# Research on the Algorithm of Text Data Classification Based on Artificial Intelligence

Ying-jian Kang and Lei Ma(✉)

Telecommunication Engineering Institute, Beijing Polytechnic, Beijing, China
kangyingjian343@163.com, malei235@tom.com

**Abstract.** In view of the low recall of the traditional network text data classification algorithm, an artificial intelligence based network text data classification algorithm is designed. Before feature extraction, text information is preprocessed first, and word stem is extracted from English. Because there is no inherent space between Chinese words, word segmentation is carried out to complete the preprocessing of network text data. On this basis, an evaluation function is constructed to evaluate each feature item in the input space independently, and to reduce the dimension of the features of the network text data. Finally, the artificial intelligence method is used to classify the network text data, and the most similar training text is found through similarity measurement in the network text data training set. The experimental results show that the designed algorithm based on artificial intelligence has higher recall than the traditional algorithm, and can meet the needs of network text data classification.

**Keywords:** Artificial intelligence · Network · Text · Data classification · Pretreatment · Recall rate

## 1 Introduction

In recent years, with the rapid development of computer technology, Internet and mobile Internet industry, the number of Internet users has shown an explosive growth. With the social platforms such as wechat and microblog, the Internet products are becoming more and more mature. With a large number of active users joining in, hundreds of millions of network text data are generated on the running platform every day, such as chat records, user comments, etc. Whether it is for government departments, scientific research institutions or Internet service providers, it is of great research significance and great application value to be able to correctly apply the network text data classification technology, so as to mine the real intention of users behind the data.

From the statistical point of view, although the traditional statistical text classification algorithms are powerful, they are always based on strong assumptions, but in most cases, these assumptions are not true in practical application. Therefore, although the results they get are accurate, they are difficult to be well connected with the actual application, that is to say, the high-precision results lead to the loss of a lot of text fixation. Some structural information leads to low recall rate. In order to solve the problem of low recall rate in the traditional network text data classification algorithm, a network text data classification algorithm based on artificial intelligence is designed. Artificial intelligence


Y.-D. Zhang et al. (Eds.): ICMTEL 2020, LNICST 327, pp. 38–52, 2020.
https://doi.org/10.1007/978-3-030-51103-6_4

is a new technology science which researches and develops the theory, method, technology and application system for simulating, extending and expanding human intelligence. Artificial intelligence includes a wide range of science, which is composed of different fields, such as machine learning, computer vision, etc. in general, one of the main objectives of artificial intelligence research is to enable machines to be competent for some complex work which usually needs human intelligence to complete.

The network text data classification algorithm designed in this paper completes the network text data classification through two aspects: network text data preprocessing and network text data feature processing. The experimental results show that the designed algorithm based on artificial intelligence has higher recall than the traditional algorithm, and has a certain practical significance.

## 2 Preprocessing of Network Text Data

Before feature extraction of documents, text information is preprocessed first, and the preprocessing process is shown in the following figure (Fig. 1):

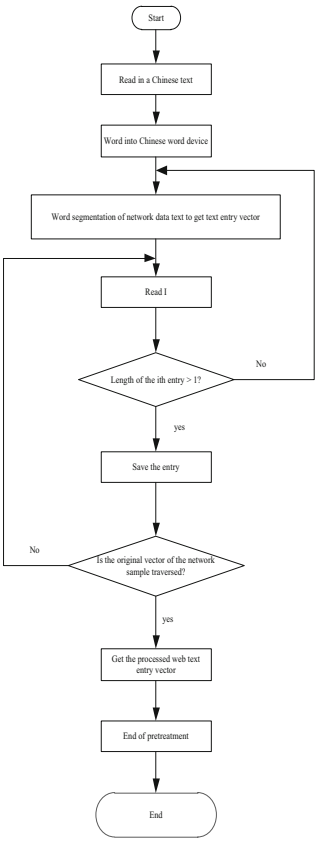


**Fig. 1.** Preprocessing process of network text data

In the process of extracting stemmed words from English, there is no inherent space between Chinese words, so word segmentation is carried out. The preprocessing of text information mainly includes word segmentation [1], elimination of symbol marks, removal of stop words and word frequency statistics.

Firstly, Chinese word segmentation is applied to the network text data, combined with the example of the network text classification system to improve the classification effect. The text classification system is shown in the following figure (Fig. 2):

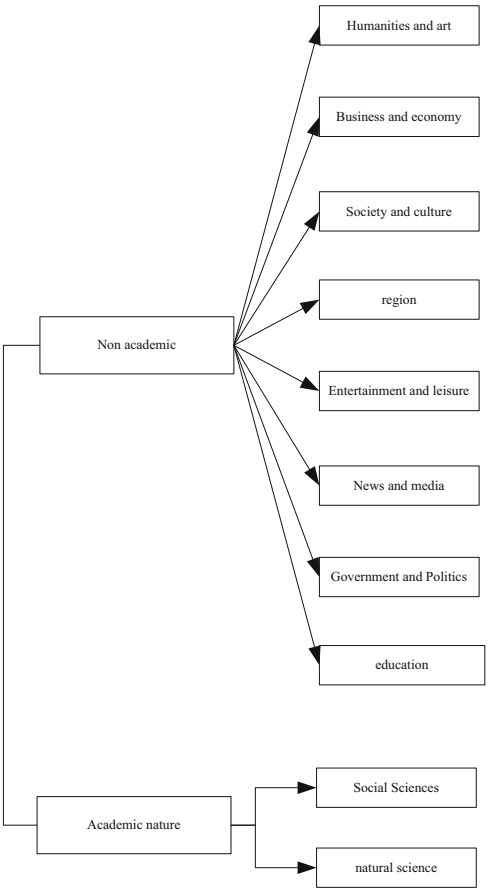


**Fig. 2.** Example of text classification system

Using the longest matching algorithm, from the first word of the text, if the dictionary cannot match the new word [2] composed of this word plus the next word, the word will be output. Otherwise, add another word on the basis of the new word to see if there is a matching entry in the dictionary. When there is no matching entry in the dictionary, you will get the correct segmentation entry, so as to repeat until the end of the text.

On this basis, the symbol mark is removed, and the method of establishing symbol dictionary is used to filter these characters through program flow. After the symbol dictionary is established, all words containing the above symbols in Chinese text are filtered out through program flow control, and the calculation formula is as follows:

$$G = \frac{K}{\sum_{x} sFH} \tag{1}$$

In formula (1), $G$ represents the network text data, $\sum_{x} s$ represents the statement containing symbols, $FH$ represents the symbol mark of the network text, and $K$ represents the network text eigenvector.

On the basis of the above elimination of symbol marks, the word frequency [3] statistics, because each word after the above-mentioned processing is basically the most representative of the text features of the attributes, the classifier must learn according to these attributes. Therefore, the more frequent the feature words appear in this kind of text, the more representative the feature of this kind, that is, the greater the weight value of its category, otherwise, the smaller the weight value, the pre statistics will be carried out according to the size of its weight value.

Based on the preprocessing of the network text data, the vector space model is established to mine the deeper information in the network text data. The mining process is as follows (Fig. 3):

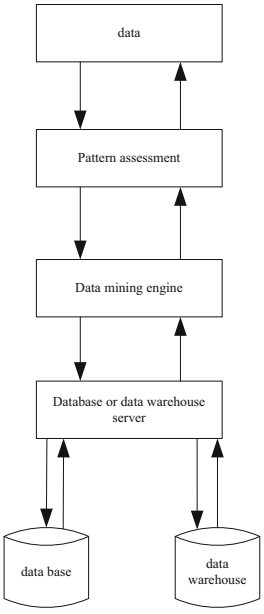


**Fig. 3.** Deep information mining of network text data

The information contained in the text is expressed by the frequency of the feature items and the order between them. With the directed pointer structure, the whole text becomes a complex graph, and the vector is used to represent the text. The specific way is as follows:

The content of the text is expressed by some characteristic items [4], which can be words, words, sentences and other language units. These items constitute a vector space, and each item represents a dimension. Use the following formula to express the frequency of the item:

$$w(t,j) = \frac{t'\bar{f} \times \log(n)}{\sum_{f} a(tg(y) \times \log(n_m + 0.007))} \tag{2}$$

In formula (2), $w(t,j)$ represents the weight of $t$ in text $j$, $\sum_{f} a$ represents the weight of each feature item, $tg(y)$ represents the importance of feature item $y$ in the text, $\log(n_m + 0.007)$ represents the total number of training texts, $t'\bar{f}$ represents the number of texts with $t'$ in training texts, and $\log(n)$ represents the normalization factor.

On this basis, a vector space model is established to discard the sequence information among each feature [5], and a text is represented as a vector, that is, a point in the feature space and a text set as a matrix, that is, a set of points in the feature space. According to the vector space model, the inner product of the corresponding vector or the cosine of the included angle is used to express the similarity between the two texts. For the measurement of the similarity of the network text data, the calculation formula is as follows:

$$We = \frac{p * x_0 K}{Y_i} \tag{3}$$

In formula (3), $We$ represents the similarity feature of network text, $Y$ represents the frequency of phrase in text, $p$ represents the network text vector, and $x_0 K$ represents the semantic unit of text content.

Through the above process, the preprocessing of network text data is completed, which provides the basis for network text data classification.

## 3 Network Text Data Classification

Because the dimension of the input space is very high, the feature of the network text data is reduced, an evaluation function is constructed, and each feature item in the input space is evaluated independently. The evaluation principle is as follows (Fig. 4):

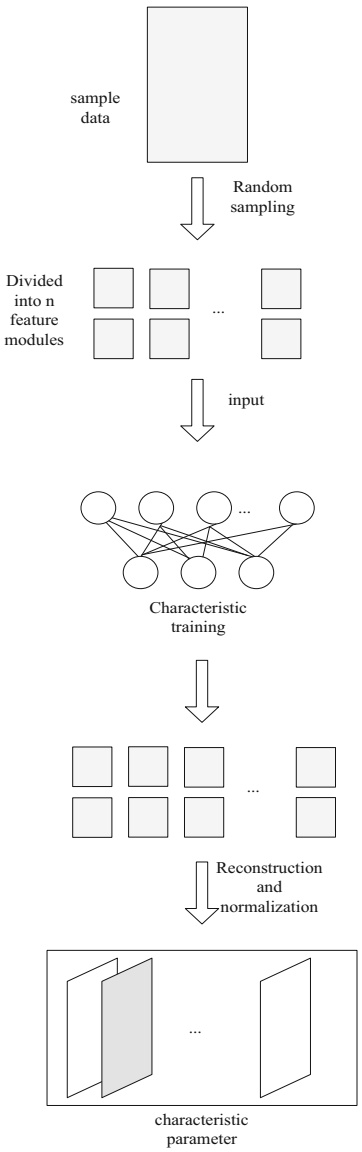


**Fig. 4.** Principle of characteristic item evaluation

Each feature item gets an evaluation score, then all feature items are sorted by size, and a predetermined number of the best feature items are selected as feature subsets. Select the method of information gain [6], and calculate it by the formula:

$$dx(f) = \sum_{i=1} k(p) + f \tag{4}$$

In formula (4), $dx(f)$ represents the probability of text data of category $f$ appearing in text set, $\sum_{i=1} k(p)$ represents the probability of text containing feature $p$ appearing in text set, and $f$ represents the conditional probability of feature.

According to the above formula, the information gain value of each feature item in the text set is calculated, from which the features lower than the preset threshold are removed, and the features higher than the threshold are retained as the optimal feature subset.

On this basis, the clustering process of feature items is as follows (Fig. 5):

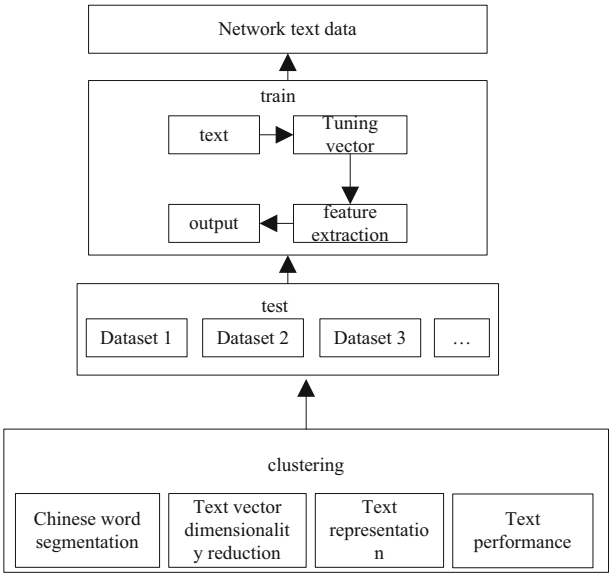


**Fig. 5.** Clustering process of characteristic terms

Analyze the relationship between feature items and feature items to group feature items, so that each group of feature items has some similar properties. Use a feature item that can represent the properties of the group to represent the group, so as to achieve the purpose of feature dimensionality reduction. Because there are many similar data in the network text data, by defining a similarity evaluation index to represent the distribution similarity between the feature items [7], the multiple feature items with similar distribution characteristics in the text are grouped into an independent event. The parameters of the event are determined and set by the weighted average of all the feature item parameters that constitute the event. This paper studies the simultaneity of the occurrence of feature terms, and then judges the possibility of their combination of feature terms. According to the similarity measure between feature

vectors, it judges the Category attribute of text, and reduces the dimension of the text matrix of feature terms, i.e. the input space. The singular value decomposition method is as follows:

$$W = D \underset{\leftrightarrow}{F} * \sum_{C} a \tag{5}$$

In formula (5), $W$ and $D \underset{\leftrightarrow}{F}$ respectively represent the left and right singular vector matrix corresponding to singular value, and $\sum_{C} a$ represents the ambiguity of semantic relationship between feature term and text.

Based on the above formula to determine the text category, classify the network text data, and the classification process is shown in the following figure (Fig. 6):

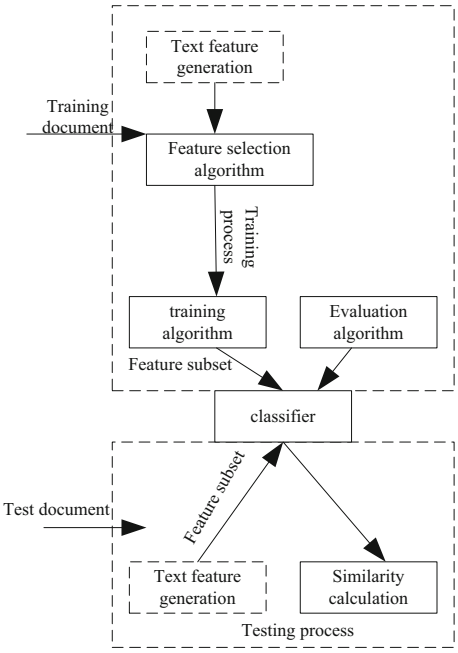


**Fig. 6.** Process of network text similarity

Rocchio algorithm is used to calculate the similarity between the network text data and the eigenvectors corresponding to all training texts in the text training set. The class center vector [8] is the weighted difference between all positive and negative eigenvectors. The calculation formula is as follows:

$$df = \exists \frac{z \sum_{c} q}{GH * x_v} \tag{6}$$

In formula (6), $df$ represents the weight of dimension $f$ of Text $d$ eigenvector, $\exists$ represents the total number of texts contained in this set, $GH * x_v$ represents the number of texts contained in category $x_v$, and $z\sum_{c} q$ represents the distance between text and various central vectors.

Finally, the network text data is classified by artificial intelligence method, and the classification principle is shown in the following figure (Fig. 7):

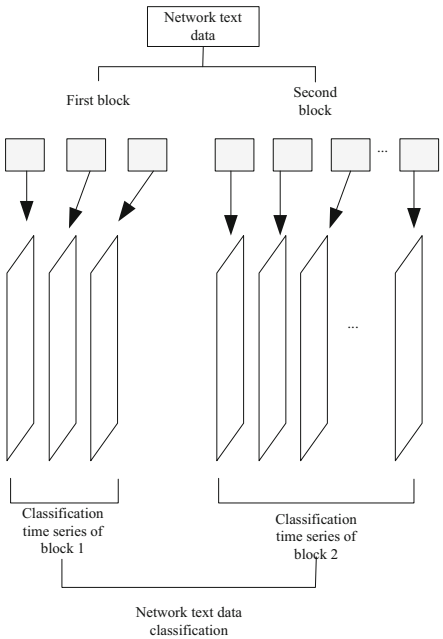


**Fig. 7.** Classification principle of network text data based on Artificial Intelligence

In the training set of network text data, the most similar training text is found by similarity measurement. On this basis, each text category is scored, and the score is taken as the sum of the similarity between the text belonging to the category in K training texts [9], and the similarity can use Euclidean distance or cosine similarity. The k-nearest neighbor text classification diagram is as follows (Fig. 8):

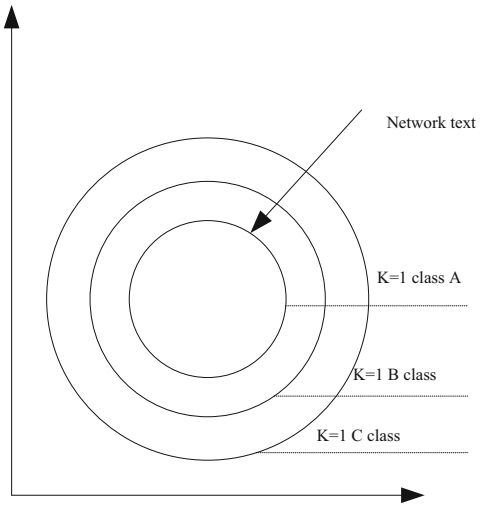


**Fig. 8.** **k** schematic diagram of nearest neighbor text classification

After the classification score of the adjacent text classification diagram is counted, it is sorted according to the score, and a threshold value is set in advance. Only the categories whose score exceeds the threshold value are considered. The test text belongs to all categories that exceed the threshold value, then:

$$dsf(d_i, c_i) = sim(f, n_d) \tag{7}$$

In formula (7), $dsf(d_i, c_i)$ represents the similarity measurement between Text $d_i$ and text $c_i$, and $sim(f, n_d)$ represents the score of category *sim*.

According to the above formula, find out an attribute of the corresponding text set in the network text data, test the text set, divide the text training set into several text subsets [10] according to different test results, each text subset constitutes a new node, repeat the above division process for the nodes, and then cycle until the specific termination conditions are reached, so as to complete the network. Classification of network text data. The specific process is shown in Fig. 9:

First of all, preprocess the network text data, establish the vector space model, mine the deeper information in the network text data, and reduce the dimension of the characteristics of the network text data, construct an evaluation function; then evaluate each characteristic item in the input space independently, cluster the characteristic item; finally, use the artificial intelligence method to classify the network text data, Complete the research of text data classification algorithm based on artificial intelligence.

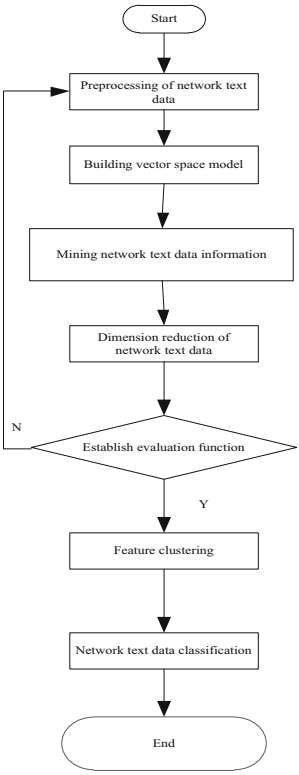


**Fig. 9.** Flowchart of network text data classification

# 4 Experimental Comparison

## 4.1 Experimental Environment

The test environment of this paper is a stand-alone environment, using two PCs. the operating system is Windows 7 64 bit professional edition and Linux CentOS 7 64 bit. The following table is the experimental configuration of the system (Table 1).

**Table 1.** Hardware and software configuration of Windows system

| System hardware configuration | | System software configuration | |
|---|---|---|---|
| CPU | Intel® 8 Nuclear i7-4790 | CPU | Intel® 8 Nuclear i7-4790 |
| Memory | 16G | Memory | 16G |
| Hard disk | 500G | Hard disk | 500G |
| Using the tool class library | Eclipse, JDK1.8, ND4J, DL4J | Using the tool class library | GCC, libsvm, word2vec |

The above is the configuration of the experimental environment, which is shown in the following figure (Fig. 10):

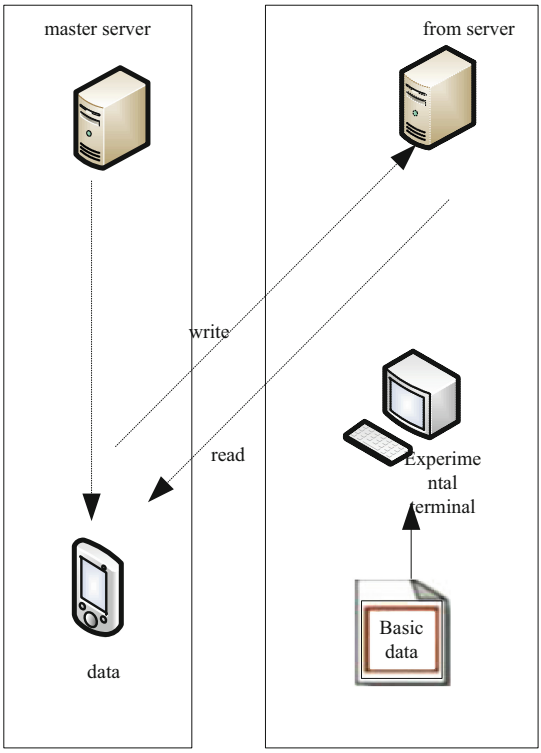


**Fig. 10.** Experimental environment

In addition, the statistical analysis of the experimental characteristics is carried out, and the statistical results are available for subsequent experiments.

### 4.2 Experimental Data Preparation

The experimental data of this paper come from three datasets: the corpus of Natural Language Processing Laboratory of Fudan University, the corpus of Chinese text classification of Tan Songbo, the Key Laboratory of network data science and technology of Computer Research Institute of Chinese Academy of Sciences, and the internet corpus of Sogou laboratory. For the secondary classification in the text classification experiment, the text establishment classification system is as follows (Table 2):

**Table 2.** Experimental text classification system

| Serial number | Class I | Class two |
|---|---|---|
| 1 | Sports | Football, basketball, table tennis, track and field |
| 2 | Entertainment | Games, movies, culture, fashion |
| 3 | Life | Food, tourism, education, health |
| 4 | Finance | Stocks, wealth management, funds, securities |
| 5 | Science and technology | Mobile phone, digital, exploration, automobile |
| 6 | Journalism | Military, social, domestic and international |

Due to the characteristics of Internet short text, the above experimental data does not fully meet the requirements of Internet text length. And the scale of the experimental data should not be too large, otherwise the program will crash and the training time will be too long. The actual data used in the text experiment is 8600 texts that meet the requirements selected from these data as the experimental data. The data mainly includes four categories, namely, sports, life, finance and entertainment. Each category includes two categories, namely, eight categories, namely, games, movies, football, basketball, financial management, securities, health and education. The distribution of the amount of text data for each category is shown in the following table (Table 3):

**Table 3.** Distribution of experimental data sets

| | Class two | Text quantity |
|---|---|---|
| Entertainment | Game | 1000 |
| | Film | 1000 |
| Sports | Football | 1000 |
| | Basketball | 1300 |
| Finance | Conduct financial transactions | 1000 |
| | Negotiable securities | 1000 |
| Life | Healthy | 1200 |
| | Education | 1100 |

Use the following formula to calculate the recall rate of the traditional algorithm and the designed network text data classification algorithm based on artificial intelligence. The recall rate is the proportion of correctly classified text in the text of the artificial classification result. The higher the value, the better the classification effect. The mathematical formula is as follows:

$$\text{Recall rate} = \frac{\text{Number of correct texts classified}}{\text{Expected number of texts}} \tag{8}$$

The experimental data is recorded in real time by the third-party software, and the corresponding experimental results are generated. The recall ratio of the two methods is calculated by the above recall ratio formula.

### 4.3 Analysis of Experimental Results

Recall rate is an important factor in network text data classification. Therefore, compare the recall rate of two network text data classification algorithms. The experimental results are shown in the table below (Table 4):

**Table 4.** Experimental comparison results

| Class I | Traditional algorithm (recall/%) | The design algorithm (recall rate/%) |
|---|---|---|
| Entertainment | 20 | 96 |
| Sports | 35 | 95 |
| Finance | 40 | 92 |
| Life | 22 | 93 |

Analysis of the above experimental comparison results shows that the recall rate of traditional methods in the classification of entertainment data, sports data, financial data and life data is lower than 50%, while the designed network text data classification algorithm based on artificial intelligence has a higher recall rate in the classification of entertainment data, sports data, financial data and life data. Compared with all the experimental results, the recall rate of the design method is 76% higher than that of the traditional method in the classification of entertainment data; 60% higher than that of the traditional method in the classification of sports data; 52% higher than that of the traditional method in the classification of financial data; and the recall rate of the design method in the classification of life data. The recall rate is 71% higher than that of traditional methods.

Therefore, through the above experiments, it can be proved that the designed algorithm based on artificial intelligence has a higher recall rate than the traditional algorithm, and has a certain practical significance.

## 5 Concluding Remarks

In view of the low recall of the traditional network text data classification algorithm, an artificial intelligence based network text data classification algorithm is designed. The network text data classification is completed from two aspects: network text data preprocessing and network text data feature processing. Experimental results show that the designed algorithm based on artificial intelligence has higher recall than the traditional algorithm.

There are still some shortcomings in the design method. It is necessary to further combine the natural language understanding technology to better obtain the semantic knowledge of text and text class, as well as the semantic relevance value between

words and phrases, such as the words in the beginning, the end, the introduction and the conclusion of the paragraph to give a higher degree of subordination. At the same time, we should study all kinds of classification technology, further strengthen the relevant theoretical and application research, strengthen the research of automatic classification and utilization of online information, so that the available information can play a maximum role.

## References

1. Zhang, Z., Ji, J.: Classification method of fMRI data based on convolutional neural network. Pattern Recogn. Artif. Intell. **30**(6), 549–558 (2017)
2. Wang, H., Hu, X., Li, P.: Semi-supervised short text stream classification based on vector representation and label propagation. Pattern Recogn. Artif. Intell. **31**(7), 634–642 (2018)
3. Fang, Fang, Wang, Y., Wang, S.: Knowledge acquisition from Chinese records of cyber attacks based on a framework of semantic taxonomy and description. J. Chin. Inf. Process. **33**(4), 48–59 (2019)
4. Fu, P., Lin, Z., Yuan, F., et al.: Convolutional neural network and user information based model for microblog topic tracking. Pattern Recogn. Artif. Intell. **30**(1), 77–84 (2017)
5. Chang, Shen, Junzhong, Ji: Text sentiment classification algorithm based on double channel convolutional neural network. Pattern Recogn. Artif. Intell. **31**(2), 158–166 (2018)
6. Li, Y., Xie, M., Yi, Y.: Fine-grained sentiment analysis for social network platform based on deep-learning model. Appl. Res. Comput. **34**(3), 743–747 (2017)
7. Du, H., Yu, X., Liu, Y.: CNN with part-of-speech and attention mechanism for targeted sentiment classification. Pattern Recogn. Artif. Intell. **31**(12), 1120–1126 (2018)
8. Chen, Q., Zheng, S., Chen, H.: Research on automatic detection of bad vocabulary in online media based on AlphaGo algorithm]. Comput. Dig. Eng. **46**(8), 1589–1592 (2018)
9. Huang, B., Liu, Q., He, Q., et al.: Towards automatic smart-contract codes classification by means of word embedding model and transaction information. Acta Automatica Sinica **43** (9), 1532–1543 (2017)
10. Liu, S., Cheng, X., Fu, W., et al.: Numeric characteristics of generalized M-set with its asymptote. Appl. Math. Comput. **243**, 767–774 (2014)