Journal on Big Data
DOI: 10.32604/jbd.2022.027717
Article

Tech Science Press

Chinese News Text Classification Based on Convolutional Neural Network

Hanxu Wang and Xin Li*

Department of Information Technology and Cyber Security, People's Public Security University of China,
Beijing, 102623, China
*Corresponding Author: Xin Li. Email: lixin@ppsuc.edu.cn

Received: 25 February 2022; Accepted: 26 March 2022

Abstract: With the explosive growth of Internet text information, the task of text classification is more important. As a part of text classification, Chinese news text classification also plays an important role. In public security work, public opinion news classification is an important topic. Effective and accurate classification of public opinion news is a necessary prerequisite for relevant departments to grasp the situation of public opinion and control the trend of public opinion in time. This paper introduces a combined-convolutional neural network text classification model based on word2vec and improved TF-IDF: firstly, the word vector is trained through word2vec model, then the weight of each word is calculated by using the improved TF-IDF algorithm based on class frequency variance, and the word vector and weight are combined to construct the text vector representation. Finally, the combined-convolutional neural network is used to train and test the Thucnews data set. The results show that the classification effect of this model is better than the traditional Text-RNN model, the traditional Text-CNN model and word2vec-CNN model. The test accuracy is 97.56%, the accuracy rate is 97%, the recall rate is 97%, and the F1-score is 97%.

Keywords: Chinese news text classification; word2vec model; improved TF-IDF; combined-convolutional neural network; public opinion news

1 Introduction

1.1 Research Background and Significance

The rapid development of the Internet heralds the arrival of the big data era. The wide application and rapid development of the Internet will bring a large number of written materials. It is an important and arduous task to classify these text data systematically. Chinese news text is an important form of text data. The classification of Chinese news text is also an important topic in the task of text classification. As an information processing technology for efficient information retrieval and large-scale data development, text classification technology plays a very important role in the organization and management of text materials and information.

In public security work, in the face of endless news text data every day, news text classification is of great significance to the work of public security departments [1–3]. By classifying the news text, we can effectively and accurately extract the news related to the case and public opinion, so that the personnel of the public security department can grasp the public opinion situation faster and more effectively, and respond and deal with it in time, so as to stabilize public opinion and reduce the risk of major social events.

Text classification technology is a classical topic in the field of natural language processing. Its main purpose is to assign labels to target sentences. With the development of mobile Internet, the number of texts increases exponentially, and automatic text classification has gradually become the mainstream method. Automatic text classification can be roughly divided into three types: rule-based, machine learning and deep learning, and hybrid methods.

In recent years, machine learning, especially deep learning related methods, have become popular in text classification. When learning text features, deep learning model can find some hidden rules or patterns that are difficult to define. This kind of method usually includes two main steps: first, construct an appropriate word vector to represent the text input in the task; second, select an appropriate model to train and extract text features, and classify the text through these features. This paper mainly analyzes that word2vec is used to train the word vector of each word, and convolutional neural network (CNN) model is used as a text classification and analysis model to classify and analyze the Chinese news text classification data set, and has successfully achieved good practical learning results.

1.2 Journals Reviewed

Text classification began in the 1950s. At that time, the main principle of text classification was based on the rules and rules defined by experts themselves. In the 1980s, the expert system based on knowledge engineering was widely used in the processing of text classification tasks. In the 1990s, the classification analysis method of machine learning was widely used in the task of text classification. This method realizes text classification based on artificial feature engineering and shallow classification model. Traditional machine learning classification methods include nearest neighbor classification (KNN) [4], naive Bayesian classification (NB) [5] and vector machine classification (SVM) [6].

Now, word vectors and deep learning [7] are mostly used for text classification. In 2014, the application of convolutional neural network in text and sentence classification was proposed [8], which has important practical significance for the first application of convolutional neural network (CNN) in text classification task. CNN model can capture the characteristics of local correlation, which makes it successful in the field of computer vision. In the task of text classification, CNN [9–13] can be used to extract key information similar to n-gram in sentences. Zhang et al. [14] conducted a large number of comparative experiments on the proposed CNN model, obtained some experience on the setting of super parameters, and compared the effects of different super parameters on the stability and performance of CNN model. Johnson et al. [15] proposed a new CNN structure DPCNN, which can effectively extract the remote relationship features in the text and avoid the stacking of complexity. In order to improve the precision of Chinese news text classification, Zhang et al. [16] proposed and implemented a combined-convolution neural network model to study the impact of different convolution and pooling operations on the classification results. By convolution and recombination of word vectors respectively, more comprehensive local text feature information can be extracted, which performs well in text classification. RNN model is a serialized neural network model, which can be directly used to extract the data in the text context. However, in some long serialized or complex texts, it is prone to some problems of gradient disappearance and gradient explosion. Based on RNN model, a long short-term memory network model (LSTM) [17] is proposed to solve the problems of gradient disappearance and gradient explosion of traditional cyclic neural network. Chung [18] proposed the cyclic gate unit (GRU) model based on the LSTM model. GRU is an improvement based on the LSTM and is trained by the GRU neural network model. The results show that the algorithm plays an obvious role in improving the performance of text classification. With the introduction of attention mechanism, the classification model has been further improved and optimized. Sun [19] combined GRU model with attention mechanism and established GRU attention model to identify the importance of words in the text through training word vector and give weight, so as to extract important features in the text. Lan et al. [20] integrated LSTM model with attention mechanism and established LSTM attention model. Based on LSTM model, it made up for the defects of classical model and further improved the effect of text classification.

In terms of text representation, Mikolov et al. [21,22] proposed a continuous word bag model (CBOW), which uses context words to predict intermediate words. CBOW model obtains the hidden layer by adding the word vector in the context, and uses the hidden layer to predict the probability of intermediate words. Mikolov et al. Also proposed a Skip-Gram model. Different from the principle of CBOW model, Skip-Gram model predicts the probability of words in the context through intermediate words. Wang et al. [23] proposed the improved TF-IDF algorithm of class frequency variance, and analyzed the weight of each word vector in the text, and constructed the text vector representation based on word vector and weight by combining word vector and weight, so as to improve the effect of text classification.

1.3 Research Method

In this paper, word2vec is used to train the word vector, and the improved TF-IDF algorithm of class frequency variance is used to construct the text vector representation which can better represent the text content based on the word vector. The deep learning framework of combined-convolution neural network is used to classify the news texts. The classification results are compared with GRU model, LSTM model, text CNN model and CNN model of word2vec pre-trained word vector. Through the comparative experiments, the performance of the combined-convolution neural network text classification model based on word2vec and improved TF-IDF in Chinese news text classification task is compared and analyzed.

2 Relevant Technical Basis

2.1 Chinese Word Segmentation

All Chinese news texts are composed of several words and sentences. To realize the function of computer recognizing Chinese text through in-depth learning, we should correctly divide the text. This is the process of text word segmentation. Unlike English word segmentation, English words are separated by spaces. However, in Chinese sentences, words are closely connected, so it is more difficult and challenging to segment Chinese text.

Word segmentation technology It belongs to the application category of natural language processing technology. People can identify and distinguish words in a sentence through their own experience and common sense, and the main task of word segmentation algorithm is to make computers understand and process sentences like people. At present, there are three main types of automatic word segmentation technology based on Chinese: word segmentation method based on string matching, word segmentation method based on understanding and word segmentation method based on statistics.

2.1.1 Word Segmentation Method Based on String Matching

This method first establishes a unified dictionary table. When dividing a sentence, divide the sentence into several parts, and each part matches the dictionary. If the word is in the dictionary, complete the word segmentation. If it is not in the dictionary, continue the segmentation until the matching is successful. This method is fast and easy to implement, but the processing effect of ambiguous words and non entered words is not good.

2.1.2 Understanding Based Word Segmentation

This method simulates human's understanding of sentences, so that the computer can analyze the semantic information and grammatical rules of sentences, so as to complete word segmentation. Because the basic knowledge of modern Chinese is extremely complex, it is very difficult to transform all kinds of language materials into a form that can be read directly by machines. Therefore, this method can only be tested and cannot be widely used.

2.1.3 Word Segmentation Method Based on Statistics

Based on the corpus, the method segments sentences and counts the occurrence probability of words composed of adjacent words. The more adjacent words, the higher the probability of occurrence. Word segmentation of sentences according to probability.

2.2 Tensorflow

Tensorflow is a digital symbolic intelligent mathematical programming system based on data information flow. It is widely used in various data programming with machine intelligent learning programming algorithms. Its predecessor is Google's neural network algorithm library.

Tensorflow has a multi-layer structure, which can be directly run and deployed on various network servers, PC mobile terminals and various web pages. It supports various GPUs and TPUs, and can carry out various high-performance numerical analysis and calculation. It has been widely used in various fields, including information product technology research and development and information science and technology exploration within Google Group.

With the help of tensorflow deep learning framework, this paper constructs a convolutional neural network model to realize data training and testing. Tensorflow provides many functional modules with superior performance, which greatly improves the programming efficiency.

3 Text Classification Theory

3.1 Basic Steps of Text Classification

Text classification technology is the process of automatically distributing the given text content and providing it to one or several categories. As a supervised learning method, the labeled samples are used as the training set to train the optimal classifier, so as to achieve the purpose of automatic classification. The text classification process can be defined as a set of text data a = {A1, A2, . . . , an}. Each text should have its own corresponding text category (one or more). The predefined text category set is b = {B1, B2, . . . , BM}, where m is the total number of all predefined text categories, and all text data are divided into text training set and text test set. The training samples are trained with the classification model, and the samples of unknown categories are predicted with the trained model, so as to provide a security guarantee for the prediction samples to continue to predict their corresponding categories. The specific process is shown in Fig. 1 below.

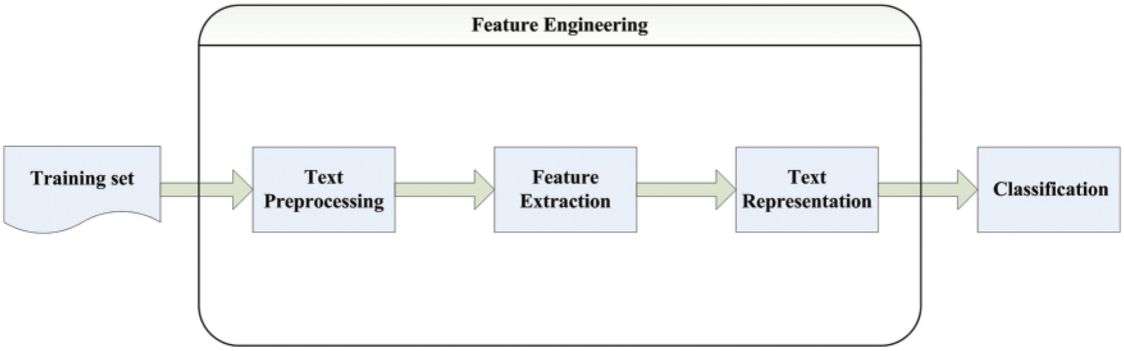


Figure 1: Basic flow of text classification task

The text classification process mainly includes several important links such as text preprocessing, feature extraction, Chinese text vector representation and text classification. Finally, the test set also needs to undergo text preprocessing, feature extraction and other operations to test the classification performance of the classifier, and the parameter values of the classifier can be adjusted according to the test results, Make the classifier better realize the function of text classification. This chapter will introduce the above links step by step.

3.2 Text Preprocessing

Text preprocessing is a basic work of text classification task. Through text preprocessing, all text data sets are processed and transformed into the required format for subsequent processing.

3.2.1 Coding Specification

Chinese codes mainly include GBK, GB2312, Big5 and UTF-8. There are many text coding models in the sample set. In order to facilitate subsequent text processing, it is necessary to standardize the characteristics of text coding methods. In this paper, the data in all sample sets are stored in UTF-8 format.

3.2.2 Remove Illegal Characters

Today's basic research on text classification mainly selects a data set, which is a data set in web page format. The common acquisition form of news text is to crawl news web page information through crawlers. In addition to news text information, the text crawled in web pages may also include some interference information irrelevant to classification, such as links, pictures, garbled codes and so on, Compared with the pure text information we need to extract, this kind of interference information is regarded as noise information. Therefore, this kind of noise information should be removed in the text preprocessing stage.

3.2.3 Chinese Word Segmentation

The common means of text word segmentation are: using word segmentation tools to segment the text directly, using existing dictionaries to segment words, and establishing word segmentation model through algorithm. This paper mainly uses a Jieba sentence word segmenter based on Hidden Markov algorithm to accurately segment each sentence. Jieba's sentence word segmentation is based on the dictionary in dictionary statistics. Through calculation, a set of dictionary table with prefix derived by construction method is established for sentence segmentation. In this way, all sentence segmentation word formats can be calculated directly. According to the correct position of segmentation, a directed acyclic graph is established. The path with maximum probability is calculated by using the algorithm of dynamic segmentation planning graph, and the final segmentation form is obtained.

3.2.4 Delete Stop Word

Delete stop words, that is, delete words that appear frequently in the text but have little practical significance for the text features, such as common adverbs such as "you", "I", "we", "nothing", "Bu Du", quantifiers, prepositions, exclamations, numerals, etc. These words are easy to cause noise. After word segmentation, according to the pre-established stop words list, these stop words are filtered out from the word segmentation results by character matching scanning.

3.3 Feature Extraction

After text preprocessing, text data exists in the form of feature item set. If a feature item set is adopted directly, it will cause the disaster of dimension of the model and have a great impact on the final classification results. At the same time, in the process of text classification, not all feature words play a role. Only in special semantic cases can they fully reflect the value of their practical application. The whole process of feature extraction is a process of regularizing text semantics. Only a regularized semantics can be expressed by computer. Therefore, the dimensionality reduction of feature set data, the selection and extraction of feature item data with large contribution in text information, and the final classification effect also play an important role. At present, the common text feature selection methods mainly include: information increment method, document frequency mutual information method, evidence weight and chi square test.

3.4 Chinese Text Vector Representation Model

The problem of text representation is to transform a text represented in the form of string into a form that can be recognized by computer (i.e., an easy to process vector). The problem of directly transforming a string into a vector is an important core of the problem of text representation. When training the model, the text needs to be expressed in the form of digital vector or matrix in advance. The text representation methods can be roughly divided into four categories: Boolean model, vector space model, probability model and word embedding model.

3.4.1 Boolean Model

Boolean model is considered to be the simplest model for text representation. It marks and writes text in the form of a set of binary variables, that is, the value is 1 or 0. If a feature word appears in the current document, the feature value is marked as 1. If it does not appear, it is marked as 0.Although the implementation of Boolean model is simple and fast, its language expression ability for the text is relatively poor. It cannot accurately express the frequency of feature words in the context, and can not fully reflect the importance of feature words to the text in the context.

3.4.2 Vector Space Model

VSM model converts text into vector form. Each element in the vector represents weight value, which can be calculated by TF-IDF and other algorithms. Words are defined by the dictionary. If a word exists in the dictionary, its weight value is calculated. If it is not in the dictionary, the weight value is recorded as 0. Therefore, the obtained text vector dimension depends on the number of words in the dictionary, but it may lead to sparse text vectors. In addition, the vectorization of text in VSM Model ignores the positional relationship between words.

3.4.3 Probability Model

The probability model comprehensively considers the correlation between each feature word and between each feature word and the text, and gives the probability that each feature word may appear in the relevant text or irrelevant text. The system can calculate these probabilities and make decisions according to all statistical data. The conditional probability model of the document is obtained, and the maximum category in the conditional probability of the document is taken as the output category of the document.

3.4.4 Word Embedding Model

A simple representation of word vectors is One-Hot representation, which uses a very long vector composed of 0 and 1 to represent words. This method uses the number of words in the dictionary to describe the length of the vector. The position of 1 in the vector corresponds to the position of the word in the dictionary, and the other positions are 0. However, it is prone to dimensional disasters and cannot describe the similar information and positional relationship between words, which is the main problem of this method.

Word2vec is one of the word embedding models first proposed by Google in 2013. It can train word vectors very quickly, make certain connections between words, and deeply explore the relationship between words. The basic design idea of word2vec is to use a three-layer neural network to train the input words. After calculating the mapping relationship of the model, the word vector is generated by transformation matrix. In general, word2vec can be subdivided into CBOW (continuous word bag) and Skip-Gram. CBOW model predicts the generation probability of current feature words according to context related words, and Skip-Gram model predicts the generation probability of each word in the context according to the current word.

This paper adopts the CBOW model in word2vec word vector coding. The goal and task of CBOW model is to predict the maximum probability that the output word is the center word w_t of the window according to the given words $w_{(t-1)}$, $w_{(t-2)}$, $w_{(t+1)}$ and $w_{(t+2)}$ in a certain relative neighborhood and radius of a center word W_t (usually referred to as windows window, where the window size is 5, that is, the given center word is the center of the window and 2 is the relative neighborhood radius). In the implementation process, because the word order relationship is not considered, it is called the word bag model, as shown in Fig. 2 above.

In Fig. 2, the symbols $w_{(t-1)}$, $w_{(t-2)}$, $w_{(t+1)}$ and $w_{(t+2)}$ are used to represent the input words. In fact, they are vectors represented by unique heat coding. It is easy to establish the input word index by establishing a dictionary by yourself. The dimension of the vector is the same as the number of words in the dictionary. Only one dimension has a value of 1, which corresponds to the position of the word in the dictionary index, and other dimensions have a value of 0. The PROJECTION layer is calculated through the query table. First, you need to initialize a word vector matrix W. W is a two-dimensional word vector matrix, the number of rows is equal to the number of words in the dictionary, and the number of columns is a super parameter set by a person for formulation.

The input is the context word of the central word w_t, which is coded by One-Hot representation. Assuming that windows size is 5, there are four input vectors, $w_{(t-1)}$, $w_{(t-2)}$, $w_{(t+1)}$ and $w_{(t+2)}$, coded by One-Hot representation .Let the input layer matrix be W_{tn}, and the size can be expressed as $|V| \times d$ by the formula. $|V|$ is the size of the dictionary, d is the dimension of the word vector, and v vector is a line of W_{tn}. The lookup process is as follows:

$$w_t^T W_{tn} = v_t \tag{1}$$

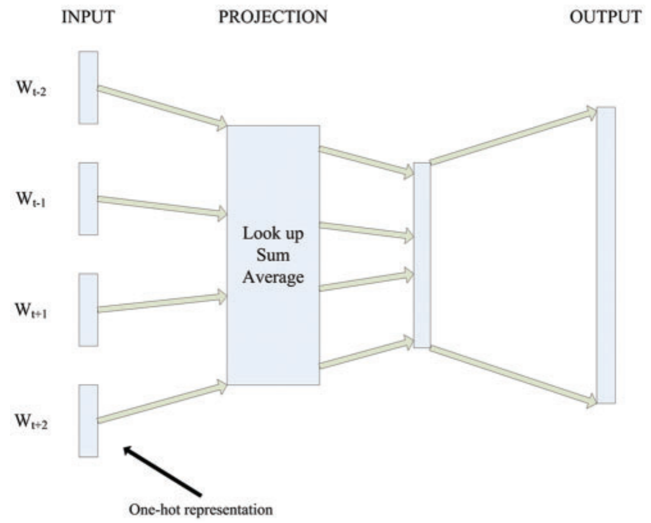


Figure 2: Word vector processing flow of CBOW model

Vector multiplication between a One-Hot coded vector w_t and W_{tn} is to take out a row corresponding to w_t in the input matrix W_{tn}, and the row number of this row is the index number of the word w_t in the dictionary. The vector obtained through the input layer operation is a dense vector, which is assumed to be P_{middle}. The matrix vector of the output layer is set to W_{out}, and the size of the matrix in the output layer is $d \times |V|$, then the output vector is as follow:

$$P_{out} = P_{middle}^T W_{out} \tag{2}$$

P_{out} has a size of $1 \times |V|$, the value of each dimension can be understood as the logit probability of each word in the dictionary under the current context $w_{(t-1)}$, $w_{(t-2)}$, $w_{(t+1)}$ and $w_{(t+2)}$. After softmax, it is the probability of each word. The above description is also a forward propagation process of the CBOW model.

3.5 Text Classification Model

This paper classifies Chinese News Texts Based on the convolutional neural network model in deep learning, so I will focus on the convolutional neural network classification model in detail.

Convolutional neural network model is widely used in various image signal processing technology fields. Convolutional neural network is mainly composed of multilayer neural network connection. Based on the two-layer design of data input layer and data output end, it is also necessary to design a feature extraction layer for extracting data features and a layer that can map the feature extraction of the input layer to. Compared with other traditional neural networks, convolutional neural network can better realize the two basic functions of weight information sharing and local weight perception, and greatly reduce the number of processing parameters that may be required by traditional neural networks in specific training. In the main task of studying natural language processing, convolutional neural network inputs and operates in the unit of word vector. In their paper, Zhang et al. Proposed a

CNN model that can be used to complete the task of text classification and analysis. Its basic structure is shown in Fig. 3 below.

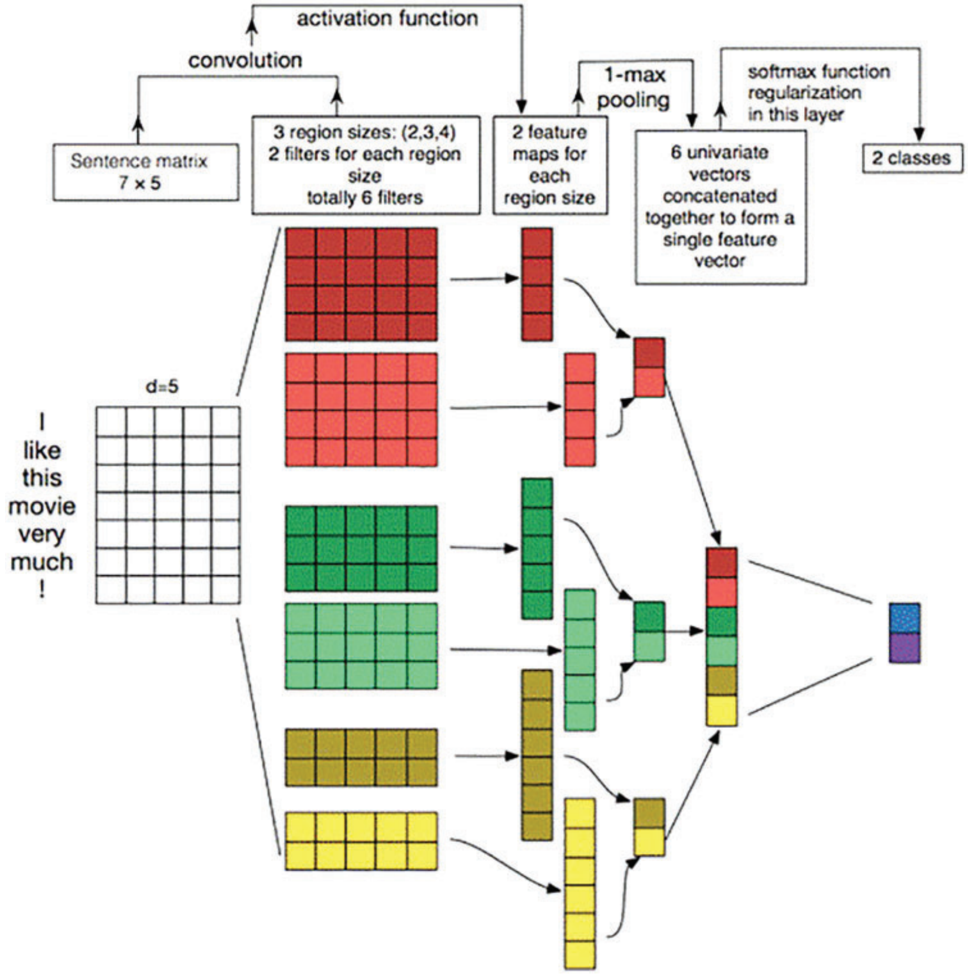

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Figure 3: CNN model for text classification task

Through in-depth study of existing methods and in order to achieve the goal of fast and accurate text classification, this paper integrates word2vec word vector coding method with convolutional neural network (CNN) model, so as to reduce the word vector dimension, improve the efficiency of CNN model training and save model training time, At the same time, compared with One-Hot coding, word2vec word vector coding pays more attention to the relationship between text context, deepens the relationship between word vectors, enriches the feature representation, and improves the effect of text classification to a certain extent.

As shown in the figure above, the hierarchical structure of convolutional neural network can be divided into word embedding layer, convolution computing layer, pooling layer and FC layer. Each layer will be introduced below.

3.5.1 Word Embedding Layer

In the field of image recognition, the input layer of convolutional neural network is the pixel of the image, and the pixel of a single object in the picture is continuous. When the input layer slides, the convolution kernel can easily extract the object features in the picture; In addition, the pixels of the picture form a matrix, which is convenient for convolution kernel sliding to extract features. However, for text data, the words in the text are discrete, such as "although... But...". Although "and" but "are obviously related, but they may be far apart in the sentence. It is difficult to extract the relationship between such words with long-distance dependence by convolution kernel.

In addition, compared with the natural density of the elements in the pixel matrix of the image, the bag of words model is generally used to represent the text vector in the traditional text classification method in machine learning. It has the characteristics of high-dimensional and sparse, which is not suitable for CNN input. After the exploration of experts and scholars, we finally put forward the concept of word embedding layer. Word embedding is actually to map each word through space and convert One-Hot representation into distributed representation. Therefore, we can get low-dimensional and dense word embedding vectors to represent each single word discourse. In this way, every single word can only be a one-dimensional vector, and a discourse sentence can only be expressed by using several one-dimensional vectors.

3.5.2 Convolution Computing Layer

A sentence passing through the embedding layer actually forms a matrix. For example, "We, cannot, be, late" can be transformed into a matrix of $4 \times N$, and N is the word vector dimension. Different from two-dimensional convolution in image processing, one-dimensional convolution is used in text processing.

One dimensional convolution requires multiple filters with different widths to obtain different local receptive fields. As shown in Fig. 3, the network is equipped with convolution cores of 2, 3 and 4 sizes, and there are two convolution cores of each size. When the general convolution structure is used for text classification, the features it can extract are n-gram features. In Fig. 3, the input layer obtains six characteristic graphs with dimension 1 through the convolution window.

3.5.3 Pool Layer

Pooling is to convert inputs of different lengths into outputs of uniform length.

In this paper, the pooling method adopted by the convolutional neural network model is maximum pooling. This method simply puts forward the maximum value from the previous one-dimensional characteristic graph. It is explained that the maximum value represents the most important signal. It maximizes the pool of vectors obtained from each channel of the convolution layer to obtain a scalar. Therefore, it can be seen that the convolution network can only extract whether there is an n-gram feature in a sentence. In this way, there will be as many maximum scalars as the convolution kernels, and then these scalars are spliced into a vector. The vectors obtained by convolution kernels of all sizes are spliced again to obtain a final one-dimensional vector.

3.5.4 Full Connection Layer

The one-dimensional vector output by the pooling layer is connected to a softmax layer through full connection, and the final result is output after being activated by softmax.

4 Construction of the Combined-Convolution Neural Network Model Based on Word2vec and Improved TF-IDF

4.1 Improved TF-IDF Algorithm

TF-IDF is a classical method for calculating the weight of words in the text library. It is composed of term frequency (TF) and inverse document frequency (IDF). The word frequency calculation is shown in Eq. (3):

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{k} n_{k,j}} \tag{3}$$

Among them, $\mathrm{tf}_{i,j}$ represents the frequency of word w_i in document d_j, where the numerator represents the number of times of words in document d_j, the denominator represents the total number of times of all words in the document, and k represents the number of kinds of words in the document. The calculation of inverse document frequency is shown in Eq. (4):

$$idf_i = \log \frac{n_d}{df\left(d, w_i\right) + 1} \tag{4}$$

Among them, idf_i represents the inverse document frequency of the word w_i in the library d, n_d is the total number of documents in the text library d, and df (d, w_i) represents the number of documents containing the word w_i in the library. In order to prevent the result of denominator df(d, w_i) from being zero, the method of adding one is adopted. The calculation result of TF-IDF is shown in Eq. (5):

$$tf - idf_{i,j} = \log \frac{tf_{i,j} \times idf_i}{\sqrt{\sum^{w_i} \in d_j \left[tf_{i,j} \times idf_i\right]^2}} \tag{5}$$

It can be concluded that the weight of a word is directly proportional to its frequency in the document and inversely proportional to the number of documents containing the word in the library.

In the text classification task, the text of the text library is marked as multiple different categories. The traditional TF-IDF method calculates the weight of words in the whole text library. Here, the importance of words to a certain category and the distribution of words to different categories are ignored, resulting in the loss of some words that make an important contribution to category judgment. Therefore, TF-IDF algorithm with class frequency variance is adopted in this paper, class frequency variance reflects the distribution of words in different categories. The calculation is shown in Eq. (6):

$$\tau_i = \frac{\sqrt{\sum_{j=1}^{N} \left(\frac{df\left(d, w_i\right)}{N} - df\left(c_j, w_i\right)\right)^2}}{N} \tag{6}$$

Among them, τ_i is the class frequency variance of word w_i, N is the number of text categories, df(d, w_i) is the number of documents containing word w_i in the text library, and df(c_j, w_i) is the number of documents containing word w_i in the documents of category c_j. As can be seen, the size of τ_i represents the fluctuation degree (distribution) of word w_i in different categories. The larger τ_i is, the greater the fluctuation of the word in the category and the more uneven the distribution is, which indicates that the word has different contribution to different classes of texts and plays a greater role in judging text classification. The algorithm formulas of all improved TF-IDF are shown in Eq. (7):

$$tf - idf - \tau_{i,j} = tf - idf_{i,j} * \tau_i \tag{7}$$

4.2 Model Construction

The news text classification model in this paper is different from the classical Textcnn model. Based on the classical Textcnn model, this paper implements a six-layer combined-convolutional neural network model, as shown in Fig. 4 above.

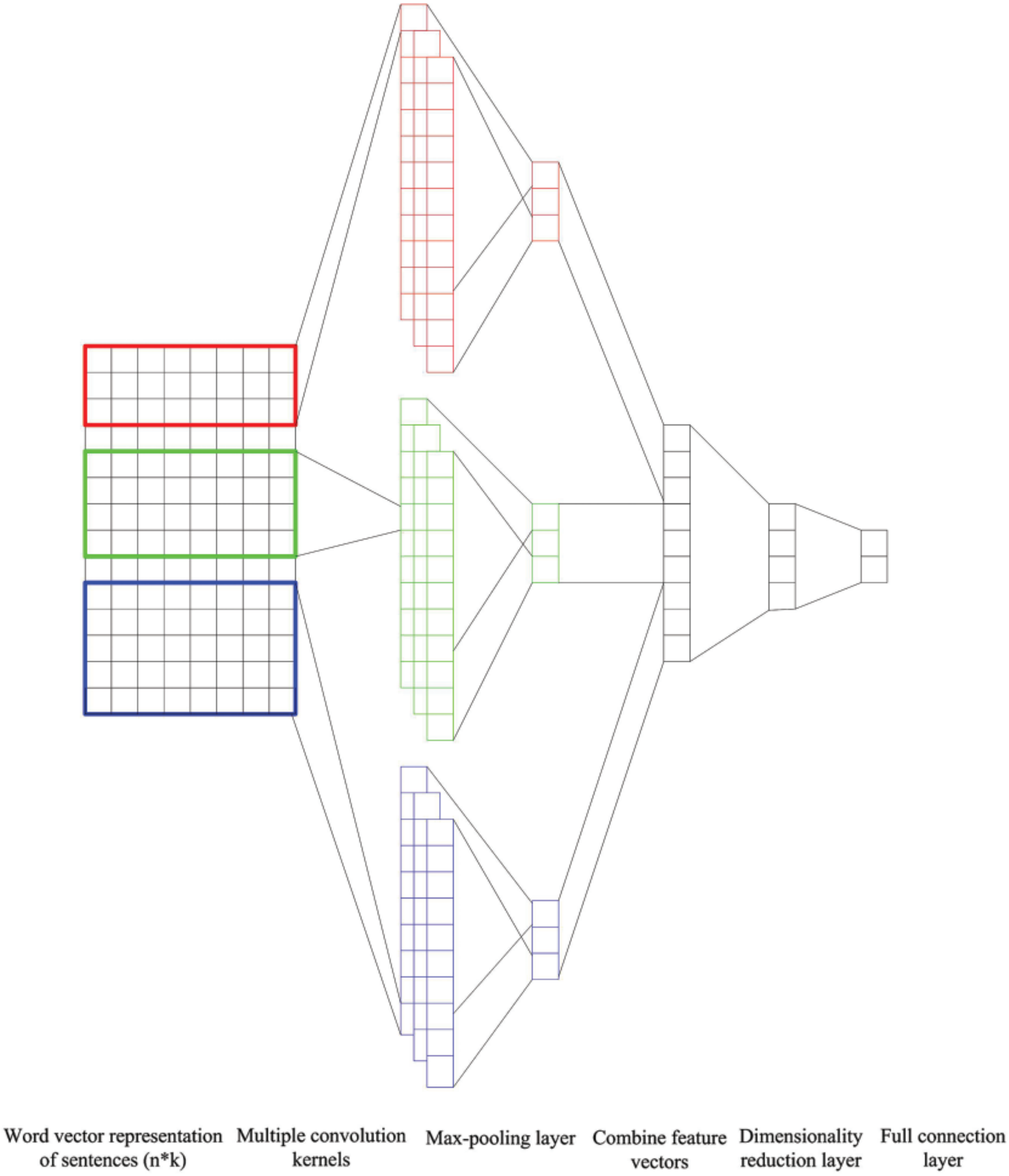

Word vector representation of sentences (n*k)
Multiple convolution kernels
Max-pooling layer
Combine feature vectors
Dimensionality reduction layer
Full connection layer

Figure 4: Six-layer combined-convolutional neural network

Layer 1: word2vec model transforms the segmented texts from high-dimensional and sparse one hot coding to low-dimensional coding. However, the word vectors cannot describe the importance of words to the texts. The improved TF-IDF algorithm is used to calculate the weight of word vectors.

The text representation adopted in this model is shown in (8):

$$vec\left(d_i\right) = \sum_{t \in W_i} V_t * tf - idf - \tau_{t,i} \tag{8}$$

Layer 2 and layer 3: convolution layer and pooling layer. Classical CNN models have different situations of single-layer convolution and multi-layer convolution. In the single-layer convolution model, the local feature information extracted by a single convolution kernel is limited. In the multi-layer convolution model, the text vector matrix is convoluted for many times, but the semantics of cutting words has no practical significance, which is not conducive to text classification. The relevant results also show that the single-layer convolution structure is better than the multi-layer convolution structure for text classification. In order to extract more local text features, this model uses three convolution kernels with different sizes to extract multiple n-gram features of texts. In order to extract the main features and reduce the number of feature parameters, the maximum pooling layer is used to maximize the output of the convolution layer.

Layer 4 and layer 5: there are no these two hidden layers in the classical CNN model. The output of the third layer is the result of three pooling operations. There are a large number of convolution kernels set for each size. Therefore, the feature vectors obtained through the pooling layer is large. A hidden layer is added to this model for dimensionality reduction.

Layer 6: full connection layer, in which dropout layer is added to prevent overfitting and improve generalization ability. Secondly, the model adopts activation function to increase the nonlinearity of the model and avoid the problem of gradient disappearance. Finally, news text classification is realized by using softmax.

5 Experiment

5.1 Experimental Preparation

This experiment uses a subset of THUCNews to traCin the model and test the classification results. The content of THUCNews mainly includes 740000 historical news material manuscripts and source files (2.19GB). All the content is completely in UTF-8 simplified text format. It is formed by optimizing and filtering all Sina historical news material content of Sina News RSS subscription channel from 2005 to 2011.This paper extracts 10 kinds of relevant news materials about global sports, finance, real estate, home, education, science and technology, fashion, current affairs, games and entertainment from 10 different news fields. In this paper, the relevant news of the training set is 50000 (5000 for each news category, a total of 50000), the relevant news of the test set is 10000 (1000 for each news category, a total of 10000), and the relevant news of the verification set is 5000 (500 for each news category, a total of 5000). The sample example is shown in Tab. 1.

Table 1: Sample data

Category	News text example
Sports	Kobe Bryant scored 24 points, the Lakers won 4–2 and advanced to Paul 10 + 11 + 8. Sad farewell to sina sports news. On April 29 Beijing time, the defending champion Lakers beat the Hornets 98–80 away . . .

(Continued)

Table 1: Continued

Category	News text example
Finance and Economics	The world top 25 private equity fund investment managers made a net income of 10 billion US dollars last year...
House Property	The chairman of Huayuan Real Estate (enterprise zone, real estate) Zhiqiang Ren said at the financial forum held by Sohu...
Home Furnishing	Sanitary ware promotion is tricky and hypes the concept of selling high-end price sanitary ware...
Education	The first "water test" of CET-4 and CET-6 increased the following reading link, and the listening ratio was significant. Yesterday, CET-4 and CET-6 began twice a year. This time, some students got two exams...
Science and Technology	Professional SLR Camera Nikon D3000 double head set sold 5710 Author: Zhang Yue Nikon D3000 is an SLR camera designed for entry-level users...
Fashion	Group pictures: jeans also have new patterns. The trend of dress up in early spring is skyrocketing. Introduction: they are at the forefront of fashion all year round. Of course, they are the first...
Current Politics	On August 13, Jinrong Liao, deputy director of the Criminal Investigation Bureau of the Ministry of public security, said that guns are illegal and criminal activities, which is very harmful to the social economy...
Game	The mobile online game pocket elf is an excellent pet. In the world of Pocket ELF, there are extremely rich pet types waiting for players to collect...
Entertainment	DreamWorks is optimistic about the prospect of the sequel to Kung Fu Panda (picture) Sina entertainment news since the global release of the Hollywood animated film Kung Fu Panda...

5.2 Evaluation Criteria

In this paper, precision, recall, F1 score and accuracy are used as evaluation indexes to evaluate the classification performance of the model.

True positive (TP): predicts the number of positive classes as the number of positive classes.

True negative (TN): predicts the number of negative classes as the number of negative classes.

False positive (FP): the number of negative classes predicted as positive classes is false positive (type I error).

False negative (FN): predict the number of positive classes as negative classes (type II error).

(1) Precision is a statistical measurement, and its calculation formula is shown in Eq. (9):

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

(2) The calculation formula of recall is shown in Eq. (10)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

(3) F1 score is an index used to measure the classification accuracy and accuracy of classification instruments. Its basic calculation formula is shown in Eq. (11):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

(4) Accuracy is the most commonly used evaluation index, that is, the number of correctly classified samples divided by the number of all samples. Generally speaking, the higher the accuracy, the better the classifier. The calculation formula is shown in Eq. (12):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

5.3 Experimental Design

This paper uses the CNN model of word2vec word vector coding to classify Chinese news texts, and evaluates the classification effect through accuracy, accuracy, recall and F1 value.

(1) The mainstream direction of deep learning text classification algorithm is cyclic neural network and convolutional neural network. Here, the traditional Text-CNN model and traditional text RNN model (GRU model and LSTM model) are compared. As the control variable, the input of the word embedding layer of the two models defines the input vector through random initialization.

(2) In order to illustrate the superiority of word2vec training word vector, the CNN model of word2vec training word vector and the traditional Text-CNN model with random initialization of word vector are compared for the classification effect of common data sets to form a control experiment.

5.4 Result Analysis

(1) This paper compares the text classification effects of RNN model and CNN model. The comparison results are shown in Tab. 2. It can be seen that CNN is better than RNN model in text classification.

Table 2: Comparison of text-CNN model with LSTM model and GRU model

Model	Accuracy	Precision	recall	F1-Score	Training time
Text-CNN	96.67%	0.97	0.97	0.97	9 min and 53 s
LSTM	77.75%	0.78	0.78	0.77	3 h, 44 min and 36 s
GRU	95.19%	0.95	0.95	0.95	6 h, 1 min and 46 s

Training:

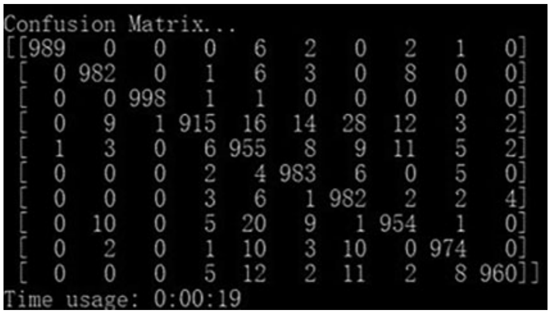
After four iterations of CNN model, the accuracy of the verification set will not be improved within 1000 rounds, and the training will be completed in advance. In the 2000 round, the best effect can be obtained in the verification set. The accuracy rate of the verification set can reach 95.86% and the loss value of the verification set is 0.16. In the training process, the accuracy rate of the training set was as high as 100% and the minimum loss value of the training set was 0.0021. The effect is very ideal. The total training time of CNN model is 9 min and 53 s, with relatively less time and faster training speed.

After three rounds of iteration of LSTM model, the accuracy of the verification set can not be improved within 1000 rounds, and the training can be completed in advance. In the 1100 round, the best effect is obtained in the verification set. The accuracy rate of the verification set can reach 73.78%, and the loss value of the verification set is 1.0. In the training process, the highest accuracy rate of the training set is 85.94%, and the lowest loss value of the training set is 0.51. Compared with the CNN model, the effect has obvious shortcomings. The total training time is 3 h, 44 min and 36 s. The training speed is slow and inefficient.

After five rounds of iteration of GRU model, the accuracy of verification set will not be improved within 1000 rounds, and the training will be completed in advance. In the 2700 round, the best effect is obtained in the verification set. The accuracy rate of the verification set can reach 92.64% and the loss value of the verification set is 0.28. In the training process, the highest accuracy rate of the training set is 100% and the lowest loss value of the training set is 0.022. The training effect is obviously better than the LSTM model. The total training time is 6 h, 1 min and 46 s. The training time is very long, about 1.5 times that of LSTM model and 40 times that of CNN model.

Test:

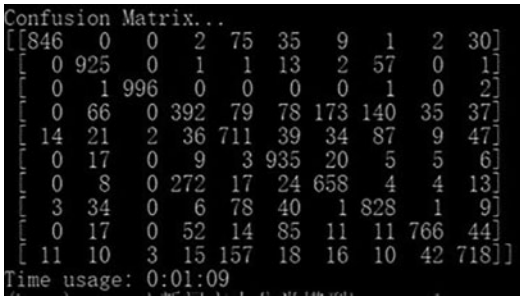
The loss value of CNN model on the test set is 0.12, the test accuracy is as high as 96.67%, the accuracy rate is 0.97, the recall rate is 0.97, and the F1 value is 0.97. In the confusion matrix, the data are mostly concentrated on the main diagonal of the matrix, the classification effect is obvious, and the total time is 19 s. The confusion matrix is shown in Fig. 5 below.



```
Confusion Matrix...
[[989   0   0   0   6   2   0   2   1   0]
 [  0 982   0   1   6   3   0   8   0   0]
 [  0   0 998   1   1   0   0   0   0   0]
 [  0   9   1 915  16  14  28  12   3   2]
 [  1   3   0   6 955   8   9  11   5   2]
 [  0   0   0   2   4 983   6   0   5   0]
 [  0   0   0   3   6   1 982   2   2   4]
 [  0  10   0   5  20   9   1 954   1   0]
 [  0   2   0   1  10   3  10   0 974   0]
 [  0   0   0   5  12   2  11   2   8 960]]
Time usage: 0:00:19
```

Figure 5: Confusion matrix in the test of the text-CNN model

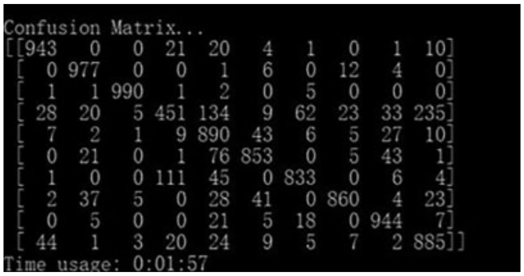
The loss value of LSTM model on the test set is 0.76, the accuracy rate of the test set is 77.75%, the accuracy rate is 0.78, the recall rate is 0.78, and the F1 value is 0.77. In the confusion matrix, the categories of “fashion” and “home”, “entertainment” and “education” intersect, taking a total time of 1 min and 9 s. The confusion matrix is shown in Fig. 6 below.



```
Confusion Matrix...
[[846   0   0   2  75  35   9   1   2  30]
 [  0 925   0   1   1  13   2  57   0   1]
 [  0   1 996   0   0   0   0   1   0   2]
 [  0  66   0 392  79  78 173 140  35  37]
 [ 14  21   2  36 711  39  34  87   9  47]
 [  0  17   0   9   3 935  20   5   5   6]
 [  0   8   0 272  17  24 658   4   4  13]
 [  3  34   0   6  78  40   1 828   1   9]
 [  0  17   0  52  14  85  11  11 766  44]
 [ 11  10   3  15 157  18  16  10  42 718]]
Time usage: 0:01:09
```

Figure 6: Confusion matrix in the test of the LSTM model

The loss value of GRU the test accuracy is as high as 95.19%, the accuracy rate is 0.95, the recall rate is 0.95, and the F1 value is 0.95. In the confusion matrix, the data distribution is mostly concentrated on the main diagonal, and the classification effect is better than LSTM model, but there is still a gap with CNN model, with a total time of 1 min and 5 s. The confusion matrix is shown in Fig. 7 below.



```
Confusion Matrix...
[[943   0   0  21  20   4   1   0   1  10]
 [  0 977   0   0   1   6   0  12   4   0]
 [  1   1 990   1   2   0   5   0   0   0]
 [ 28  20   5 451 134   9  62  23  33 235]
 [  7   2   1   9 890  43   6   5  27  10]
 [  0  21   0   1  76 853   0   5  43   1]
 [  1   0   0 111  45   0 833   0   6   4]
 [  2  37   5   0  28  41   0 860   4  23]
 [  0   5   0   0  21   5  18   0 944   7]
 [ 44   1   3  20  24   9   5   7   2 885]]
Time usage: 0:01:57
```

Figure 7: Confusion matrix in the test of the GRU model

In conclusion, the classification effect of CNN model is better than RNN model. The characteristics of parallel training make the training efficiency of CNN model high and time-consuming short. At the same time, it can also be proved that CNN model can learn more classification features on this data set, which is more advantageous than RNN model. Five samples are randomly selected from the test data to classify and predict the input text. The prediction results are consistent with the original text label, which proves that CNN model is feasible in the task of news text classification and worthy of further promotion.

(2) This paper compares the text classification effects of the classical Text-CNN, the word2vec-CNN and the combined-convolution neural network model based on word2vec and improved TF-IDF (W2V-ITI-Combined-CNN) in this paper. The comparison results are shown in Tab. 3.

Table 3: Comparison between text-CNN model and word2vec-CNN model

Model	Accuracy	Precision	recall	F1-Score
Text-CNN	96.67%	0.97	0.97	0.97
Word2vec-CNN	97.24%	0.97	0.97	0.97
W2V-ITI-Combined-CNN	97.56%	0.97	0.97	0.97

Training:

The training of word2vec-CNN model is divided into two parts: word vector training and CNN model training. Five rounds of iteration were conducted for CNN model training. The accuracy of the validation set was not improved within 1000 rounds, and the training was ended in advance. The W2V-ITI-Combined-CNN adds the process of word weight calculation on the basis of word2vec-CNN, and the training time is relatively increased, but it also achieves good results in the validation set.

Test:

The test accuracy of the word2vec-CNN model on the test set is as high as 97.24%, which is half a percentage point higher than the traditional Text-CNN model. The precision rate is 0.97, the recall rate is 0.97, and the F1 value is 0.97. The test accuracy of the W2V-ITI-Combined-CNN is 0.32% higher than word2vec-CNN.

In conclusion, the W2V-ITI-Combined-CNN has a certain improvement in the classification effect of Chinese news text compared with word2vec-CNN model and classical Text-CNN model. Compared with the traditional Text-CNN model, the word2vec-CNN model that introduces word2vec to train word vectors has improved the effect of Chinese news text classification. Therefore, we can show that the word vector trained by word2vec can well and accurately describe the characteristics of a text, so that words are not isolated or unrelated, and the influence of context on words is fully considered. The W2V-ITI-Combined-CNN model further calculates the weight of words on the basis of word vector training, and integrates word weight and word vector to construct a new text vector representation. Through the comparative experiment, the three models without word vector weight, calculating word vector weight based on traditional TF-IDF and calculating word vector weight based on improved TF-IDF in this paper are compared. It is found that after introducing word vector weight, the performance is improved, because the introduction of weight enhances the feature of word vectors. The performance of the improved TF-IDF algorithm is improved compared with the traditional TF-IDF algorithm, which shows that the improved TF-IDF algorithm is effective. Compared with Text-CNN with one-layer convolution and Text-CNN with multi-layer convolution, the combined-CNN model also has a certain improvement in classification effect.

6 Conclusion

Focusing on the classification of news text, this paper trains and tests Chinese news text with the help of convolution neural network model and word2vec word vector training model, so as to realize the classification of news text.

Starting from reality, this paper expounds the diversity and complexity of texts in today's information age. As a form of expression of Internet texts, news texts often carry a lot of information. In public security work, the text classification of public opinion news plays an important role in public security organs to grasp the dynamics and trend of public opinion in time, Efficiently and accurately classify the existing news texts, so as to further analyze and deal with some news purposefully, which plays an important role for the public security organ to control and grasp the public opinion situation in time, and can further prevent the occurrence of major social accidents.

Deep learning is the most popular text classification processing method. The cyclic neural network model (RNN model) has been successfully applied to text classification tasks and achieved good results. With the development of technology, CNN model has been widely used in the field of natural language processing in recent years. This paper introduces Chinese word segmentation technology into practice. Firstly, the news text information in the subset of thucnews is preprocessed. After data processing and cleaning, according to the basic steps of text word segmentation task, the text data information in the text set is extracted. The word vector is trained by word2vec model, the word weight is calculated by improved TF-IDF algorithm, and the word vector and word weight are fused to construct text vector representation. The combined-CNN model is used to train and test text data.

In this paper, precision, recall, F1 score and accuracy are introduced as evaluation indexes, and the classification effect is intuitively displayed and comprehensively analyzed by using confusion matrix. At the same time, in order to fully prove the effect and superiority of this model, two groups of comparative experiments are designed. Comparing the traditional Text-CNN model with the traditional Text-RNN model (including LSTM model and GRU model), and comparing the traditional Text-CNN model with the CNN model of word2vec training word vector and the combined-convolution neural network model based on word2vec and improved TF-IDF, it is concluded that the CNN model has better performance and less time-consuming than the RNN model in the text classification task of the text dataset in this paper. In this paper, the improved TF-IDF algorithm is combined with word vector to construct text vector representation, which further enhances the representation ability of word vector and retains the original text information to the greatest extent. Finally, the powerful learning ability of combined-CNN is used to deeply learn a large number of text vectors. Experiments show that the classification effect of this model is further improved compared with word2vec-CNN. However, during the experiment, it is also found that the W2V-ITI-Combined-CNN has long training time and high training cost. At the same time, the data set distribution in this paper is relatively balanced, and the classification result is relatively high. However, the real news data cannot be balanced, so there are some problems here, such as ideal news text data set and insufficient generalization. In the next step, we should consider more efficient text classification algorithm, further simplify the model and reduce the training cost. At the same time, we should also train the model on multiple data sets to reduce the dependence of the model on the data set through further optimization.

Acknowledgement: We would like to thank teachers Yuan Deyu and Sun Haichun for their helpful discussion and feedback, as well as the team's technical support.

Funding Statement: This work was supported by Ministry of public security technology research program [Grant No.2020JSYJC22ok], Fundamental Research Funds for the Central Universities (No.2021JKF215), Open Research Fund of the Public Security Behavioral Science Laboratory, People's Public Security University of China (2020SYS03), and Police and people build/share a smart community(PJ13-201912-0525).

Conflicts of Interest: The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

[1] J. Q. Zhao, "Research on opinion monitoring method based on automatic text classification," *Software Guide*, vol. 15, no. 3, pp. 133–135, 2016.

[2] S. P. Li, "An empirical analysis of news topic classification for network public opinion monitoring," *Journal of News Research*, vol. 7, no. 23, pp. 26–27, 2016.

[3] K. Zheng, X. M. Shu, H. Y. Yuan and S. K. Jin, "On classification method of network public opinion triggered by incidents," *Computer Applications and Software*, vol. 27, no. 5, pp. 3–5,+37, 2010.

[4] C. Huang and J. H. Chen, "Chinese text classification based on improved k-nearest neighbor algorithm," *Journal of Shanghai Normal University (Natural Sciences)*, vol. 48, no. 1, pp. 96–101, 2019.

[5] T. Y. Jiang, S. Wang and W. Xu, “Chinese text classification based on naive Bayes,” *Computer Knowledge and Technology*, vol. 15, no. 23, pp. 253–254,+263, 2019.

[6] Y. He, “Research on opinion monitoring method based on automatic text classification,” *Henan Science and Technology*, no. 29, pp. 8–10, 2019.

[7] S. Minaee, N. Kalchbrenner and E. Cambria, “Deep learning based text classification: A comprehensive review,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[8] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing(EMNLP)*, Doha, Qatar, pp. 1746–1751, 2014.

[9] R. Dubey and J. Agrawal, “An improved genetic algorithm for automated convolutional neural network design,” *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 747–763, 2022.

[10] T. T. Leonid and R. Jayaparvathy, “Classification of elephant sounds using parallel convolutional neural network,” *Intelligent Automation & Soft Computing*, vol. 32, no. 3, pp. 1415–1426, 2022.

[11] J. Xu and W. Chen, “Convolutional neural network-based identity recognition using ECG at different water temperatures during bathing,” *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1807–1819, 2022.

[12] S. Habib and N. F. Khan, “An optimized approach to vehicle-type classification using a convolutional neural network,” *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3321–3335, 2021.

[13] S. Mustajar, H. Ge, S. A. Haider, M. Irshad, S. M. Noman *et al.,* “A quantum spatial graph convolutional network for text classification,” *Computer Systems Science and Engineering*, vol. 36, no. 2, pp. 369–382, 2021.

[14] Y. Zhang and B. Wallace, “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification,” in *Proc. of the 8th Int. Joint Conf. on Natural Language Processing*, Taipei, Taiwan, pp. 253–263, 2017.

[15] R. Johnson and T. Zhang, “Deep pyramid convolutional neural networks for text categorization,” in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 562–570, 2017.

[16] Y. Zhang, K. F. Liu, Q. X. Zhang, Y. G. Wang and K. L. Gao, “A Combined-convolutional neural network for Chinese news text classification,” *Acta Electronica Sinica*, vol. 49, no. 6, pp. 1059–1067, 2021.

[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. Chung, C. Gulcehre, K. H. Cho, Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv:1412.3555, 2014.

[19] M. M. Sun, “Chinese text classification based on GRU-attention,” *Modern Information Technology*, vol. 3, no. 3, pp. 10–12, 2019.

[20] W. F. Lan, W. Xu, D. Z. Wang and P. C. Pan, “Text classification of Chinese news based on LSTM-attention,” *Journal of South-Central University for Nationalities(Nat. Sci. Edition)*, vol. 37, no. 3, pp. 129–133, 2018.

[21] T. Mikolov, K. Chen and G. Corrado, “Efficient estimation of word representation in vector space,” arXiv:1301.3781, 2013.

[22] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proc. of the 31st Int. Conf. on Machine Learning*, Beijing, China, pp. 1188–1196, 2014.

[23] G. S. Wang and X. J. Huang, “Convolution neural network text classification model based on word2vec and improved TF-IDF,” *Journal of Chinese Computer System*, vol. 40, no. 5, pp. 1120–1126, 2019.