Journal of Computers Vol. 33 No. 4, August 2022, pp. 1-14
doi: 10.53106/199115992022083304001

Chinese News Text Classification and Its Application Based on Combined-Convolutional Neural Network

Kai-Feng Liu[1], Yu Zhang[2,3*], Quan-Xin Zhang[4], Yan-Ge Wang[2], Kai-Long Gao[2]

[1] Jiangsu Vocational College of Finance and Economics, Huaian 223003, China
liu923683429@163.com
[2] School of Electrical and Information Engineering & Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[3] State Key Laboratory in China for GeoMechanics and Deep Underground Engineering, China University of Mining & Technology, Beijing 100083, China
[4] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Received 12 July 2021; Revised 11 November 2021; Accepted 2 March 2022

Abstract. A method based on combined-convolutional neural network (Combined-CNN) for Chinese news text classification is proposed. First of all, in order to solve the problem of a lack of special term set for Chinese news classification, a vocabulary suitable for Chinese long text classification is made by constructing a data index method. The Word2Vec pre-trained model was used to embed the text features word vectors. Second, by optimizing the structure of the classical convolutional neural network (CNN) model, a new idea of Combined-CNN model is proposed, which solves the problem of incomplete feature extraction of local text blocks and improves the accuracy rate of Chinese news text classification. Effective model regularization and RAdam optimization algorithm are designed in the model to enhance the model training effect. The experimental results show that the precision of the Combined-CNN model for Chinese news text classification reaches 93.69%. Compared with traditional machine learning methods and deep learning algorithms, the accuracy rate is improved by a maximum of 11.82% and 1.9%, respectively, and it is better than the comparison model in Recall and F-Measure. Finally, the Chinese news classification algorithm of the Combined-CNN is applied to realize a personalized recommendation system.

Keywords: Combined-CNN, Chinese news, text classification, recommendation system

1 Introduction

The Internet and big data industries are booming, accompanied by endless text information. Since the end of the 1990s, more than 200 news websites have been officially approved by the state. There are many kinds of mobile news APPs, resulting in massive news text data. News text data is an important form of social information and has become an important part of people's access to social information resources. In order to efficiently obtain and manage valuable news information, news text classification has become a hot research field in the world [1]. News classification is helpful to the management of text information, the realization of news order and the mining of news data [2].

Due to the influence of global economic integration and the strategy of "The Belt and Road", Chinese has already occupied an important position in the world language system. However, the classification of Chinese news text is very few, especially the classification of Chinese long text [3]. Chinese is much more complex than western languages, and it is difficult to extract features by traditional methods. At the same time, there are fewer related corpora for studying Chinese text classification. This is also the reason for the slow development of Chinese news text classification [4-5].

Firstly, to address the problems encountered in Chinese news text classification, this paper uses the method of constructing data index to produce a term set suitable for Chinese text classification. Secondly, by improving the classical CNN model [6-8], this paper proposes a new idea of Combined-CNN model, which makes the local feature extraction of text blocks better and improves the classification effect of Chinese news text. Finally, the Chinese news classification algorithm of the Combined-CNN is applied to realize a personalized recommendation system [9-10].

The main contributions of this paper are summarized as follows:

* Corresponding Author

First, we propose a method of constructing data index, which can make a vocabulary suitable for Chinese long text classification and provide convenient word vector mapping for feature input.

Second, we propose a Combined-CNN model, which makes the local feature extraction of text better and greatly improves the accuracy of Chinese news text classification.

Third, we use model regularization and Rectified Adam (RAdam) optimization algorithm, which can effectively prevent model training overfitting and optimize the results of model training.

Fourth, we have implemented a personalized recommendation system for news classification, and fully applied the Chinese news classification algorithm of the Combined-CNN.

This paper is organized as follows. Section II introduces the related work of text classification algorithms. Section III explains the details of the Combined-CNN model algorithm. Experimental results and discussion are presented in Section IV. In Section V, the specific application of Chinese news classification algorithm is described. Finally, Section VI concludes with future work.

2 Related Work

Text classification is one of the basic problems of natural language processing. Solving this problem has opened many doors for natural language processing, such as information retrieval, machine translation and automatic summarization. The common machine learning algorithms for news text classification are: Naive Bayes (NB) [11], k-Nearest Neighbor (KNN) [12], Decision Tree (DT), Neural Networks (NNs), Maximum Entropy Model (ME) and Support Vector Machine (SVM) [13].

The distributed representation of words was first applied to statistical language models by Bengio et al. [14], and neural language models began to gain widespread attention. In 2008, Collobert et al. [15] proposed and used a neural network approach to represent text words as tensor data. Similar words are mapped to similar positions in the vector space, and the meaning of a word is determined by its contextual vocabulary, but its shared word embedding can only collaborate low-level information through matrices. In 2013, Mikolov et al. [16] proposed two models, the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. CBOW model input the context related word vector of a feature word and output the word vector of a specific word in the way of a priori probability. The prediction method of the Skip-gram model is opposite to CBOW model, and the word vector of the context is predicted by inputting the vector of the middle word. Skip-gram model can better handle rare words, but when the amount of data is large, the training time is too long [17]. To solve the problem of efficiently training on millions of dictionaries and hundreds of millions of data sets, Google has open sourced a tool for word vector computing — Word2Vec [18]. This tool mainly maps words to low-dimensional space, and uses these lower dimensional word embedding vectors into the classifier. Moreover, the word embedding obtained by the Word2Vec can well measure the similarity between words and words. Barakat et al. [19] proposed that multi-layer neural networks have strong feature learning ability, which can map the real meaning of original data more accurately after training.

In 2014, Deep Recurrent Neural Network (RNN) was used for the evaluation of sentiment classification tasks [20-21]. Since then, because RNN has the problem of gradient explosion and gradient disappearance in processing long text, Tai et al. [22] proposed improved semantic representations from long short-term memory neural network (LSTM). The convolutional neural network model was originally invented for computer vision and was later proven to be effective for NLP by Meek et al. [23], which has achieved good results in semantic analysis. In 2015, Some researchers proposed a character-level convolutional network model, using different classification data sets for semantic analysis and topic classification tasks [24]. However, the effect of this method in Chinese text classification is general, and the training and work of the classifier are slow, because the term set and N-gram of Chinese text classification are much larger than that of English text classification [25]. In 2017, Google proposed the Transformer framework to capture the global relationship between input and output based on the attention mechanism, which was widely applied to various pre-training models [26]. In 2018, Devlin et al. [27] proposed BERT, a deep bidirectional Transformer-based pre-training model. The feature extractor used by BERT is the Transformer encoder part, and two strategies, Masked Language Model (MLM) and Next Sentence Prediction (NSP), were used for model pre-training.

Based on the previous research results, this paper proposes a supervised learning model of the Combined-CNN, which improves the structure of the classical CNN model by the way of separate convolution and recombination. The Combined-CNN enhanced the extraction of local features of the text, and finally achieved good results of Chinese news text classification.

3 Combined-CNN for Chinese News Text Classification

According to the processing steps of English text classification, the process of Chinese News Classification in this paper includes: data set preprocessing, text feature representation, feature extraction and classifier training.

3.1 Constructing Data Index and Integration

The data set used in this paper is THUCnews, which is derived from the historical data filtered and generated by Tsinghua University according to Sina News RSS subscription channel from 2005 to 2011. It contains 836075 news documents (2.04GB), all in UTF-8 plain text format. On the basis of the original Sina News classification system, it is divided into 14 categories, such as technology, stocks, sports, entertainment, etc.

In order to construct the index of the whole data set more conveniently, this paper performs a big data visualization analysis of THUCnews. Thus, the optimal text sequence length is determined and set, and its also used as a standard for sentence padding length in the later model. According to statistics, the average number of words per news is 941. It can be seen from Fig. 1 that most of the news are within 2000 words, and the cumulative distribution of the text appearance frequency is shown in Fig. 2, 90% of the quantile points correspond to a text length of 1857. Hence, according to the results of the visual analysis, this article sets the read text length as 2000.

Because the computer processes more than 800000 text files and takes a long time to read them, this paper exploits the pickle standard module in Python to store complex data types and transform text information into binary data streams. The loading speed of binary files is more than 50 times that of text files. This kind of information is stored in the hard disk, which is very convenient when the experiment reads the file data. The original data can be obtained by deserializing it. A certain number of files are saved for each integration to avoid memory overflow.

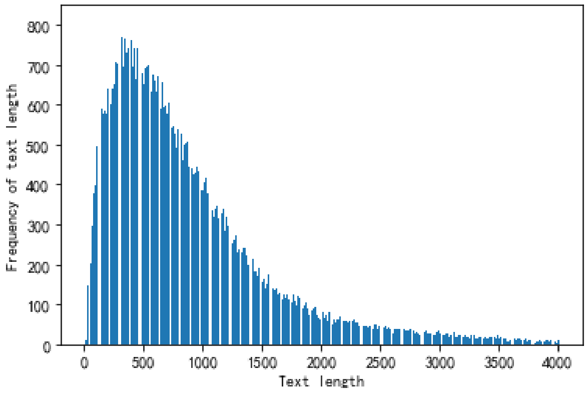

Fig. 1. Frequency statistics of text length

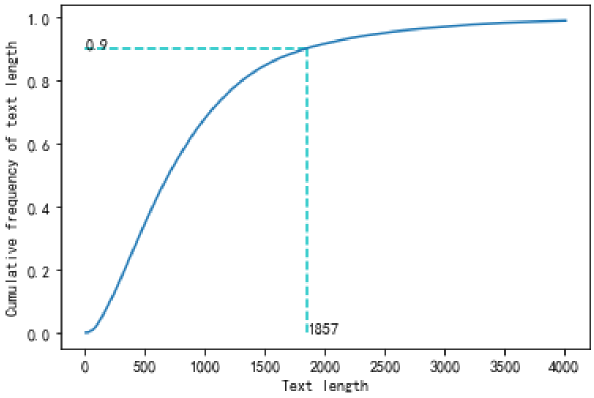

Fig. 2. Cumulative distribution of text length

3.2 The Representation of Text Features

A vocabulary list is made to prepare for the standardization of text data. The reason for removing the stop words in the vocabulary list is that they are used too frequently and have little semantic impact, such as these words “的”, “了”, “在”, “是”, “也”, etc. If there are a large number of such words in the vocabulary, it is equivalent to wasting a lot of resources. So, the stop words in the news text are removed from the vocabulary to give more space for keywords. According to statistics from Beijing Guoan Information Equipment Co., Ltd., there are 91,251 Chinese characters in the Chinese character database. Commonly used Chinese characters are only a few thousand words, about 2500 to 7000. Therefore, the words of all Chinese news texts are counted statistically, and the words whose occurrence frequency ranks in the top 7000 are made as vocabulary corpus.

Using the prepared vocabulary to standardize the data:

(1) Data standardization of text content. First, traverse the index sequence of the vocabulary, list the data and data subscripts, and use the dictionary method to convert the list. Second, the mapping of words and word ids is implemented using list comprehensions and lambda anonymous functions. Finally, the word mapping is used for each sample content to obtain standardized data. The word is converted into vector data of the word ids.

(2) Data standardization of text label. The One-Hot encoding, which is widely used for categorical data, is used

to vectorize the text labels. Each label is represented as an all-zero vector, and only the element corresponding to the label index is one.

3.3 An Improved Combined-CNN Model

3.3.1 Combined-CNN

In order to classify news texts, a six-layer Combined-CNN model is designed and implemented in this paper on the basis of the classical CNN model, as shown in Fig. 3.

Layer 1: embedding layer for receiving input. Because the input data of news classification is text data, the text data needs to be converted into real number vector data to be input. Therefore, the Word2Vec is used in the input layer to map the vocabulary semantics into a real number vector. Then the word embedding is performed on the sample content of the data standardization, and the word vector representation of the sentence is obtained as the input of the next layer.

Layer 2 and 3: convolutional layer and pooling layer. Compared with the classical CNN model, the Combined-CNN model mainly improves the convolution and pooling operations. The classical CNN model has different situations of single-layer convolution and multi-layer convolution. In terms of single-layer convolution, the local text feature information extracted by a convolution kernel is limited and not complete. In terms of multi-layer convolution, the text features extracted by multi-layer convolution operation in a superposition way are often too abstract. It is not conducive to expressing the true meaning of the text. Therefore, the Combined-CNN model uses three different sizes of convolution kernels to extract more complete local text block features in layer 2. At the same time, to extract the main features and reduce the number of feature parameters, the output of convolution is maximally pooled by using the feature of down sampling of maximum pooling layer. Thus, more and more important text features are extracted without deepening the depth of neural network.

Layer 4 and 5: middle hidden layer, there are no two hidden layers in the classical CNN. Because the output of the third layer is the result of three pooling operations, the hidden layer is used to combine the feature vectors extracted by different convolution kernels. In this model, the number of each convolutional kernel is set to be large, and the dimension of the combined feature vector through the fourth layer is too large, so a hidden layer is added to reduce the dimension.

Layer 6: fully connected layer. First, Dropout layer is added to the fully connection layer to prevent the model from over fitting and improve the generalization ability of the model. Secondly, the model uses ReLU as the activation function to increase the nonlinearity of the neural network model and avoid the problem of the disappearance of the neural network gradient. Finally, Softmax function is used to classify and predict news text.

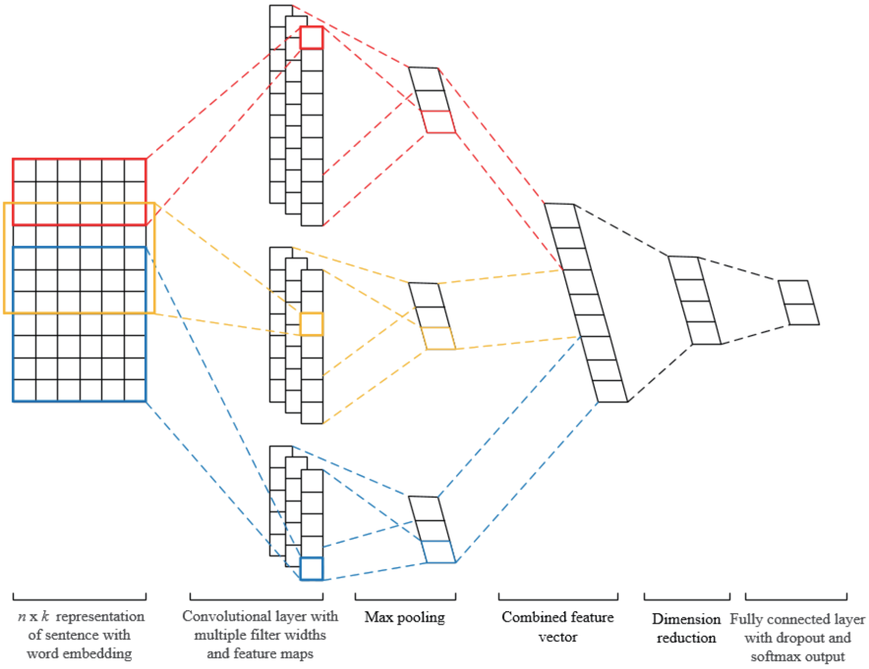
n x k representation of sentence with word embedding
Convolutional layer with multiple filter widths and feature maps
Max pooling
Combined feature vector
Dimension reduction
Fully connected layer with dropout and softmax output

Fig. 3. Combined-CNN model architecture

The following is a detailed description of the Combined-CNN principle:

The embedding layer is a dictionary lookup that maps integer indexes to dense vectors, as shown in Fig. 4. It takes an integer as input, looks up the integers in the internal dictionary, and returns the associated vector. The Word2vec was used for word vector mapping in this layer. Word embedding was performed on the input data to obtain the word vector input into the convolution layer.

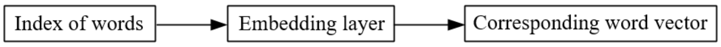

Fig. 4. Embedding layer

The vectorized Chinese text after mapping is a k -dimensional word vector R^k , Assuming that $x_i \in R^k$ is the vector representation of the i -th word, a sentence of length n can be represented as

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n , \quad \textbf{(1)}$$

Where $\oplus$ represents the connection operation, and $\boldsymbol{x}_{i:i+h-1}$ is the word vector matrix in the i -th to $(i+h-1)$ -th windows of the input. The convolutional layer uses a convolution kernel of different sizes to convolute a continuous window of width k , and the convolution kernel is a matrix of $h*k$. In this paper, the height h values of the three convolution kernels are set to 3, 5, and 7 respectively. There are r convolution kernels of each size and the value is set to 256. The weight matrix $\boldsymbol{W}_1 \in R^{h*k}$ performs feature extraction on the text blocks of h words, and a feature o_i extracted by $\boldsymbol{x}_{i:i+h-1}$ is defined as (2).

$$o_i = f\left(\boldsymbol{W}_1 \cdot \boldsymbol{x}_{i:i+h-1} + b_1\right), \quad \textbf{(2)}$$

Here $f(\cdot)$ is a non-linear function of ReLU, and $b_1 \in R$ is a bias term. The convolution operation applied to the word vector $\{\boldsymbol{x}_{1:h}, \boldsymbol{x}_{2:h+1}, \cdots, \boldsymbol{x}_{n-h+1:n}\}$ of a complete news text produce a feature map (3).

$$\boldsymbol{o} = \left[o_1, o_2, \ldots, o_{n-h+1}\right], \quad \textbf{(3)}$$

Where $\boldsymbol{o} \in R^{n-h+1}$. To simplify the computational complexity of the network, the maximum pooling operation is used to compress each feature map and extract the main features. The maximum value in each feature map is taken as the most important feature extracted from the text vector, and a feature vector with the dimension of $1*r$ is obtained. The result of the max-pooling operation $\hat{\boldsymbol{o}}$ is defined as (4).

$$\hat{\boldsymbol{o}} = max\{\boldsymbol{o}\}, \quad \textbf{(4)}$$

The above content describes a convolution kernel to extract features. In this model, multiple convolution kernels with different window sizes are used to obtain multiple features. The maximum value obtained from max-pooling operation of different convolution kernel is combined to produce a new feature vector $\boldsymbol{a} \in R^{1*3r}$, which is defined as (5).

$$\boldsymbol{a} = \hat{\boldsymbol{o}}^3 \oplus \boldsymbol{o}^5 \oplus \boldsymbol{o}^7 , \quad \textbf{(5)}$$

Where $\hat{\boldsymbol{o}}^h$ represents the feature vector after max-pooling using the convolution kernel of height h . Then, a hidden layer is added for nonlinear dimensionality reduction to become a feature vector $\boldsymbol{z} \in R^{1*d}$ (d is the number of neural units in the hidden layer, which is set as 128).

The feature vector z is transferred to the full connection layer, and the regularization method Dropout layer is added to reduce the interaction between hidden layer nodes, so as to reduce the over fitting phenomenon. In each training batch, the output value of hidden layer node is cleared to 0 with a certain probability of p. Finally, the probability distribution of 14 category labels is output through the Softmax layer. Taking the category corresponding to the maximum probability, and the predicted category label value y_i is defined as (6).

$$y_i = max\left[softmax(\boldsymbol{W}_2 \cdot z + b_2)\right], \tag{6}$$

Where $\boldsymbol{W}_2 \in R^{m*d}$, m is the number of classes, $b_2 \in R$ is a bias term.

3.3.2 Regularization and RAdam Optimization Algorithms

(1) The amount of data in this article is close to one million levels. It is very important to reasonably divide the training set, validation set and test set. In the case of a large amount of data, more sample data should be given to the training set to reduce the proportion of the validation set and the training set. Therefore, this paper randomly divides 686075 Chinese news samples for training, 50,000 verification sets for model verification and optimization, and uses 100,000 test sets to evaluate the classification effect of the model. As shown in Fig. 5.

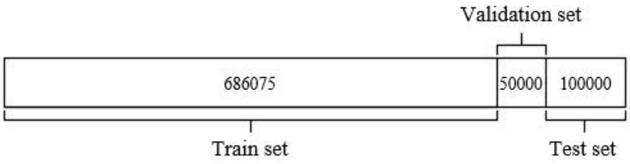

Fig. 5. The division of data set

(2) A suitable optimizer must be selected to obtain the best model training results. In this paper, Rectified Adam (RAdam), Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) optimization algorithms are used for experimental comparison on the Combined-CNN model. In order to set an appropriate initial learning rate, multi-group learning rate of different optimization algorithms are used to compare the effects of experiments on the same data set.

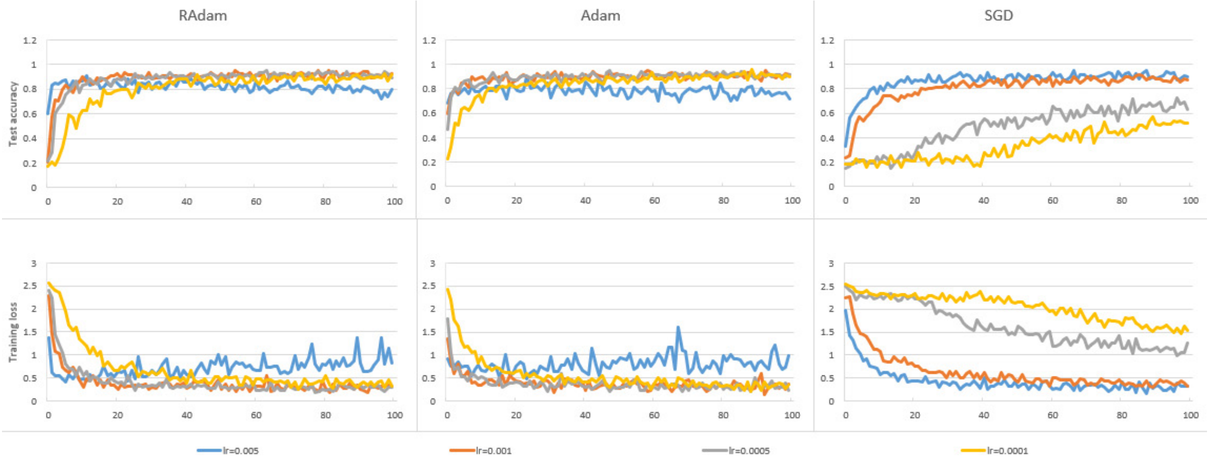

Fig. 6. Multi-group learning rate of different optimization algorithms

It can be found from Fig. 6 that the multiple sets of learning rates of different optimization algorithms have varying degrees of impact on the classification results. First: Compared with the SGD optimization algorithm, the RAdam and Adam optimization algorithms have good robustness to the initial learning rate. It can adapt to a wider range of changes. In a wide range from 0.0001 to 0.005, the RAdam and Adam optimization algorithms show consistent performance, and the end of the training curve is highly coincident, which can bring better training performance. Second, the convergence speed of the RAdam and Adam optimization algorithms are obviously better than that of SGD optimization algorithm. The initial learning rate has a great influence on the SGD algorithm, and it is easy to converge to a local optimal solution. The SGD optimization algorithm is suitable for large initial learning rate, but its convergence speed is slow. Third, the overall average accuracy of the RAdam algorithm classification results is 0.28% higher than that of Adam algorithm, and the overall average loss value is 0.3% lower than that of the Adam algorithm. The overall optimization effect of the RAdam algorithm is better than the Adam algorithm. Finally, the model training optimizer uses the RAdam optimization algorithm.

3.3.3 Model Training

In this paper, the Combined-CNN model is trained by minimizing the loss function on the training set. The loss function uses multiple classification cross entropy. The logarithmic loss function is defined as (7).

$$L(Y,P(Y\mid X))=-\frac{1}{N}\sum_{i=0}^{N-1}\sum_{m=0}^{M-1}y_{i,m}\log p_{i,m}\ , \tag{7}$$

Where $L(\cdot)$ is the loss function, X is the input variable, and Y is the output variable. $y_{i,m}$ is a binary indicator that indicates whether category m is the true category of input X_i . $p_{i,m}$ represents the probability that the i -th sample of N samples is predicted as the m -th tag value. The loss value is used to measure the distance between the probability distribution of the network output and the true probability distribution of the label. Training network can make the output as close to the real label as possible. The RAdam optimization algorithm automatically and dynamically adjusts the adaptive learning rate. Model training adopts the method of model saving and loading in Tensorflow, and it takes about 30 minutes to complete 10,000 iterations of training.

4 Experiments

The experiment is carried out under Windows 10 system. The CPU is Inter(R) Core(TM) i7-8750H 2.20GHz, and the memory size is 16GB. The programming language is Python-3.7.2, the development tool is Jupyter_notebook-6.0.1, and the deep learning framework used is Tensorflow_gpu-1.13.1.

4.1 Hyperparameters

During the experiment, the setting of adjustable parameters in the Combined-CNN model is consistent, as shown in Table 1. To speed up the convergence speed, a small batch of sample gradient descent is used, each batch is 64. In addition, the number of hidden neurons in the fully connected layer is 128.

Table 1. The setting of adjustable parameters

Parameter	Value
Text sequence length	600
Word vector dimension	128
Filter sliding window size h	3,5,7
Number of filters	256
Pooling strategy	Max pooling
Learning rate	5e-4
Dropout rate	0.5
Activation function	ReLU

4.2 Experiment Methods

This paper uses different models to classify Chinese news texts. In order to evaluate the effect of the classification model, the classification results are measured by the overall average of Precision, Recall and F-Measure. To verify the classification performance of the Combined-CNN model, we select multiple baselines for comparison.

(1) Combined-CNN is compared with traditional machine learning algorithms including NB, KNN and SVM.

To prevent the experimental results from being incomparable due to different feature construction methods, the feature construction of traditional machine learning methods is also based on word vectors.

(2) Combined-CNN is compared with the classical CNN and other deep learning algorithms. The classical CNN includes single-layer convolutional neural network (CNN-1) and superimposed multi-layer convolutional neural network (CNN-3); other deep learning algorithms include Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU).

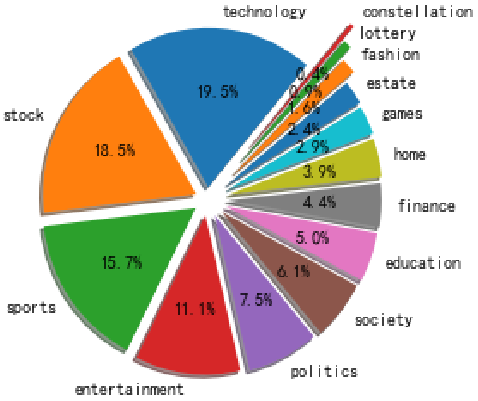

Fig. 7. The proportion of various samples

(3) To further improve the precision of the model and reduce the impact of unbalanced sample data on the classification results, the data set is balanced. For example, as shown in Fig. 7, there are too few samples of “constellation”, “lottery”, “fashion” categories, less than 3% of the total, while there are too many samples of “technology”, “stock”, “sports” categories, only three categories exceed 50% of the total. Therefore, it will lead to the poor classification effect of the former. It can be seen from the confusion matrix that some samples of the former will be classified into the latter. The data set that has been randomly divided again has a total of 65000 sample data, which is divided into 10 categories, including 5000×10 training sets, 500×10 development sets and 1000×10 test sets. Based on different data sets, the classification results of the Combined-CNN model are compared.

4.3 Results and Discussion

(1) The overall average of Precision, Recall and F-Measure of the Combined-CNN and machine learning algorithms are shown in Table 2.

Table 2. Classification results of Combined-CNN and machine learning algorithms

Model	Precision	Recall	F-Measure
NB	0.8187	0.8162	0.8140
KNN	0.8548	0.8526	0.8494
SVM	0.8735	0.8763	0.8739
Combined-CNN	0.9369	0.9368	0.9368

According to the comparison in Table 2, it can be found that first, the Word2Vec word bag model pre-trained word vector is used for feature construction. On the same data set, each classification model has achieved a precision of more than 80%, indicating that the word vector can describe text features well. Second, the classification results obtained by the Combined-CNN are better than the three traditional machine learning algorithms, indicating that the Combined-CNN model can learn more classification features than the traditional machine learning models. Third, the Precision of the Combined-CNN model for Chinese news text classification reached 93.69%. Compared with the classification results of NB, KNN, and SVM, the Precision were increased by 11.82%, 8.21%, and 6.34%, respectively. At the same time, the Recall and F-Measure value of the Combined-CNN model are also better than comparison models. It shows that the Combined-CNN model can improve the text classification effect very well.

For further analysis, we compared the Combined-CNN model and the SVM model with the best classification results among the three machine learning algorithms. The test model outputs a set of training accuracy and loss values every 100 iterations. With the change of iteration times, the training accuracy and loss of different models are shown in Fig. 8.

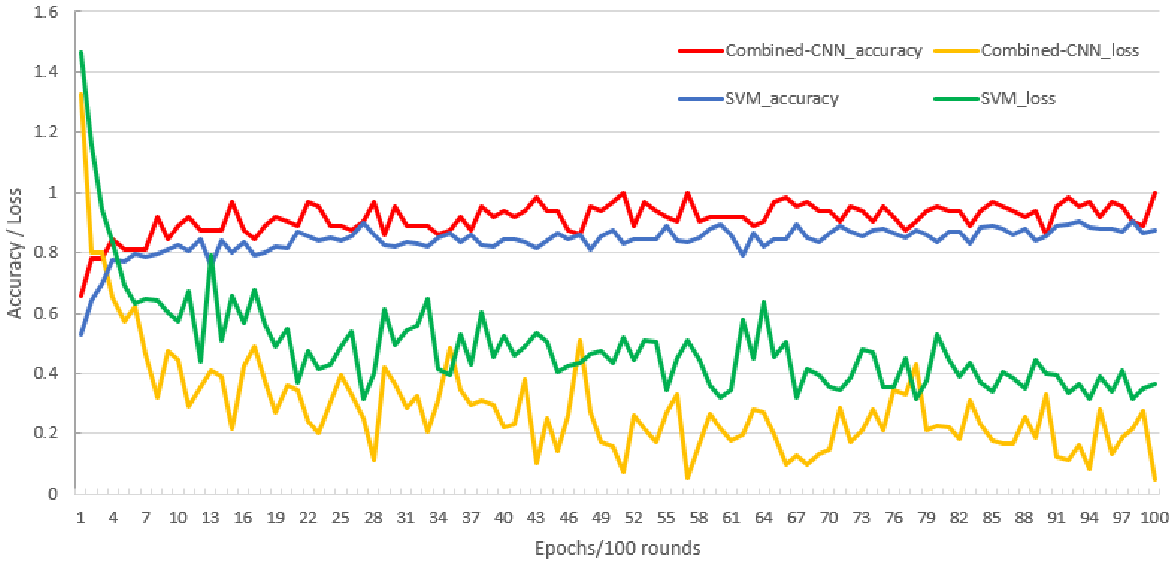

Fig. 8. Training accuracy and loss of Combined-CNN and SVM

It can be seen from the Fig. 8 that the model accuracy value rises quickly with the increase of the number of iterations and tends to stabilize. Because of the role of the RAdam optimization algorithm, the loss value gradually decreases, and finally stabilizes in a small interval fluctuation. The accuracy value of the Combined-CNN model is higher than that of the SVM model, the convergence speed of the loss value is obviously faster, and the loss value is lower than the SVM model. It can be seen that the Combined-CNN model algorithm has huge advantages over traditional machine learning algorithms.

(2) The overall average of Precision, Recall and F-Measure of the Combined-CNN and deep learning algorithms are shown in Table 3.

Table 3. Classification results of Combined-CNN and deep learning algorithms

Model	Precision	Recall	F-Measure
CNN-1	0.9250	0.9240	0.9243
CNN-3	0.9192	0.9178	0.9167
RNN	0.9179	0.9179	0.9175
LSTM	0.9229	0.9226	0.9226
GRU	0.9266	0.9265	0.9264
Combined-CNN	0.9369	0.9368	0.9368

According to the comparison in Table 3, it can be found that first, each deep learning algorithm has achieved a precision of more than 90% on the same data set, indicating that these deep learning algorithms are more effective than machine learning methods. Second, the CNN-3 model has a worse classification effect than the CNN-1 model, which shows that blindly deepening the number of layers of the neural network on the basis of the classical CNN model cannot achieve better results. Compared with the classification results of the CNN-1 model, the Precision of the Combined-CNN model is improved by 1.19%. It shows that the method of convolution and recombination of word vectors can extract more comprehensive local text block feature information. Third, compared with the classification results of the RNN, LSTM, and GRU models, the Precision of the Combined-CNN model is improved by 1.9%, 1.4%, and 1.03%, respectively. The model training time is about 0.25 times the average time of these three models, indicating that the Combined-CNN model can also save resources and cost effectively with the best precision.

For further analysis, we compared the Combined-CNN model and the GRU model with the best classification results among the five deep learning algorithms. The training accuracy and loss of different models are shown in Fig. 9.

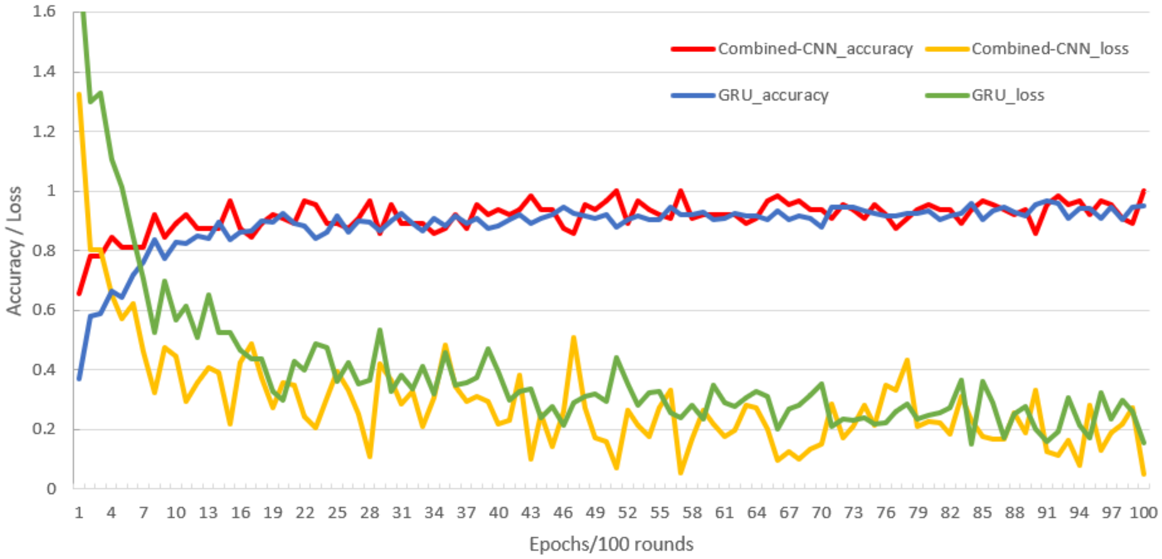

Fig. 9. Training accuracy and loss of Combined-CNN and GRU

It can be seen from Fig. 9 that the accuracy values of the Combined-CNN model and the GRU model are very close. The overall accuracy value of the former is slightly higher, and the convergence speed of the loss value is faster. In addition, it is found in the experiment that the Combined-CNN model saves a lot of time in training time compared to the GUR model. Thus, Combined-CNN model algorithm is effective in Chinese news text classification.

(3) The overall average of Precision, Recall and F-Measure of the Combined-CNN model on different data sets are shown in Table 4.

Table 4. Classification results of different data sets

Data set	Precision	Recall	F-Measure
Unbalanced	0.9369	0.9368	0.9368
Balanced	0.9557	0.9544	0.9540

According to the comparison in Table 4, in the case of using the Combined-CNN model, the Precision obtained on the balanced data set is as high as 95.57%. Compared with the unbalanced data set, the classification effect obtained with the balanced data set is better, and the Precision increased by 1.88%, the Recall increased by 1.76%, and the F-Measure increased by 1.72%. This shows that the unbalanced data set is processed again to obtain a balanced data set, which can solve the problem of misclassification of a category with a small proportion of samples into a category with a large proportion of samples. Therefore, the data set is too unbalanced to have a greater impact on the classification results, and the equalization of the data set can further improve the accuracy of news classification.

For further analysis, we compared the classification results of the Combined-CNN model on balanced and unbalanced data sets. The training accuracy and loss of different data sets are shown in Fig. 10.

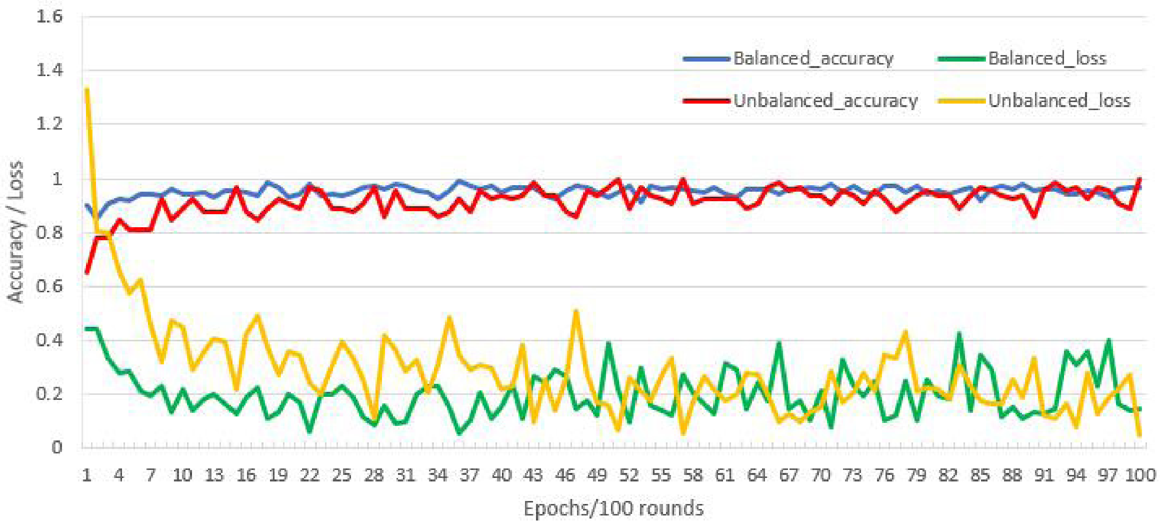

Fig. 10. Training accuracy and loss of different data sets

It can be seen from Fig. 10 that the initial accuracy of the Combined-CNN model is very high on different data sets. It rises rapidly to stable, and the loss value gradually decreases to stable, but the convergence rate is different. Due to the optimization of the proportion of data samples in the balanced data set, the overall accuracy value is very good. The loss value is obviously better, the convergence speed is faster and the fluctuation is smaller. Therefore, the classification effect of the Combined-CNN model on balanced data set is better than that on unbalanced data set, which can improve the accuracy of news text classification on balanced data.

5 Personalized Recommendation System for News Classification

Based on the realization of Chinese news text classification, the Combined-CNN is applied to realize a personalized recommendation system for news classification. The system uses WeChat applet as the visualization platform of the recommendation system, and establishes an application combining news reading and music recommendation. The system can not only automatically categorize news, but also meet the habitual needs of contemporary people to listen to music when reading news. The recommendation system helps people to select potentially interesting music from a large number of songs. The time cost of the recommendation service is low, and the expense is small. Therefore, the personalized recommendation system is easily accepted by users.

5.1 The Framework of the System

Development environment: WeChat developer tool. The architecture and design of the applet are shown in Fig. 11. The structure of the applet is mainly divided into three levels: rendering layer, logic layer, and system layer. The rendering layer builds the skeleton of the page through WXML, and WXSS builds the style of the page. The logic layer uses JS scripts to write the logic of the program. The JS of the applet is used for logic processing, data request, API interface call, etc. The system layer mainly has the functions of communication, storage, network request, etc. The rendering layer and the logic layer run in two threads respectively, and communicate through the “JSBridge” of the system layer. The rendering layer notifies the triggered events to the logic layer for business processing, and the logic layer notifies the data changes to the rendering layer and triggers the update of the WebView pages.

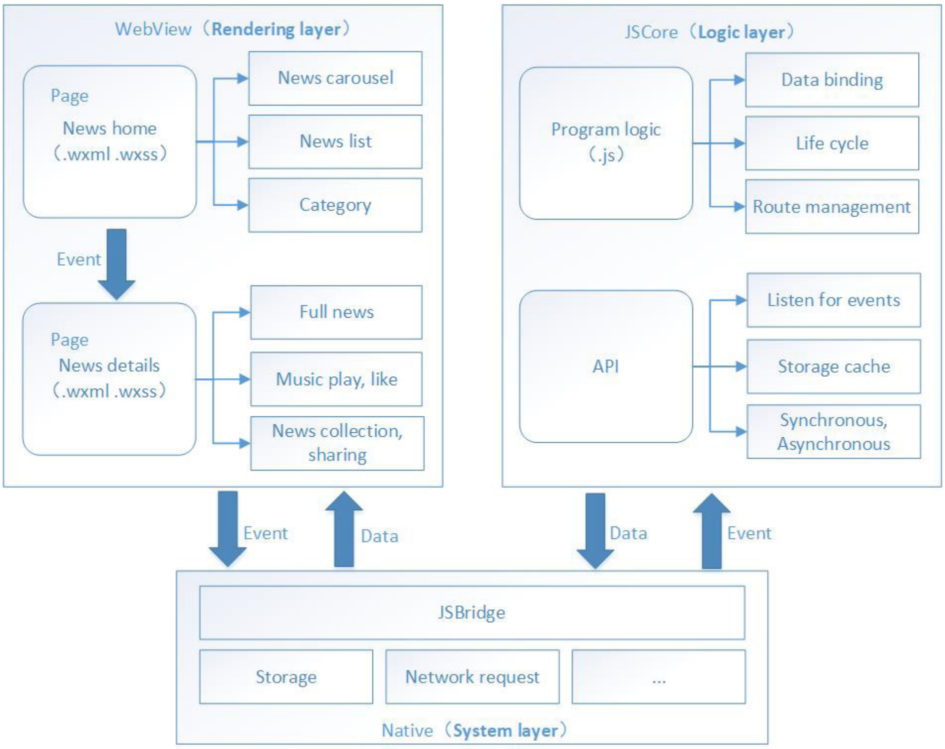

Fig. 11. The architecture and design of the applet

5.2 The Realization of the System

The personalized recommendation system for news classification applies the Combined-CNN algorithm to the automatic classification of Chinese news texts, which solves the problem of information overload caused by a large number of news data. Moreover, a visual news and music interface is established to improve the user experience in practical application.

The recommendation system of the applet realizes the news home page and news detail page. In the news home page, users can see the overview of the news, and the news is displayed in the way of carousel and list. At the same time, the Combined-CNN algorithm is used to predict the category of the news text in the data set for display. The result of the news text classification is matched with the classified music library, and a song is randomly recommended to the user from the music library of the same category. In the news detail page, users can browse the complete news, and can bookmark and share the news. Whether the recommended background music is played or not is determined by the user. For recommended background music, users can add it to “My favorite music” if they like it. In the future, you browse the news will give priority to recommending the corresponding category music users like. The background of the applet is mainly through the database storage and event monitoring function for statistics and implementation.

6 Conclusion

In this paper, we use the method of constructing data index to make a vocabulary suitable for Chinese long text classification, and use the Word2Vec to map the vocabulary semantics to real number vector. Based on the classical CNN model, we propose an improved Combined-CNN model. The classical superimposed convolution is improved to separate convolution and recombination, which makes the extraction of local features of text blocks more comprehensive. From the experimental results, it can be seen that the Combined-CNN model has a good improvement on news text classification with a precision rate of 93.69%. The Combined-CNN model improves the accuracy rate by up to 11.82% and 1.9% compared to the traditional machine learning methods and deep learning algorithms, and it is better than the comparison model in Recall and F-Measure. In addition, the Combined-CNN model has achieved better classification results on the balanced data set. However, since the news data in reality cannot be balanced, there is a shortcoming of an ideal data set and insufficient generalization. Finally, the

Combined-CNN algorithm is applied to the automatic classification of Chinese news texts, and a personalized recommendation system for news classification is realized. In the next step, we will try to use the model on more data sets, and calculate the weight of the sample data for the training of the classification model to reduce the dependence of the model on the data set. Then, we also intend to combine the GRU model or the BERT model with the Combined-CNN model to build a classification system for ensemble learning.

Acknowledgement

This work was supported by National Key R&D Program of China (No. 2016YFC0600901), and State Key Laboratory for GeoMechanics and Deep Underground Engineering & Institute for Deep Underground Science and Engineering (XD2021022), and BUCEA Post Graduate Innovation Project (No. PG2020051), and Scientific Research Project of the Beijing Educational Committee (No. KM201910016008), and Fundamental Research Funds for Beijing Universities of Civil Engineering and Architecture (No. X20069 and ZF15073), and Institute of Energy and Environmental Sciences of Beijing Coal Mine Innovation Alliance (No. 00362016055), and Excellent Teachers Development Foundation of BUCEA (No. 21082718041). We are grateful for the support of these organizations.

References

[1] Y. Jiang, Y. Song, High-Accuracy Offline Handwritten Chinese Characters Recognition Using Convolutional Neural Network. Journal of Computers 31(6)(2020) 12-23.

[2] Y. Wang, H. Li, Z. Wu, Attitude of the chinese public toward off-site construction: a text mining study, Journal of Cleaner Production 238(11)(2019) 117926.

[3] C. Liu, X. Wang, Quality-related English Text Classification Based on Recurrent Neural Network, Journal of Visual Communication and Image Representation 71(8)(2020) 102724.

[4] T.H. Ling, D.Q. Sheng, Y.L. Ping, S.Y. Jie, L.M. Yu, Slda-tc:a novel text categorization approach based on supervised topic model, Acta Electronica Sinica 47(6)(2019) 1300-1308.

[5] W. Liao, Y. Wang, Y. Yin, X. Zhang, P. Ma, Improved sequence generation model for multi-label classification via CNN and initialized fully connection, Neurocomputing 382(3)(2020) 188-195.

[6] Y. Liang, H. Li, B. Guo, Fusion of heterogeneous attention mechanisms in multi-view convolutional neural network for text classification, Information Sciences 548(2)(2021) 295-312.

[7] X. Yang, S. Xu, H. Wu, R. Bie, Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network, Procedia Computer Science 147(2)(2019) 361-368.

[8] W. Chen, K. Shi, Multi-scale Attention Convolutional Neural Network for time series classification, Neural Networks 36(2021) 126-140.

[9] J. Wang, Y. Yang, S. Wang, Context-aware Personalized Crowdtesting Task Recommendation, IEEE Transactions on Software Engineering 99(2021) 1-1.

[10]M. Asenova, C. Chrysoulas, Personalized micro-service recommendation system for online news, Procedia Computer Science 160(11)(2019) 610-615.

[11]P. Liu, H. Zhao, J. Teng, Parallel naive bayes algorithm for large-scale chinese text classification based on spark, Journal of Central South University 26(1)(2019) 1-12.

[12]Z. Chen, L.J. Zhou, X.D. Li, The Lao Text Classification Method Based on KNN, Procedia Computer Science 166(3) (2020) 523-528.

[13]O. Mabrouk, L. Hlaoua, M.N. Omri, Exploiting ontology information in fuzzy SVM social media profile classification, Applied Intelligence 51(11)(2020) 3757-3774.

[14]Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, Journal of Machine Learning Research 3(2) (2003) 1137-1155.

[15]R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proc. International Conference on Machine Learning (ICML), 2008.

[16]T. Mikolov, I. Sutskever, K. Chen, Distributed representations of words and phrases and their compositionality, arXiv preprint arXiv:1310.4546, 2013.

[17]F.F. Qiang, Z.J. Zhao, H.W. Zhong, Research on m-sequence estimation based on CNN, Journal of Computers 32(6) (2021) 134-143.

[18]Z. Yang, J. Zheng, Research on Chinese text classification based on Word2vec, in: Proc. IEEE International Conference on Computer and Communications (ICCC), 2016.

[19]B.K. Barakat, A.R. Seitz, L. Shams, The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted, Cognition 192(2)(2013) 205-211.

[20]R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training Recurrent Neural Networks, in: Proc. International Conference on Machine Learning (ICML), 2013.

[21]O. Irsoy, C. Cardie, Deep recursive neural networks for compositionality in language, Advances in Neural Information Processing Systems 27(3)(2014) 2096-2104.
[22]K.S. Tai, R. Socher, C.D. Manning, Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, arXiv preprint arXiv:1503.00075, 2015.
[23]W.T. Yih, X. He, C. Meek, Semantic parsing for single-relation question answering, in: Proc. Annual Meeting of the Association for Computational Linguistics (AMACL), 2014.
[24]X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, arXiv preprint arXiv:1509.01626, 2015.
[25]W. Gao, Z. Yang, H. Wang, A Semantic Enhanced Topic Model Based on Bi-directional LSTM Networks, Journal of Computers 30(6)(2019) 60-72.
[26]A. Vaswani, N. Shazeer, N. Parmar, Attention is all you need, Advances in Neural Information Processing Systems 30(2017).
[27]J. Devlin, M.W. Chang, K. Lee, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018.