.

[40] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanour, ‘‘Recurrent neural network based language model,’’ in *Proc. INTERSPEECH*, Chiba, Japan, 2010, pp. 1045–1048.

SONG MAO received the M.S. and Ph.D. degrees in applied mathematics from Northwestern Polytechnical University, in 2012 and 2014, respectively. She is currently an Associate Professor with the School of Economics and Management, Shanxi University. Her research interests include data mining, recommendation systems, and applied probability and statistics.

LU-LU ZHANG received the B.S. degree in industrial engineering from the Henan University of Science and Technology, in 2018. She is currently pursuing the M.S. degree with the School of Economics and Management, Shanxi University. Her research interests include natural language processing and data analysis.

ZHEN-GUO GUAN received the B.S. degree in industrial engineering from the Jiangxi University of Science and Technology, in 2017. He is currently pursuing the M.S. degree with the School of Economics and Management, Shanxi University. His research interest includes natural language processing.