Text Classification of Potential Dangers in Coal Mine Safety Based on Convolutional Neural Network

Na Liu¹, Yanzhu Hu¹, Xinbo Ai¹, Yanchao Shao²,*

¹School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China
²Central Research Institute of Building and Construction Co., Ltd, MCC GROUP, Beijing 100088, China

*Corresponding author e-mail: syc-mail@163.com

Abstract. In recent years, with the improvement of people's safety awareness and the steady progress of safety production supervision, text classification algorithm based on data mining has been widely applied. At present, for the classification of hidden danger text in coal mine, it mainly relies on manual or machine learning. The efficiency of manual classification is too inefficient to meet the requirements of massive text classification. And the accuracy of machine learning-based classification method is low. In view of the above problems, this paper combines Word2vec and convolutional neural network to achieve accurate classification of hidden danger text in coal mine safety, and achieves great results. The results show that Word2vec can retain the semantic information between contexts. Convolutional neural network can effectively extract the high-level features of local contexts, and the classification effect is more accurate. This method can be implemented in the classification of hidden danger text in coal mine, which has very important practical significance.

1. Introduction

With the steady progress of safety production supervision, the situation of safety production in China has gradually improved. However, the situation is still serious, especially in coal mine safety production [1]. In recent years, the amount of hidden danger data in coal mine safety has increased sharply. These data are large in quantity and type, high in value. Mining important information and knowledge from a large number of potential safety hazards data is the key problem to prevent and control the potential dangers in coal mines safety. Text mining is an extremely important means to carry out the analysis of coal mine safety hidden danger information. How to use text mining algorithm to effectively analyze coal mine safety hidden danger information is of great significance [2].

In recent years, text classification methods based on traditional machine learning are widely applied. Huang Z S uses chi-square test to select text features according to word frequency [3]. Kaur R et al. proposes a probability-based text categorization method to achieve accurate text classification [4]. Yao Q, Song Z combine LDA and KNN algorithm to realize automatic text classification, and achieves good classification performance [5]. Jia SA proposed a new Chinese text categorization algorithm based on deep learning structure. Its performance is better than support vector machine [6]. Liu C, Wang W et.al put forward a new classification model, gravitational model (GM), which solves the classification
problem of species imbalance [7]. Cao J, Wang H, et.al proposes an incremental model-based support vector machine text classification algorithm, which is more effective [8]. Abdur R, Kashif J proposed a new feature ranking metric called max-min ratio (MMR), its performing text classification with higher accuracy and more efficiency [9]. However, the above algorithms ignore the semantic information between contexts and cannot automatically extract text features. Moreover, the calculation is complex, which leads to low classification accuracy.

Therefore, a text classification method based on convolutional neural network is proposed in this paper. Before constructing the classification model of convolutional neural network, Word2vec is used to characterize the similarity between semantics. Then, the high-level features of local context are extracted by convolutional neural network to achieve accurate text classification.

2. Theoretical knowledge

2.1. Convolutional Neural Network

Convolutional neural network model was proposed by Lecun in 1989[10]. It can be used as a classifier to classify the vectored text and output the corresponding classification results. In this paper, a four-layer convolutional neural network is constructed. The first layer is the input layer. The input layer is a matrix corresponding to the text to be classified. Assuming the input text is a sentence length of , is the word vector dimension, the text matrix of the sentence is:

\[ I = x_1 \oplus x_2 \oplus \cdots \oplus x_m \]  

(1)

Where, \( \oplus \) is the connection operation.

Each row of the matrix represents the vector corresponding to each word in the sentence. The number of rows is the number of words in the sentence, and is the dimension of the vector. Taking the text of “methane sensor without calibration in 1407 transport lane” as an example, the sentences are divided into “methane/sensor/without/calibration/in/1407/transport lane” according to the unit of words. And each word is vectored into word vectors with the same dimension to form a matrix.

The second layer is convolution layer. The convolution matrix window with the column number and row number is used to convolute with each matrix block of the input layer matrix from top to bottom. The column number of matrix block is , row number is . And the convolution result is obtained, that is:

\[ r_i = W \cdot I_{i+h-1} \]  

(2)

Where, \( i = 1, 2, \ldots, m-h+1 \), it represents the ith column matrix block from top to bottom. “\( \cdot \)” means point multiplication. Therefore, \( m-h+1 \) convolutions are carried out in total. Each convolution is followed by a non-linearization process, and the results of the non-linearization are obtained, namely:

\[ c_i = f(r_i + b) \]  

(3)

Where, \( b \) is the bias term, and its value can be adjusted automatically during the training process. \( f(\ ) \) is nonlinear activation function.

Finally, \( m-h+1 \) real numbers are obtained, and these real numbers are arranged in order to form the vector of convolution layer. The third layer is the pooling layer. Although the dimension of the convolution is much lower than that of the original input data, the dimension is relatively high. It is
necessary to use the pooling layer to reduce the dimension. There are two main options for pooling layer: maximum pooling and mean pooling. In this paper, maximum pooling is adopted. That is, the largest element of convolution vector obtained by convolution of each convolution window is taken as eigenvalue. The result of pooling is as follows:

\[ p_j = \max \{ c \} \] (4)

Thereby, the feature values \( p_j \) corresponding to the respective convolution windows \((j = 1, 2, \ldots, w, w)\) are extracted. All the feature values \( p_j \) are sequentially spliced to form a vector \( p \) of the pooling layer, \( p \) is the vector representing the global feature of the sentence.

The fourth layer is the output layer. The output layer is fully connected with the pooling layer. The pooling layer vector \( p \) is used as input. Softmax classifier is used to classify the vectors. The classification results are converted to the probabilistic values between 0 and 1. The class with the greatest probability is selected as the final classification result.

2.2 Evaluating indicators

The final classification results should be evaluated to verify the effectiveness of the model. When using convolutional neural network to classify the text, the most commonly used indicators to describe the quality of classification results are precision rate, recall rate, \( F_1 \) value, confusion matrix, ROC curve and so on. The specific definition is given as following:

\[ P = \frac{TP}{TP + FP} \] (5)

\[ R = \frac{TP}{TP + FN} \] (6)

\[ F_1 = \frac{2TP}{2TP + FP + FN} \] (7)

Where, \( TP \) is the positive samples predicted as positive samples by the model, \( FN \) is the positive samples predicted as negative samples by the model. And \( FP \) is the negative samples predicted as positive samples by the model.

When using data sets with unbalanced classifications, a single indicator such as accuracy does not reflect the overall situation. For example, in the data set of this paper, the number of text in equipment and facility category is 562, and the number of text in other three categories is 1438. Obviously, the difference between the number of positive and negative class labels is large, and the data set is not balanced. Therefore, the precision rate \( P \), recall rate \( R \) and \( F_1 \) value are used to evaluate comprehensively the classification performance in this paper. The precision reflects the accuracy of the classification results. The higher the precision, the more accurate the classification results are. The recall rate reflects the comprehensive degree of classification results. The higher the recall rate is, the more comprehensive the classification results are.

3. Experiment

This section mainly includes three parts: text pre-processing, word vector construction and text classification. Firstly, word segmentation, removing symbol and stop words are carried out on security hidden danger text. Secondly, word vectors are constructed with Word2vec. Finally, convolution neural
network is used to realize text classification. The principle of text classification of hidden dangers in coal mine safety is shown in Figure 1.

![Figure 1. Principle of text categorization for hidden dangers in coal mine safety](image)

The data of this paper is 2000 coal mine safety hidden danger text records of a coal mine in Sichuan Province. Each text length is 5-20, 70% of the samples are used as training set to train model, and the remaining 30% are used as test set to test model. According to the Classification Standards for Hidden Dangers of Coal Mine Safety Production, the texts of hidden dangers in coal mine safety are divided into four categories: safety management, equipment and facilities, employees and workplace. The proportion of four types of texts in 2000 text records of hidden dangers in coal mine safety is shown in Figure 2. The characters S, F, E, and W represent “safety management”, “equipment and facilities”, “employees” and “workplaces”.

3.1. Text Preprocessing and Word Vector Construction

In Before the text pre-processing, it's necessary to build a special corpus for the field of coal mine safety to improve the accuracy of word segmentation. The construction of the corpus is completed manually. Part of the corpus is shown in Figure 3. The character behind / is the label of the word, n means noun, v means verb, and so on.

In this paper, text pre-processing mainly includes word segmentation, symbol and stop words removal. The removed symbols include commas, periods, ellipses, etc. The removal of stop words is mainly based on the stop word list to remove some words in the text that have no practical effect, such as "ground", "ah", "such as", and so on. For example, after the segmentation and removal of stop words, the sentence “Air leakage and water leakage occur in pressure pipes, and cable of mechanical and electrical equipment suspended disorderly” becomes "pressure pipe/air leakage/water leakage/electromechanical/equipment/cable/suspension/disorder". The results of text pre-processing are shown in Table 1.
Table 1. Word segmentation of partial security hidden danger texts

| Word Segmentation Results | types          |
|---------------------------|---------------|
| power distribution box/one gate/ multiple uses | safety management |
| 1407/Transportation lane/methane/sensor/no/Calibration | equipment and facilities |
| stuff/didn’t/ Wear/Safety Hat | employees      |
| 1403/ ventilator /running/unmoral | equipment and facilities |
| throttle /with/ wind tube /sundries   | workplaces     |

The pre-processed text is saved and trained with Word2vec module. During the training process, words are used as text representation to form word vectors. And word vectors are spliced according to the order in which words appear in sentences to form a matrix representing sentences. In the process of word representation, the semantic correlation between words can be learned. The order in which words appear in sentences is also taken into account in the word vector stitching. Word2vec module includes CBOW model and Skip-gram model. In this paper, Skip-gram model is selected for training. Since the convolutional neural network belongs to the deep learning model, it can automatically learn and extract text features. Therefore, when the dimension of word vector changes within the appropriate range, the effect of classification is not obvious. In this paper, the dimension of word vector is set to 200, and the window is set to 5. The results are stored as binary files. Part of the saved files are shown in Figure 4.

3.2. Experiment results

When constructing the convolution neural network model, the following parameters needed to be set: the number of convolution layers, the number of convolution kernels, step length of convolution layers, the number of pooling layers, the number of pooling kernels, step length of pooling layers, and the activation function. In this paper, the parameters of convolution neural network are set as follows: the number of convolution layers is 2, the number of convolution kernels is 32 and 64, the step length of convolution layer is 1. The number of pooling layers is 2, the number of pooling kernels is 32 and 32, the step size of pooling layer is 2, and the activation function is ReLu function.

After the construction of the classification model based on convolutional neural network, the text of coal mine safety hidden danger is classified. The results of four kinds of text misclassification are shown in Table 2 and Table 3. The numbers 1-4 represent “safety management”, “equipment and facilities”, “employees”and “workplaces”.

Table 2. The results of misclassification of text in training set

| types          | 1   | 2   | 3   | 4   |
|----------------|-----|-----|-----|-----|
| 1              | 316 | 5   | 11  | 21  |
| 2              | 2   | 358 | 3   | 5   |
| 3              | 10  | 9   | 276 | 10  |
| 4              | 7   | 21  | 6   | 342 |
Table 3. The results of misclassification of text in test set

| types | 1   | 2   | 3   | 4   |
|-------|-----|-----|-----|-----|
| 1     | 128 | 2   | 4   | 9   |
| 2     | 4   | 157 | 2   | 1   |
| 3     | 9   | 3   | 118 | 6   |
| 4     | 2   | 7   | 2   | 145 |

It can be seen from Table 2 and Table 3 that the convolutional neural network in this paper can classify almost all texts correctly in "safety management", "equipment and facilities", "practitioners" and "workplaces". Only a few samples are misclassified into other classes it can achieve better classification results.

The classification evaluation indexes of the four types of security risk texts in the training set and the test set are shown in Figure 5 and Figure 6. From left to right, the order is the accuracy rate, the recall rate, and the \( F_1 \) value. The numbers 1-4 represent “safety management”, “equipment and facilities”, “employees” and “workplaces”.

![Figure 5. The evaluation index of four types of texts in training set](image1)

![Figure 6. The evaluation index of four types of texts in test set](image2)

The Figure 5 and Figure 6 show that the text classification model of hidden dangers in coal mine safety in this paper can achieve better classification results in both training set and test set. It can achieve a higher recall rate while maintaining a higher classification accuracy.

In order to compare the classification performance between the convolutional neural network model and traditional machine learning model, Naive Bayesian and Logistic regression model are selected for comparison. The results is shown in Table 4.

Table 4. The performance of different classification models

| model                | precision | Recall rate | value  |
|----------------------|-----------|-------------|--------|
| Naive Bayesian model | 88.2%     | 80.5%       | 84.2%  |
| Logistic regression model | 85.9% | 79.2%       | 82.5%  |
| Model in this paper  | 93.5%     | 87.6%       | 90.44% |

The Table 4 shows that the performance of the coal mine safety risk text classification model is better than the traditional machine learning classification model. The accuracy rate and recall rate are significantly higher than the naive Bayesian and Logistic regression models. It can be seen that this method can classify the hidden danger text of coal mine safety very well.

4. Conclusion

This paper establishes a classification model of hidden danger texts of coal mine based on convolutional neural network. The context semantic connection of hidden danger text is studied and vectored. Then,
the convolutional neural network is established to extract the high-level features of the context. The model realizes the automatic classification of hidden danger texts. The conclusions is following:

1) Introducing a deep learning-based convolutional neural network model in the classification of mine safety hazard texts achieves better performance. Comparing with the traditional machine learning classification model, the method of this paper has more high accuracy and comprehensiveness. It provides an effective method for text classification of hidden danger texts of coal mine.

2) Convolutional neural network classification model adopts convolution windows of various sizes, which can better adapt to the change of hidden text length. It effectively improves the effect of feature extraction, and thus improves the accuracy of classification.

Based on the convolution neural network model, this paper preliminarily realizes the accurate classification of hidden dangerous text, and draws some meaningful conclusions. However, in this paper, the construction of convolution neural network is to realize the automatic classification of short text. The classification effect of long text may not be very good. In addition, this paper mainly achieves automatic classification at the level of "safety management", "equipment and facilities", "employees" and "workplaces". It does not involve multi-level classification, such as secondary classification, third-level classification and so on.

Acknowledgments
This work was financially supported by Beijing Natural Science Foundation (No. 4192042) and the National Key Technology R&D Program of China (No. 2016YFC0701309).

References
[1] Fu G, Cao J, et.al. Relationship Analysis of Causal Factors in Coal and Gas Outburst Accidents Based on the 24Model [J]. Energy Procedia, 2017, 107: 314 - 320.
[2] Zhang Y, Chen M, et.al. A review on text mining [C]. International Conference on Software Engineering & Service Science. 2015.
[3] Huang Z S, Ye Z L. Application of improved CHI statistical method in text classification [J]. Application of Computer System. 2016, 25 (11): 136 - 140.
[4] Kaur R, Kumar M. Domain Ontology Graph Approach Using Markov Clustering Algorithm for Text Classification [J]. 2018.
[5] Yao Q, Song Z. Research on text categorization based on LDA [J]. Computer Engineering & Applications. 2011.
[6] Jia S, Qin S, et.al. Chinese text categorization based on deep belief networks [C] International Conference on Computer & Information Science. 2016.
[7] Liu C, Wang W, et al. A new Centroid-Based Classification model for text categorization [J]. Knowledge-Based Systems. 2017.
[8] Cao, J, Wang, H. An improved incremental learning algorithm for text categorization using support vector machine [J]. Journal of Chemical and Pharmaceutical Research. 2017.
[9] Abdur R, Kashif J, et al. Selection of the most relevant terms based on a max-min ratio metric for text classification [J]. Expert Systems with Applications, 2018, 114: 78 - 96.
[10] Zhang T, Li C, et al. Text Feature Extraction and Classification Based on Convolutional Neural Network [J]. 2017.