Application of Improved Convolutional Neural Network in Text Classification

Liu Ronghui, and Wei Xinhong

***Abstract*—Text classification is a basic problem in the field of natural language processing. The model based on deep learning is a common model to deal with this problem. However, there are two problems in the text classification application using deep learning models: one is that the important role of the text hierarchical structure is not fully considered in classification judgment; the other is that the traditional word segmentation may produce ambiguity when dealing with text. Aiming at these problems, a text classification algorithm based on deep convolution neural network and attention mechanism is proposed. The neural network language model word2vec is used to train the word vector of short text. The trained word vector is used to represent the text. The convolution neural network is used to extract the abstract features of short text. Through extracting the abstract features, the classifier is used to classify short text. The results show that the algorithm has higher accuracy than the traditional convolutional neural network and bag of word model.**

***Index Terms*—convolutional neural network, attention mechanism, short text, machine learning, deep learning**

I. Introduction

TEXT classification automatically classifies and labels text according to a certain classification system or standard. It is mainly used in the fields of information retrieval [1], spam text filtering [2], public opinion detection [3], emotion analysis [4] and so on. It has been a research hotspot in the field of natural language processing [5-7]. Text classification methods mainly include word matching [8], knowledge engineering [9], statistics and machine learning [10]. As the first step of text classification, the natural language text with unstructured character should be transformed into numerical data. Word vector is a great tool to transform the character into numerical data, which is of great significance for the research of text classification.

In recent years, with the application of deep learning model in computer vision [11,12] and speech recognition [13,14], remarkable achievements have been made. It is worth studying to bring the advantages of deep learning into text classification, which may improve the accuracy of text classification [15]. Literature [16] proposed a sentence level classification model based on convolutional neural network (CNN). CNN is used for pre training to optimize feature expression when training word vectors. Literature [17] proposed an improved Bi-LSTM, which adds 2D convolution layer and 2D pooling layer to Bi-LSTM. The proposed Bi-LSTM is used to capture the dependency between sentences. 2D convolution layer and 2D pooling layer are used to extract deep feature expression. Compared with RNN, CNN and RECC-NN, the effect is significantly improved. Literature [18] proposed the VDCNN model by adopting the deep convolution network. However, CNN model has its defects, that is, it can only mine the local information of text. Compared with CNN, region convolutional neural networks (RNN) can better capture the long-distance dependence between text. In order to model the relationship between sentences, Literature [19] proposed hierarchical RNN model at the text level. Literature [20] proposed a deep map long and short-term memory (DM-LSTM) model for multi label text classification. In its proposed model, a graphics database is used to store documents. The standard dictionary is used to preprocess the document. Then the classification dictionary is generated. Literature [21] uses CNN to train text vectors. Some sequences obtained through CNN are taken as the output of word vectorization. Then these sequences are taken into LSTM to complete the text classification operation.

In addition, attention mechanism has been widely used in text classification models and other application [22]. The attention mechanism can distinguish the importance of each word in the text classification. Literature [23] proposed the attention mechanism on words and sentences, which retains the document structure with the hierarchical type. The importance of each sentence and word is distinguished to the classification category. At present, due to the deep learning method and attention mechanism have achieved good results in many fields, text classification based on the structure of convolutional neural network and attention is proposed. Then the experiments are carried out. The experimental results show that the proposed method is effective.

II. Convolutional neural network

Convolutional neural network is a kind of deep neural network developed in recent years, which makes a great breakthrough in computer vision and speech recognition. CNN mainly includes convolutional layer and pooling layer.

A. Convolution layer

Convolution layer is the core component in the CNN, which has the characteristics of local connection and weight sharing. The convolution layer is mainly used to produce one or more outputs. The characteristic maps of the previous layer are convoluted with the convolution kernels, where

Manuscript received January 21st, 2022; revised May 29th, 2022.

Liu Ronghui is an Associate Professor of the School of Computer and Data Science, Henan University of Urban Construction, Pingdingshan, 467036, China (corresponding author, e-mail: liurh_126@126.com).

Wei Xinhong is an Associate Professor of the School of Computer and Data Science, Henan University of Urban Construction, Pingdingshan, 467036, China (e-mail: 30080805@hncj.edu.cn).

convolution kernel is represented by v, $v \in \mathbb{R}^{lm}$, l represents the window height of the convolution kernel, and m represents the dimension size of the word vector. A new characteristic value is generated when passing through a word sequence window with height l and width m. $\boldsymbol{V}_{j:l}$ represents a word sequence of length l (V_j, V_{j+1}, ..., V_{j+l}), where V_j represents a word. The calculation process of each characteristic value is as follows.

$$c_j = f(\alpha \cdot V_{j:j+l} + d) \tag{1}$$

where α is the weight parameter of the convolution kernel, d is the bias term of the convolution layer, $d \in \mathbb{R}$. The operator $(\cdot)$ represents the convolution operation. $f(\,)$ is the activation function. Generally, nonlinear functions such as sigmoid and tanh can be used as activation function. A feature map can be obtained by convoluting the word sequence ($\boldsymbol{V}_{1:l}$, $\boldsymbol{V}_{2:l}$,..., $\boldsymbol{V}_{M-l+1:M}$) in each window of the short text. The specific calculation process is as follows.

$$c = (c_1, c_2, \cdots, c_{M-l+1}) \tag{2}$$

where M represents the number of words in a short text, l represents the height of the convolution kernel window, and c represents the characteristic map formed by the short text passing through a convolution kernel. The height of the convolution kernel is l. The width is the dimension of the word vector. As a result, the characteristic map formed after convolution is a matrix with height (M - l + 1) and width 1. Different convolution kernels can extract features from different angles. By setting the number of convolution kernels, multiple different feature maps can be obtained.

B. Pooling layer

The function of the pooling layer is to down sample the characteristic map output from the convolution layer, which can be used to simplify the information output from the convolution layer and reduce the network parameters. The pooling layer scans and samples in steps of the size of the pool area, rather than continuous sampling. Assuming that the width of the pool area is v and the height is l, the input characteristic map is divided into several sub area with $v \times l$ size. After each sub area is pooled, the value is obtained after the corresponding pooling operation. The max-pooling method is used here, which extracts the maximum feature value in the pool area. The specific calculation process is as follows:

$$c_{\max} = \max(c_j) \tag{3}$$

where, c_j represents the characteristic map formed after the convolution operation of a convolution kernel with original text, $0 < j \leq N$, where N is the number of characteristic maps. The 1-max pooling operation is adopted. The height of the pooling region is M - l + 1 and the width is 1. Therefore, a characteristic map will get a value after pooling.

The number N of the convolution kernels and the size of the pooling area are set in the convolution neural network. The feature value containing more semantic and location information can be extracted from the original text. Then all the extracted feature value is spliced together to form a vector. The vector is the feature representation of short text formed after convolution neural network processing. The structure of extracting feature value by convolution neural network is shown in Figure 1.

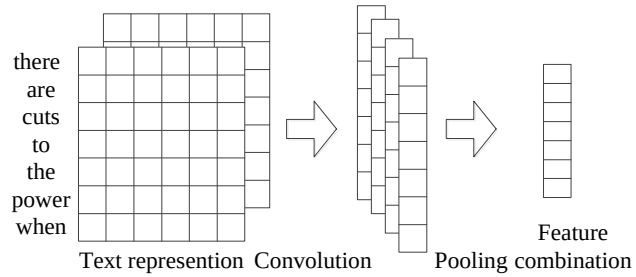


Fig. 1. Structure of CNN

III. Attention Model

Attention model is used to characterize the word correlation between text sentences and output results, and to represent the importance of each word in sentence s_j and its corresponding label t_j. The attention text generated by the attention model is expressed as b_j.

$$m_j = f(s_{ji}, t_j) \quad 1 \leq j \leq k, 1 \leq i \leq l \tag{4}$$

$$q_j = \frac{\exp(m_j)}{\sum_{j=1}^{k} \exp(m_j)} \quad 1 \leq j \leq k \tag{5}$$

$$b_j = \sum_{j=1}^{k} q_j \cdot s_j \tag{6}$$

where s_j represents a text sentence; t_j represents the label to the corresponding sentence; $f(\,)$ represents a forward network with a hidden layer; q_j and m_j represent the important information of each word in the text. After the features are obtained by convolution neural network and attention model, the feature c learned by the pooling layer relates to the attention text b_j, which is the input of the full connection layer. The output result after passing through the full connection layer is expressed as

$$o(s) = f(w \cdot (c \otimes b_j) + c) \tag{7}$$

where $o(s)$ represents the output value obtained after the model; $\otimes$ represents vector splicing operation; w represents the weight matrix of the full connection layer; c represents the offset term; $f(\,)$ indicates the selection of classifier.

IV. Text Classification

A. Text classification model

Given the text sentence data set W, which contains the text $S = \{s_1, s_2, ..., s_n\}$ and the classification tag $T = \{t_1, t_2, ..., t_n\}$ corresponding to each sentence, where each text sentence s_i is composed of m words, expressed as $\{s_{j1}, s_{j2}, ..., s_{jn}\}$. The final objective function is expressed as

$$p(T \mid S) = \arg\max_{\theta} f(T \mid S; \theta) \tag{8}$$

where θ represents all parameters involved in the model; $f(\,)$ represents the formal expression of the model.

B. Data to vector using word2vec

Word vectorization is the transformation of word representation into digital representation using computer processing. There are some outstanding word vectors using neural network, such as NNLM proposed by Bengio and log-linear model proposed by Hinton. In addition, the well-known word2vec model is a concise and efficient word vector model. Word2vec technology is a key breakthrough in

the application of deep learning technology in self-language processing.

(1) Word2vec model

Word2vec model is essentially a simple neural network. When the network training is completed, the weight matrix between the input layer and the hidden layer is the word vector mapping table. It is generally divided into CBOW and skip-gram models. Previous studies have shown that CBOW performs better in small-scale corpora, while skip-gram is more suitable for larger corpora.

(2) Data Vectorization

After extracting the text under classification, the sentence is transformed into a vector according to the word vector model. Wikipedia corpus and word2vec tool are used to train the model. Set the statement $S = \{s_1, s_2, ..., s_n\}$, where $s_j = (w_1, w_2, ..., w_d)$ is a word item and d represents the dimension of the word. The vectorization of statement S is expressed as $S = (w_1 \oplus w_2 \oplus ... \oplus w_d)$, where $\oplus$ is the join operator. Therefore, statement S is converted into a string of vectors in the order of word vectors. Similarly, for each text $W = (S_1 \oplus S_2 \oplus ... \oplus S_n)$, where n represents the important ranking of statement S, that is, S_j is the jth important statement in the text. After the text is transformed into vector form, the vectorized data can be used to train the neural network.

C. Text classification flow

In order to realize text classification, a network structure based on convolutional neural network and attention model is proposed. The model structure is shown in Figure 2. The model is mainly composed of two parts. The lower part is a typical convolution neural network structure. And the upper part is the structure of attention model. The overall process of the model is as follows: firstly, the input text sentence is encoded by word vector model, converted into word vector representation. The relevant features of the sentence are obtained after convolution neural network. Then, after combining the features obtained by the attention mechanism, the text classification is completed by using the classifier after full connection.

(1) Firstly, the model transforms characters or words into corresponding vector expressions through the operation of CNN layer. Suppose a text has n words, expressed as S_j, $j \in [1, n]$. From the formula $W = (S_1 \oplus S_2 \oplus ... \oplus S_n)$, the vector representation W of the sentence has been obtained. Then, after convolution operation, the eigenvector V is obtained

(2) The purpose of adding the attention mechanism is to find the words or words that contribute most to the meaning of the sentence. The importance of words is determined. Then, after softmax operation, a normalized attention weight matrix is obtained, which represents the weight of the jth word. Finally, after the attention weight matrix is obtained, the sentence vector is regarded as the weighted sum of these word or word vectors. After calculating the importance of each word and word to the sentence according to equations (4) - (6), the importance ranking is obtained.

(3) After the total text vector V is obtained, the text classification operation can be carried out through the softmax layer

$$p = \text{soft}\max(wV + b) \tag{9}$$

where w and b are weight matrix and offset matrix respectively. In addition, the goal for equation (9) is to minimize the loss function, i.e., negative log likelihood function, as shown in equation (10).

$$L = -\sum_{d} \log p_{dj} \tag{10}$$

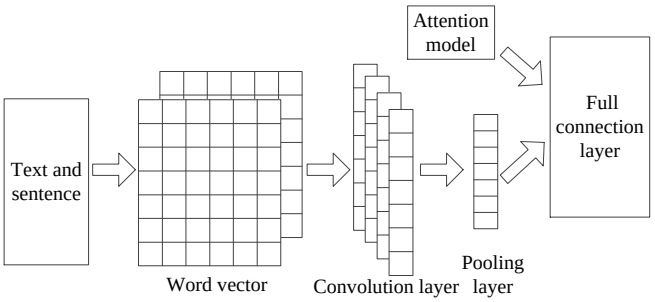


Fig. 2. Structure based on convolutional neural network and attention mechanism

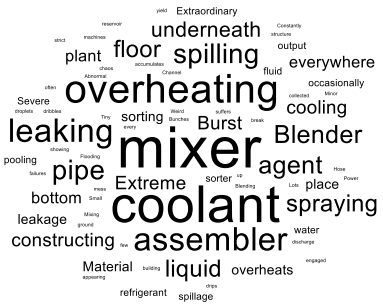
V. Experimental Analyses

A. Experimental dataset

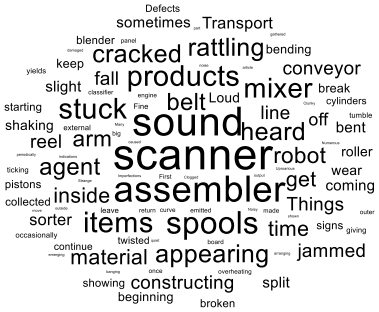
The text dataset used in this experiment comes from 480 fault records of factory equipment. Some records are shown in Table I. The text mainly consists of description part and failure type part. Failure type includes leak, mechanical failure, electronic failure, and software failure. The word cloud of description under different failure type is shown in Figure 3. 90% of the 480 text records are selected as training samples and the remaining 10% as test samples. The number of training and testing samples is shown in Table II.

TABLE I
SOME EXAMPLES OF TEXT SAMPLES

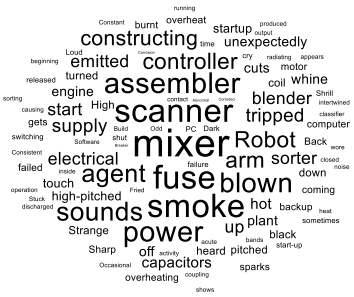
	Description	Failure type
1	Loud rattling and banging sounds are coming from assembler pistons.	Mechanical Failure
2	There are cuts to the power when starting the plant.	Electronic Failure
3	Mixing software has crashed.	Software Failure
4	Mixer tripped the fuses.	Electronic Failure
5	Burst pipe in the constructing agent is spraying coolant.	Leak
6	A fuse is blown in the mixer.	Electronic Failure



(a) Leak



(b) Mechanical Failure



(c) Electronic Failure

(d) Software Failure

Fig. 3. Word cloud of description under different failure type

TABLE II
THE NUMBER OF TRAINING AND TESTING DATA

	Training data num	Testing data num
Leak	60	8
mechanical failure	181	23
electronic failure	155	13
software failure	36	4
Total	432	48

B. Word to vector

The pretrained word2vec model is used for word embedding. The word embedding model returns 300-dimensional pretrained word embedding. The size of the text sequences is *E*-by-*W* matrix, where *E* is the embedding dimension, and *W* is the number of word vectors in the sequence. Taking the text sequences "Loud rattling and banging sounds are coming from assembler pistons" under "Mechanical Failure" for example, the vectors after word embedding are shown in Figure 4.

C. Result analysis

Experiments are designed to verify the effectiveness of the sentence level text classification scheme based on the improved convolutional neural network. In order to verify the performance of the proposed method, traditional convolutional neural network（CNN）and bag of word model (BWM) are used as the comparison method respectively, where the traditional CNN method is shown in Section 2, and the input of traditional CNN is also processed by word2vec model. The bag of word model uses word frequency counts as predictors. The confusion matrix chart is shown in Figure 5 under the BWM, CNN and the proposed method. The text classification accuracy of the three methods is shown in Table III and Figure 6. It can be seen from Table III and Figure 6 that the text classification algorithm based on the proposed convolutional neural network model has higher text recognition rate under leak type fault than BMW model under leak failure. Compared with CNN, the text classification algorithm based on improved convolutional neural network has higher text recognition rate under mechanical failure. Compared with BWM and CNN methods, the accuracy of the proposed method is higher on the whole fault type. The accuracy of different categories is more balanced. The higher accuracy in the original scheme is only slightly reduced in the new scheme, and the lower accuracy in the original scheme is greatly improved in the new scheme.

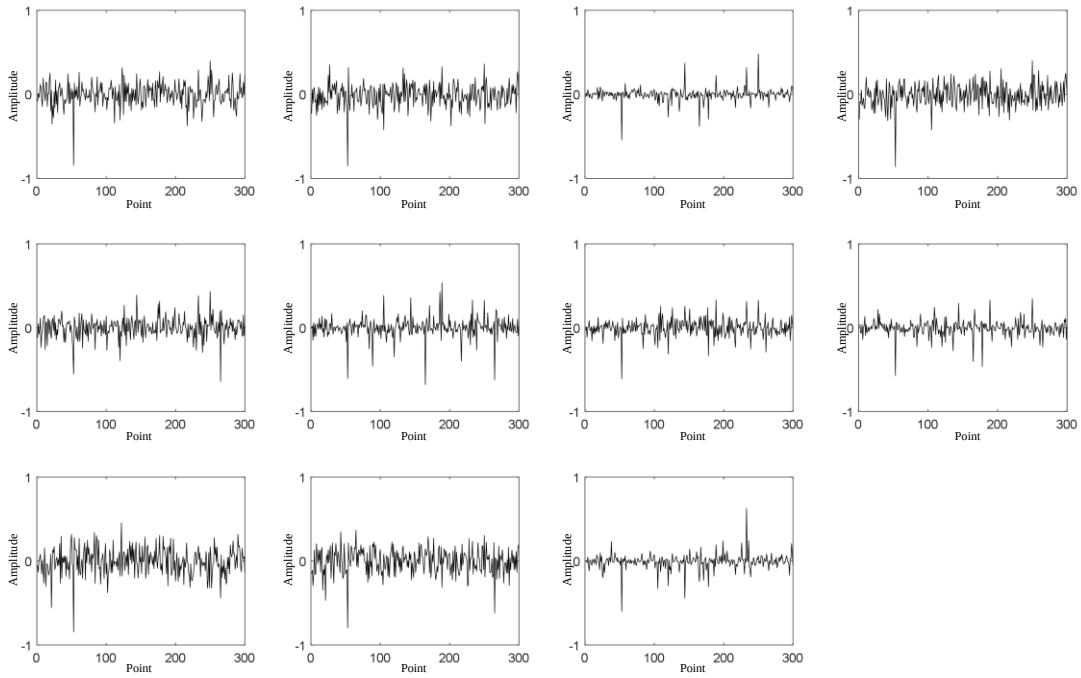
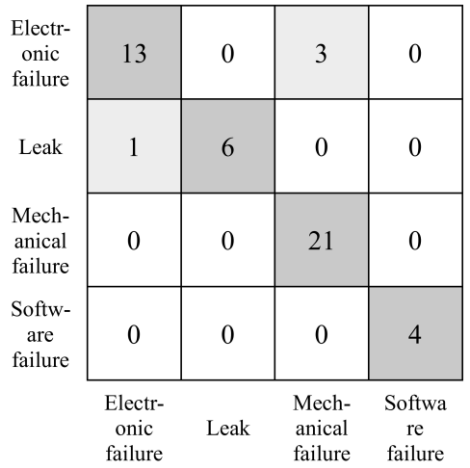
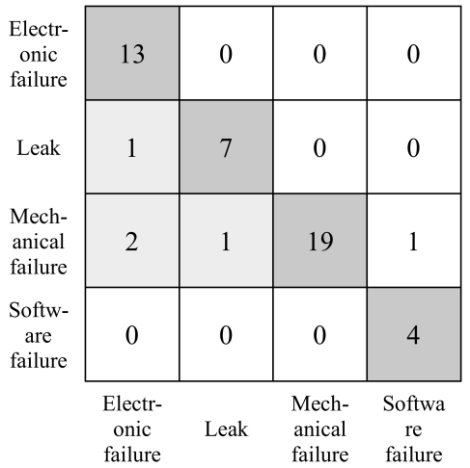


Fig. 4. The vectors after word embedding

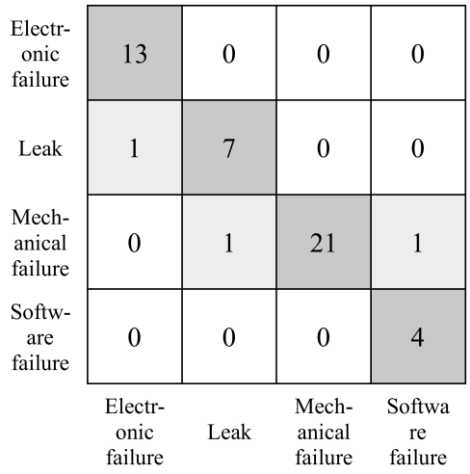
The reason is that, on the one hand, the convolution neural network model can extract richer classification features by adding convolution kernel. On the other hand, it can extract higher-level classification features by increasing the number of convolution layers. In short, convolutional neural network can extract richer features horizontally and more levels vertically, which is impossible by traditional machine learning models. Therefore, the distribution of different words is captured by improving the convolution neural network. The accuracy of convolution neural network in text classification can be further improved.



(a) BWM



(b) CNN



(c) The proposed method

Fig. 5. The confusion matrix chart

TABLE III
TEXT CLASSIFICATION ACCURACY

Type	BWM	CNN	The proposed method
Electronic failure	100%	100%	100%
Leak	75%	87.5%	87.5%
Mechanical failure	91.3%	82.6%	91.3%
Software failure	100%	100%	100%
Total	91.7%	89.6%	93.8%

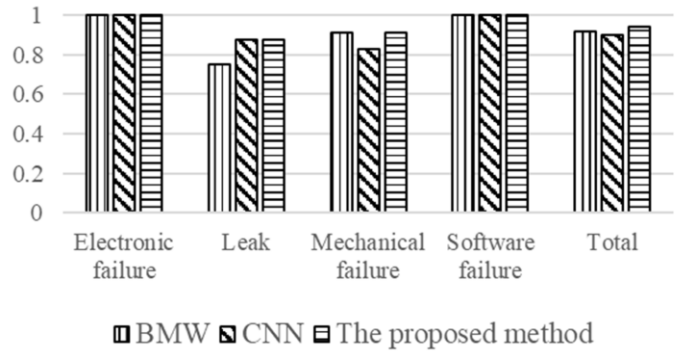


Fig. 6. Histogram of text classification accuracy

VI. Conclusion

The current short text classification algorithms do not comprehensively consider the implicit dependencies and local key information in the text. To solve this problem, a deep convolution neural network model combined with attention mechanism is proposed by using the method of deep learning. The model can capture the implicit dependency feature information of context. Then the model can pay more attention to the key information of text. As a result, the accuracy of text classification is improved. On the fault records of factory equipment data set, the experimental results show that the performance of the proposed model is better than the traditional CNN and bag of word model. In the next step, the model will be applied to the text analysis task combining text and picture information. And it will find a more suitable in-depth learning algorithm for text analysis.

References

[1] Eminagaoglu M, "A new similarity measure for vector space models in text classification and information retrieval," *Journal of Information Science*, no.1, pp1-5, 2020.

[2] Elakkiya E, Selvakumar S, and Velusamy R L, "TextSpamDetector: textual content based deep learning framework for social spam detection using conjoint attention mechanism," *Journal of Ambient Intelligence and Humanized Computing*, pp1-16, 2020.

[3] D'Andrea E, Ducange P, and Bechini A, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Systems with Application*, vol. 116, no. 2, pp209-226, 2019.

[4] Zl A, Hx B, and Gc C, "Word-level emotion distribution with two schemas for short text emotion classification," *Knowledge-Based Systems*, 2021.

[5] Minaee S, Kalchbrenner N, and Cambria E, "Deep Learning--based Text Classification: A Comprehensive Review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp1-40, 2021.

[6] Zhang J, Chang W, and Yu H, "Fast multi-resolution transformer fine-tuning for extreme multi-label text classification," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[7] Yao L, Mao C, and Luo Y, "Graph convolutional networks for text classification," *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no.1, pp7370-7377, 2019.

[8] Chakraverty S, Juneja B, and Pandey U, "Dual lexical chaining for context based text classification," *2015 International Conference on Advances in Computer Engineering and Applications (ICACEA)*, 2015.
[9] Kadhim A I, "Survey on supervised machine learning techniques for automatic text classification," *Artificial Intelligence Review*, vol. 52, no. 1, pp273-292, 2019.
[10] Charbuty B, and Abdulazeez A, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp20-28, 2021.
[11] Esteva A, Chou K, and Yeung S, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, pp1-9, 2021.
[12] Luongo F, Hakim R, and Nguyen J H, "Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery," *Surgery*, vol. 169, no. 5, pp1240-1244, 2021.
[13] Malik M, Malik M K, and Mehmood K, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp9411-9457, 2021.
[14] Hassan M D, Nasret A N, and Baker M R, "Enhancement automatic speech recognition by deep neural networks," *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 4, pp921-927, 2021.
[15] S.F. Woon, S. Karim, and M.S.A Mohamad, "On the Modification of the Discrete Filled Function Algorithm for Nonlinear Discrete Optimization," *IAENG International Journal of Applied Mathematics*, vol. 51, no.4, pp930-935, 2021
[16] Kim Y, "Convolutional Neural Networks for Sentence Classification," *Eprint Arxiv*, vol. 15, no. 6, pp1746-1751, 2014.
[17] Yao X, and Durme B V, "Information Extraction over Structured Data: Question Answering with Freebase," *Meeting of the Association for Computational Linguistics*, pp956-966, 2014.
[18] Conneau A, Schwenk H, and Barrault L, "Very Deep Convolutional Networks for Text Classification," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1, 2017.
[19] Tang D, Bing Q, and Liu T, "Document modeling with gated recurrent neural network for sentiment classification," P*roceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp1422-1432, 2015.
[20] Mittal V, Gangodkar D, and Pant B, "Deep Graph-Long Short-Term Memory: A Deep Learning Based Approach for Text Classification," *Wireless Personal Communications*, no. 4, 2021.
[21] Zhou C, Sun C, and Liu Z, "A C-LSTM Neural Network for Text Classification," *Computer Science*, vol. 1, no. 4, pp39-44, 2015.
[22] Wenhao Pan, and Kai Yang, "Enhanced Multi-Head Self-Attention Graph Neural Networks for Session-based Recommendation," *Engineering Letters*, vol. 30, no. 1, pp37-44, 2022
[23] Pappas N, and Popescu-Belis A, "Multilingual Hierarchical Attention Networks for Document Classification," *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp1480-1489, 2017.

Liu Ronghui acts as an Associate Professor of the School of Computer and Data Science, Henan University of Urban Construction. His interest is neural network and database technology.
Wei Xinhong is an Associate Professor of the School of Computer and Data Science, Henan University of Urban Construction. His interest includes data mining and computer science.