Design of Intelligent Extraction Method for Key Electronic Information Based on Neural Networks

Xiaoqin Chen, Xiaojun Cheng*
School of Intelligent Manufacturing, Chongqing Three Gorges Vocational College, Chongqing 404155, China

***Abstract*—With the rapid development of the Internet and other emerging media, how to find the needed information from massive electronic documents in time and accurately has become an urgent problem. A key electronic information extraction method based on neural network learning ideas has been proposed to solve the problems of time-consuming and difficult deep semantic feature mining in traditional text classification methods. Firstly, a weighted graph model was introduced to improve the TextRank keyword extraction algorithm, helping to capture complex data information and implicit semantics. The results indicate that the optimization method has the highest extraction accuracy (96.52%) on the CSL dataset, and its performance in feature extraction of information data is superior to other comparative models. Secondly, combining LSTM and self attention mechanism to achieve key feature extraction of contextual semantic information. The results indicate that this optimization method has relatively small training and testing errors in data classification, and tends to converge in the later stages of iteration. The accuracy of information extraction reached 94.37%, which is better than other comparative models. The keyword extraction integrity of the fusion model on the THUCNews dataset and Sogou News dataset were 86.2 and 84.1, respectively, with consistency of 96.3 and 94.7, and grammatical correctness of 92.1 and 92.2, respectively. The neural network-based extraction method proposed by the research institute can not only effectively improve the accuracy of information extraction, but also adapt to the changing data environment, and has great potential for application in the field of electronic information processing.**

Keywords*—Key electronic information; intelligent extraction; TextRank; LSTM; context

I. INTRODUCTION

The development of Internet information technology and the popularity of mobile intelligent devices make information interaction possible. According to the statistical report released by the Internet Network Information Center, the Internet penetration rate has reached more than 70% by 2022. With the help of mobile devices, people can access and produce different types of electronic information on media and social platforms, such as electronic documents, emails, social media data, etc. [1]. The generation of massive information data not only facilitates people's lives and work, but also invisibly increases the difficulty of information processing. For information receivers and users, it is necessary to filter, extract and analyze the massive data to achieve higher quality services for users. Electronic information often exists in various channels such as email, social media, and news reports in the form of text, voice, and images. How to accurately and quickly classify and extract key electronic information from a large and complex amount of network information has become one of the important tasks of natural language processing. The electronic information generated through online media has characteristics such as complexity. Classifying its content can not only solve the problem of information disorder to a large extent, but also play an important role in personalized recommendation, information retrieval, and other fields [2]. Traditional electronic information extraction techniques often rely on specific models, such as regular expression matching, which performs well in processing structured data, but often struggle to handle unstructured data such as images, audio, and natural language text.

Statistical analysis methods such as K-nearest neighbor algorithm, support vector machine, and decision tree have certain classification advantages, but they require manual design of rules and feature selection to achieve text classification, which consumes a lot of time and is difficult to mine deep semantic features of the text. Deep network models such as recurrent neural networks, convolutional neural networks, and pre trained models have advantages in information extraction, enabling large-scale dataset learning and automatic feature extraction and semantic association analysis. They can effectively improve the accuracy and efficiency of key electronic information extraction. Current research often relies on keyword extraction algorithms to achieve text data classification, which can reduce sparsity by shortening the length of text sequences. Keyword extraction algorithms are usually more intuitive and easy to implement, as they do not require complex network structure design and long-term model training. Compared to deep learning models that require a large amount of computing resources and data training, this method has lower hardware requirements, faster processing speed, and is easy to deploy [3]. Therefore, the research focuses on the learning approach of neural networks based on keyword extraction algorithms to extract key electronic information. At the same time, considering that keyword extraction algorithms often ignore the dependency relationships and complex semantic expressions of information keywords when processing data, as well as their weak ability to capture complex patterns and implicit semantics, this study proposes improvements to the algorithm. By introducing the TextRank algorithm based on weighted graph model and self attention mechanism, keyword extraction and semantic feature grasping can be achieved, thereby improving the classification performance of key electronic information extraction. TextRank is a widely used keyword extraction algorithm that constructs co-occurrence relationships between words based on a graph model. However, its performance in extracting keywords from Chinese text data is not ideal, as it cannot

*Corresponding Author

consider the positional information of words and the contextual information of the entire corpus. The weighted graph model takes into account the relationships between words and can solve the problem of sparse semantic feature distribution in electronic information text data. Traditional Long Short Term Memory (LSTM) neural networks have the same time series structure as text data and are widely used in natural language processing tasks. However, when it comes to feature extraction of text data, it cannot effectively combine contextual information to extract correct semantic features. The self attention mechanism can reduce the dependency relationship of sequence data and extract correct semantic features while combining contextual information. Therefore, research is being conducted on LSTM with improved self attention mechanism for text classification.

This study is mainly divided into following sections. Section II is Related Works, which summarizes the current research results in information extraction and other aspects. Section III is the research methodology, including key electronic information extraction based on optimized TextRank algorithm and information extraction based on LSTM network. Section IV is result analysis, which mainly tests the research methods. Section V is the discussion. Last Section VI is the conclusion, which summarizes the research results and shortcomings.

II. Related Work

Key information extraction is of great significance in improving information processing efficiency, supporting decision-making, and knowledge management. It can help researchers filter out valuable and relevant information from a large amount of information, providing more efficient, accurate, and intelligent information processing methods. Scholars from different fields have conducted extensive research and achieved some research results. Chi L et al. proposed a graph based and lightweight automatic key phrase extraction method for the automatic extraction of key information. Iterative sentences were used to sort words, generating a more accurate list of key phrases. The research method could effectively improve the extraction accuracy and significantly reduce the number of iterations [4]. Li T et al. designed an optimizer based on auto-regressive method to improve the accuracy of unsupervised key phrase extraction. They integrated any graph based unsupervised key phrase extraction model to enhance stability. This method could improve accuracy on different datasets by 50%, which was beneficial for unsupervised method key phrase extraction [5]. Singh Y et al. proposed an extraction method based on ternary block truncation encoding and binary bat algorithm to improve the efficiency of video summarization and key frame extraction. The method extracted and processed static images in frame form from the input video database, and measured the similarity measure between two consecutive frames. The research method had better extraction accuracy and F-metric value than traditional methods. It was more conducive to the effective and accurate extraction of video summaries and key frames [6]. Sachan M et al. proposed a method based on discourse and text layout features to extract geometric knowledge more effectively from multimedia textbooks. This method could more effectively extract knowledge, thereby improving the solution for geometric knowledge [7].

Many scholars have applied algorithms such as neural networks to information extraction. Zheng J et al. designed a semantic feature extraction method that integrated convolutional neural networks to address the low efficiency of traditional picking methods in identifying phase waves. The key parameters of the network were refined to ensure the extraction accuracy. It could effectively improve extraction efficiency and accuracy [8]. Sonntag D et al. proposed an automatic method for extracting text information in the complex integration process of medical data. This method could simplify the integration process of medical data and improve the accuracy of information extraction, which was beneficial for doctors in clinical diagnosis [9]. Nasar Z et al. proposed a method based on named entity recognition and relationship extraction to extract important information from text data on online platforms. Deep learning methods were used to construct joint models. The joint model based on deep learning had significant advantages. Named entity recognition and relationship extraction were beneficial for extracting key information from text data [10]. Zhang X et al. proposed a method based on multiple information extraction and support vector data description to optimize the process monitoring and fault diagnosis performance of key performance indicators, balancing local process information and mining hidden information. The maximum information coefficient algorithm selected key performance indicator information, extracted local information, and then extracted observed values, cumulative errors, and rate of change information. The verification results indicated that it could effectively improve the extraction accuracy, thereby promoting process monitoring and fault diagnosis of key performance indicators [11].

The above content indicates that effective key information extraction can significantly improve the efficiency of information processing. Some scholars have achieved intelligent information extraction by using fusion convolutional neural networks, entity recognition and relationship extraction, or data description and extraction perspectives. However, there is still room for improvement in the processing and classification of text data, making it difficult to ensure both data processing efficiency and extraction accuracy. The research aims to address the shortcomings of existing technologies in processing electronic information text data by integrating the weighted graph model's TextRank algorithm and self attention mechanism. The TextRank algorithm is a widely used keyword extraction algorithm that does not require complex network design and resource conditions to achieve accurate classification. Its improved idea can also effectively extract contextual semantic information features, especially when facing the problem of sparse distribution of semantic features in electronic information text data. It can effectively address the shortcomings of existing technologies in processing electronic information text data. By implementing the methods proposed in the research, it is expected to achieve more efficient and accurate information processing and classification effects, which will have a profound impact on promoting the development of information technology, supporting decision-making, and other related fields.

III. Key Electronic Information Extraction Based on Optimized TextRank and LSTM Networks

For the key electronic information extraction, TextRank is used to construct a graph model and optimize the TextRank algorithm. To better extract the semantic features of key electronic information by combining context, LSTM network and attention mechanism are introduced to achieve the extraction and classification of key electronic information.

A. Key Electronic Information Extraction based on Optimized Textrank

The TextRank algorithm is a widely used algorithm for extracting critical electronic information. It represents text data by constructing a graph model and establishes co-occurrence relationships between words in the graph. Specifically, this algorithm abstracts unstructured text into a graph structure. Each word is connected to a fixed number of adjacent words. Then, according to a certain formula, the weights of each node in the graph are recursively calculated. Finally, the nodes that rank higher during convergence can be used as keywords for electronic information. One of the advantages of this algorithm is its ability to simplify unstructured text data and extract useful information. It establishes connections between words by using co-occurrence relationships, thereby helping to identify the most important words or phrases in the text [12]. This is very useful for many natural language processing tasks, such as text classification, information retrieval, and summary generation. The TextRank algorithm determines the weights of words by calculating their correlation, as shown in Eq. (1).

$$TR(V_i)=(1-\alpha)+\alpha\times\sum_{V_j\in In(V_i)}\left(\frac{1}{\left|Out(V_j)\right|}\right)\times TR(V_j) \quad (1)$$

In Eq. (1), $In(V_i)$ and $Out(V_j)$ represent the predecessor and follower node sets of V_i and V_j, respectively. $\left|Out(V_j)\right|$ is the number of rear drive nodes for node V_j. $TR(V_i)$ is used to describe the weights of all nodes. α represents the damping coefficient, usually taken as 0.85. The TextRank algorithm does not rely on labeled data and does not require pre-training, but it has some limitations in keyword extraction. One limitation is that it only considers the connection between local words, without fully utilizing the global perspective to analyze the dependency characteristics between words [13]. To optimize TextRank, a key electronic information extraction model can be constructed by combining weighted graphs, which introduces the relationship between words.The text features of electronic information are considered. The word vector algorithm is used to train the word vectors of all electronic information data. The evaluation formula for TextRank algorithm is improved through the mutual information between words. This is the key electronic information extraction method based on optimized TextRank. The specific process is shown in Fig. 1.

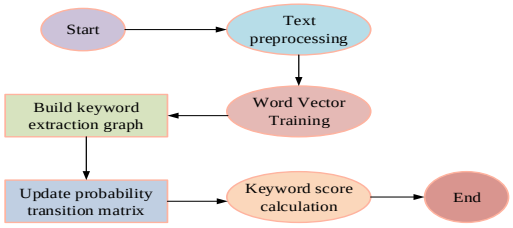

Fig. 1. Key electronic information extraction process based on optimized TextRank.

From Fig. 1, the improved key electronic information extraction method based on optimized TextRank first preprocesses the phone information. Then, word vector training is used to train the preprocessed words into word vectors, which are introduced to calculate the mutual information between words, thereby constructing the probability transition matrix of the graph. Then, the sliding window size of the improved algorithm is set. The weight of all words is calculated by combining the transition probability matrix. The word set is sorted according to the descending value of the weight value. The top ranked words are used as the extracted key electronic information [14], [15]. Among them, language models are generally used to introduce character level combination information to train word vectors. Eq. (2) is the logarithmic natural function of the information.

$$\sum_{t=1}^{T}\sum_{c=C_t}\log p\left(w_c\left|w_t\right.\right) \quad (2)$$

In Eq. (2), w_t represents the target word. w_c is the contextual word for w_t. C_t refers to the index set of contextual words. $p\left(w_c\left|w_t\right.\right)$ is used to describe the probability of w_c occurring on the basis of setting w_t, as shown in Eq. (3).

$$p\left(w_c\left|w_t\right.\right)=\frac{e^{s(w_t,w_c)}}{\sum_{j=1}^{W}e^{s(w_t,j)}} \quad (3)$$

In Eq. (3), $s\left(w_t,w_c\right)$ represents a rating function that maps words w_t and w_c to a rating, which can be combined with the rating value to determine the matching degree between the word and the context [16]. The mapping process of the rating function is generally calculated using the scalar product of the current word vector u_{wt} and the context word vector v_{wc}, as shown in Eq. (4).

$$s\left(w_t,w_c\right)=u_{wt}^{T}v_{wc} \quad (4)$$

After completing word vector training, the mutual information between nodes can be calculated using word vectors to update the probability transition matrix. The probability transition matrix is shown in Eq. (5).

$$W=\begin{bmatrix}W_{1,1} & W_{1,2} & \cdots & W_{1,n}\\ \vdots & \vdots & \ddots & \vdots\\ W_{n,1} & W_{n,2} & \cdots & W_{n,n}\end{bmatrix} \quad (5)$$

In Eq. (5), W_{ij} represents the weight values between nodes. n refers to the words in the corpus. The proportion of each contextual word in the target word should not be the same. The proportion of peripheral words with high relevance to the target word is relatively high. Therefore, the cosine similarity between nodes can be calculated using word vectors, which are used as edge weights. The similarity is shown in Eq. (6).

$$S_{V_1,V_2} = \frac{\sum_{i=1}^{d} V_{1,i} \times V_{2,i}}{\sqrt{\sum_{i=1}^{d} (V_{1,i})^2} \times \sqrt{\sum_{i=1}^{d} (V_{2,i})^2}} \quad (6)$$

In Eq. (3), S_{V_1,V_2} represents the cosine similarity between the word vectors V_1 and V_2. $V_{1,i}$ and $V_{2,i}$ represent the components of word vectors V_1 and V_2, respectively. The word vector calculation obtains cosine similarity to update the probability transition matrix. The updated probability transition matrix is shown in Eq. (7).

$$W = \begin{bmatrix} S_{V_1,V_1} & S_{V_1,V_2} & \cdots & S_{V_1,V_n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{V_n,V_1} & S_{V_n,V_2} & \cdots & S_{V_n,V_n} \end{bmatrix} \quad (7)$$

In Eq. (7), the range of cosine similarity values is [-1, 1]. To standardize the model, the probability transition matrix is normalized. The range of element values in the probability transition matrix is limited to [0,1]. The calculation process is shown in Eq. (8).

$$W_{i,j} = \frac{S_{V_i,V_j}}{\sum_{j=1}^{n} S_{V_i,V_j}} \quad (8)$$

The graph used for keyword extraction is first preprocessed for each word. Then each word is added to the graph in the form of a sliding window. The distance between words is used to indicate the coexistence between two words. When both words are included in a window, it indicates a connection between the two words. The constructed probability transition matrix is used as the weight between nodes. The key electronic information extraction based on optimized TextRank is shown in Fig. 2.

From Fig. 2, the key electronic information extraction graph based on optimized TextRank has a window size of 3. All words are connected to adjacent words before and after [17]. By constructing a weighted graph, the scoring formula for extracting key electronic information can be obtained. The expression is shown in Eq. (9).

$$TR(w_i) = (1-\alpha) + \alpha \times \sum_{w_j \in In(w_i)} \left(W_{i,j} \times TR(w_i) \right) \quad (9)$$

In Eq. (9), $In(w_i)$ represents the predecessor node sets of node w_i. $TR(w_i)$ represents the weights of all nodes. $W_{i,j}$ represents the transition probability between nodes.

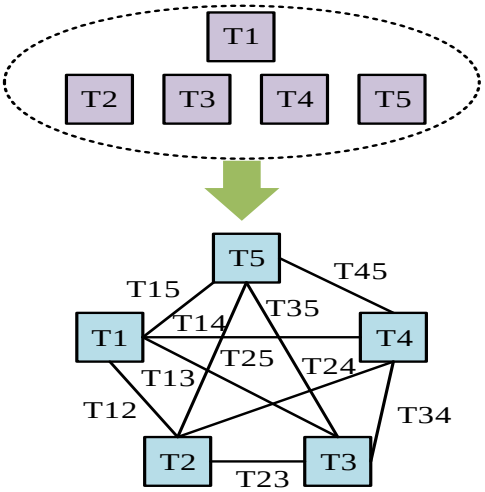

Fig. 2. Key electronic information extraction based on optimized TextRank.

B. Feature Extraction based on LSTM Network and Attention Mechanism

The semantic complexity of key electronic information can affect the extraction performance of TextRank. To better combine context to extract the semantic features of key electronic information, LSTM and attention mechanism are introduced to achieve the extraction and classification of key electronic information. LSTM is a special type of recurrent neural network (RNN) used to solve the gradient vanishing and exploding faced by traditional RNNs. The LSTM network controls the inflow and outflow of information by introducing a structure called a "gate". This network mainly includes input gates, forget gate, output gate (OG), and cell state. The input gate determines which information can enter the cellular state. The forget gate determines which information needs to be forgotten from the cellular state. The OG determines the output of information in the cell state after activation. The key to LSTM lies in their ability to "remember" and "forget" information. Through the forget gate, LSTM can selectively remove information from the cell state, avoiding irrelevant information from interfering with subsequent calculations. The input gate can control which newly inputted information can be added to the cell state. This memory ability to retain long-term dependencies makes LSTM perform well in processing sequence data. Attention mechanism is a technique that mimics human attention behavior. This method is used to assign different weights and attention levels to different parts of the input in deep learning models. In traditional neural networks, all input information is processed simultaneously. Attention mechanisms can enable the model to selectively focus on certain parts of the input, thereby improving the performance and effectiveness. The core idea of attention mechanism is to determine the importance of each input based on its input. Then, weighted processing is performed based on these importance levels. In deep learning models, attention mechanisms can be used at different levels and time steps. The key electronic information extraction based on LSTM network and attention mechanism is shown in Fig. 3.

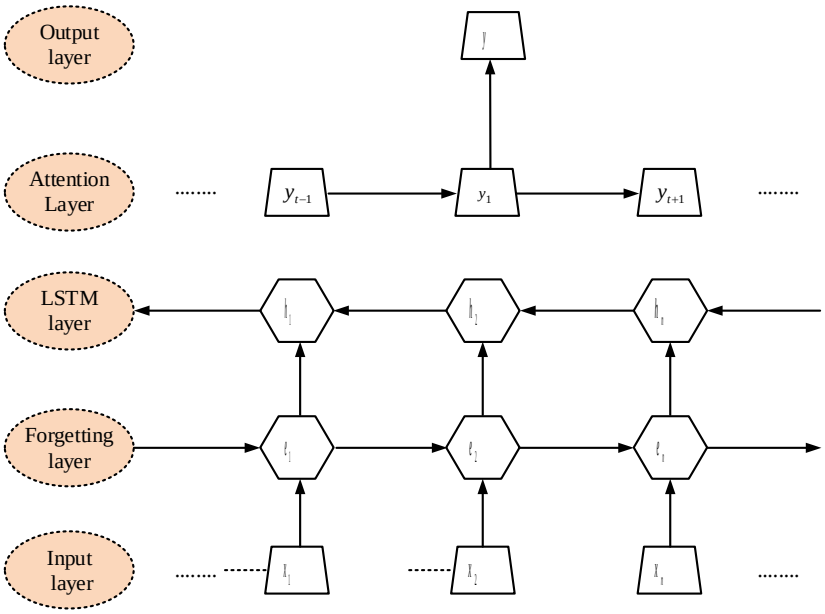

Fig. 3. Key electronic information extraction based on LSTM network and attention mechanism.

From Fig. 3, the extraction model has five parts. Time series data is used as the input layer to extract word vectors using LSTM. The self attention mechanism is applied to calculate the weight of the text features output by LSTM layer, so that it focuses on the main content of the overall text. The output layer uses sequence level feature vectors to achieve text classification. The LSTM network is used for feature extraction of input text. Based on this, the semantic features of each word are described. Compared to traditional RNNs, LSTM has added three gating mechanisms, namely, forget gate, input gate, and OG [18]. The forgetting gate determines how many past states one should maintain in their current state. The input gate determines how much input information has been retained within the current time period. The OG determines the size of the current state output. The unit structure of the LSTM network is displayed in Fig. 4.

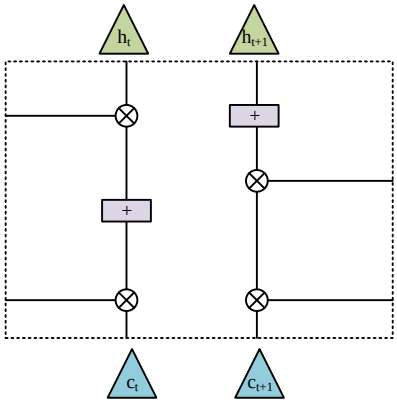

Fig. 4. The unit structure of LSTM networks.

The forgetting gate can control the proportion of information to be forgotten. The input is the output h_{t-1} of the previous moment and the text information input x_t of the current moment. The calculation process of the forget gate is shown in Eq. (10).

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{10}$$

In Eq. (10), σ represents the sigmoid function. W_f and U_f are the weight matrices of the forget gate. b_f represents the bias term of the forget gate. The input gate is mainly the information stored in the state unit, which can calculate the memory information of the current unit [19]. The unit state at the current moment is calculated by the sum for the product of the forgetting gate input and the previous moment state, as well as the input. The calculation process is shown in Eq. (11).

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \tag{11}$$

In Eq. (11), c_t and c_{t-1} represent the current and previous unit states, respectively. The OG function can determine the current hidden state and save the previous input information in the hidden state. The input of the current time and the hidden state for the previous time are respectively inputted into the sigmoid function. The current unit state is inputted into the tanh function. The output result of the sigmoid function is multiplied by the tanh function to obtain the hidden state in the current time. Afterwards, the text sequence is processed using the LSTM network to obtain the implicit state h_t at each moment, which characterizes the semantic features of each word in the text and achieves text classification. In the process of extracting electronic news information, it is also necessary to combine contextual information. The semantic feature extraction based on a single word has one sidedness, which has a certain impact on the model's judgment and thus reduces the classification accuracy [20]. Therefore, the study intends to introduce attention mechanisms into the text. The word features of LSTM are analyzed to achieve text feature extraction that integrates context. This study draws on the self attention mechanism of the Transformer model. Semantic feature extraction that integrates with context can achieve automatic mining of target text without relying on source data.

IV. Analysis of Intelligent Extraction Results for Key Electronic Information

In order to improve the intelligent extraction of key electronic information, an improved TextRank algorithm and LSTM-SAttention model were designed to enhance data classification performance and feature extraction accuracy. In the results section, the study mainly analyzed from two aspects: technical performance evaluation and application result verification.

A. Test Data Source and Experimental Environment Parameter Design

The algorithm was tested using the Chinese Scientific Literature (CSL) keyword extraction dataset and the NLPCC2017 keyword extraction dataset publicly available at the 2017 Academic Annual Meeting of the Natural Language Processing Professional Committee of the Chinese Computer Society. The CSL dataset mainly involves key electronic information content in the computer field, while the NLPCC2017 dataset mainly involves news articles and their keywords. De duplication and filtering were performed on these two datasets to remove duplicate, formatting errors, missing key fields, or incomplete content, resulting in 7562 CSL data and 7635 NLPCC2017 data. In the performance test results section, the study introduces LSTM network and attention mechanism into optimizing the TextRank algorithm to construct a key electronic information intelligent overall model. The THUCNews dataset published by Tsinghua University is taken as the research object, and ten categories of Chinese information (including social, technological, educational, etc.) are selected, each containing 20000 pieces of information, forming a balanced dataset. In addition, the "Sogou News" dataset released by Sogou Lab also contains ten categories of news, which are divided into training and validation sets in a 7:3 ratio. The parameter settings for the experimental environment are shown in Table I.

B. Experimental Test Results

To test the effectiveness of sliding windows in extracting key electronic information, the study first analyzes the keyword extraction data set and the extraction algorithm based on optimized TextRank. Firstly, the test results of two datasets at different window sizes are shown in Fig. 5.

TABLE I. System Parameter

Number	Testing environment	Parameter
(1)	Processor	Intel(R) Core (TM)i5-7300HQ CPU@2.50GHz
(2)	Operating system	Windows 10
(3)	Programming Language	C++
(4)	Memory	32GB
(5)	GPU	GTX 1050Ti
(6)	Programming Language	Python 3.7.3

From Fig. 5, when only considering the impact of association distance on keyword extraction results, a window size of 2 had the best extraction effect on keywords. To fully utilize the global text information and fully consider the order characteristics of the text, the optimized TextRank algorithm is used to represent words in vector form. The similarity between words is calculated. Then it is introduced as a weight into the model. Fig. 6 displays the specific results.

From Fig. 6(a-b), the keyword extraction performance of the research method was superior to other methods. When the window was set to 2, the F1 values of TextRank before optimization, TextRank after optimization, and traditional methods all reached their highest. The F1 values on the CSL data set were 0.59, 0.51, and 0.32. On the NLPCC2017 data set, they were 0.56, 0.47, and 0.17. From Fig. 6(c-d), the extraction accuracy of the three algorithms increased with time. Their accuracy on the CSL data set exceeded the NLPCC2017 data set. The optimized TextRank algorithm is tested on the CSL data set. When the extraction time was 10s, the extraction accuracy stabilized at 92%, while the pre-optimized algorithm stabilized at over 70% after 30s. The traditional method stabilized at over 55% after 30s of extraction.

Next, the performance of the proposed text extraction algorithm that integrates keyword extraction and attention mechanism is validated. The proposed algorithm is compared with other deep learning based text extraction algorithms. The experiment aims to construct an LSTM-SAttention feature extraction network. It is applied to two datasets to test the effectiveness. The results are shown in Fig. 7.

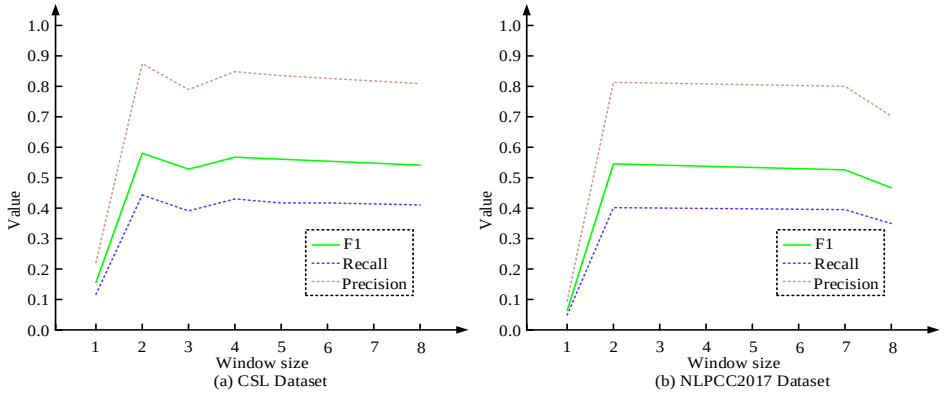

Fig. 5. Results of different datasets under different windows.

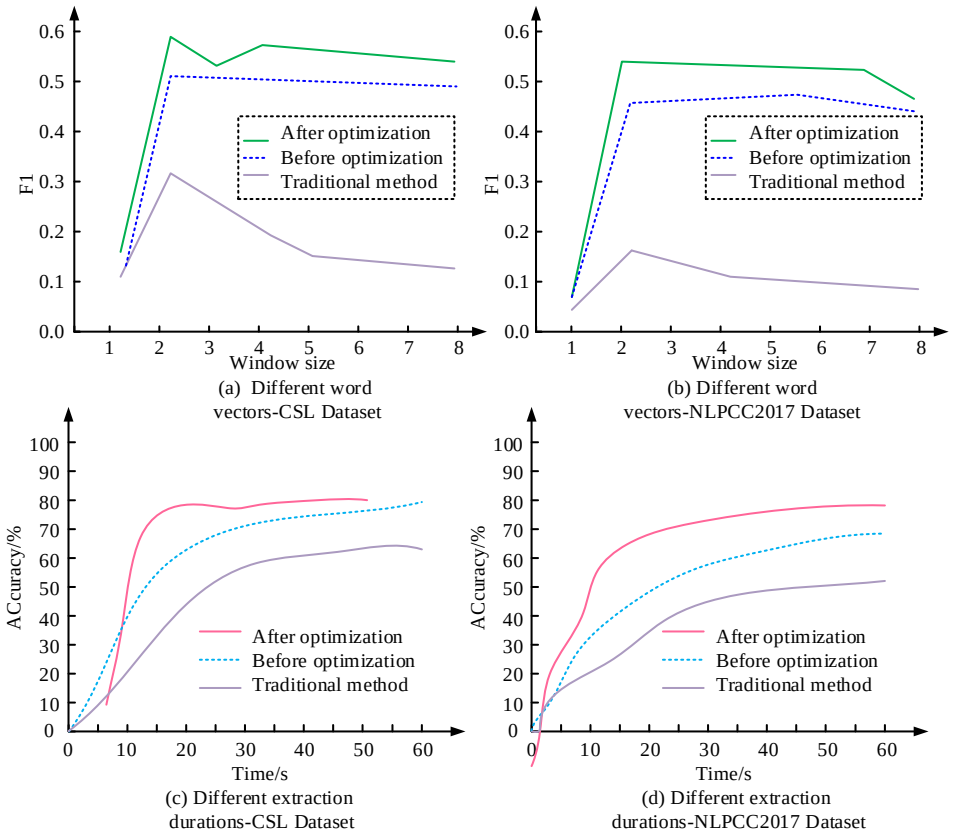

Fig. 6. Keyword extraction performance under different word vectors and extraction times.

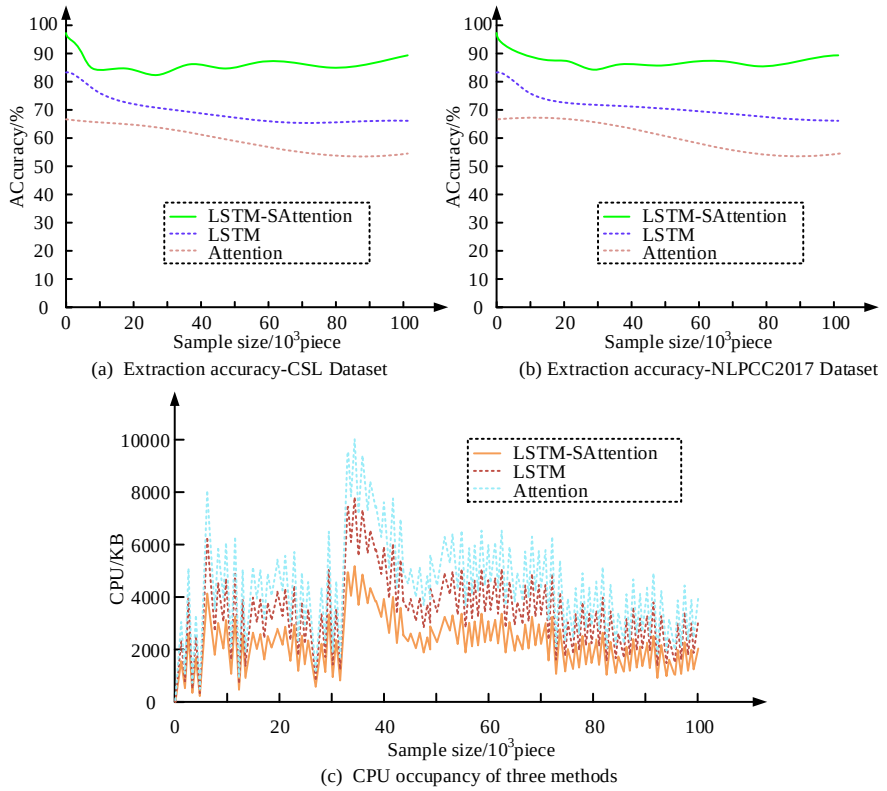

Fig. 7. Extraction accuracy and CPU usage.

From Fig. 7(a-b), as the sample size increased, all three algorithms showed a trend of decreasing extraction accuracy. However, the research method stabilized faster showed high accuracy and stability on both datasets. Among them, the highest extraction accuracy of LSTM-SAttention on the CSL data set was 96.52%. The highest extraction accuracy on the NLPCC2017 data set was 71.21%. Fig. 7(c) shows the result of memory operation. From Fig. 7(c), as the sample size increased, the CPU occupancy of all three algorithms gradually increased. However, the occupancy rate of research methods was the lowest. The increase rate was also the slowest. Overall, the key electronic information extraction method combining attention mechanism and LSTM network performs the best. Subsequently, the application effect of the LSTM-SAttention model proposed in the study was analyzed and compared with the Enhanced Support Vector Machine (E-SVM) algorithm, Graph Convolutional Networks (GCN-FCN), Large Language Model Meta AI, and Multi level Semantic Alignment Image Text Matching Algorithm (MLS-ITM). The results are shown in Table II.

The results in Table II indicate that the LSTM-SAttention model outperforms other algorithms in terms of performance evaluation on both databases. Its maximum accuracy feature extraction results on the CSL dataset differ from those of the E-SVM algorithm by over 10%, while the difference between the LSTM-SAttention model and the MLS-ITM algorithm is within 5%. The accuracy, recall, and F-value of the LSTM-SAttention model all exceed 90, followed by the well performing MLS-ITM algorithm and GCN-FCN algorithm. The Meta AI model's values do not exceed 85. On the NLPCC2017 dataset, the LSTM-SAttention model (90.47)>MLS-ITM (87.12)>GCN-FCN (83.25)>Meta AI (80.07)>E-SVM (70.32). The above results indicate that the LSTM-SAttention model can achieve good feature extraction.

C. Actual Inspection Results

Using Chinese news text classification as the test object, the proposed TextRank optimization extraction model and the Vocabulary Semantic Map Attention Mechanism (VSA), SENet and Convolutional Neural Network fusion method (Squeeze and Excitation Networks Convolutional Neural Network, SENet CNN), as well as the Convolutional Neural Network Bi directional Long Short Term Memory (TC Ablstm) fusion parallel neural network model were analyzed for information extraction results. Fig. 8 shows the training and testing errors of data processing.

From the testing process in Fig. 8(a), it can be seen that the testing errors of the four algorithms all show a decreasing trend with the increase of iteration times. Before the iteration times are less than 150, the average testing error results from large to small are: VSA algorithm>SENet CNN algorithm>TC Ablstm algorithm>the proposed algorithm. As the number of iterations gradually increases, the testing errors between algorithms are all less than 0.025, and the error difference between algorithms does not exceed 5%. The overall variation of the testing error of the TextRank optimization extraction model proposed in the study is relatively small, and tends to converge in the later stages of iteration. From the testing process in Fig. 8(b), it can be seen that the error comparison between different algorithms is relatively large, with the order of error from small to large: the proposed algorithm>TC Ablstm algorithm>SENet CNN algorithm>VSA, with an average testing error of 2.36%>5.64%>13.26%>33.64%. At the same time, the algorithm proposed in the study showed a greater slope of decrease in the training error curve when the number of iterations was less than 75, and in the later stage, the error curve became more stable, further improving the classification accuracy. Subsequently, the information extraction results of the proposed fusion algorithm were analyzed, and the results are shown in Fig. 9.

The results in Fig. 9 indicate that the PR curves of the TextRank optimization extraction model and the TC Ablstm model are closer to the bottom right corner. The accuracy of VSA, SENet CNN, and TC Ablstm algorithms are 78.24%, 85.69%, and 86.78%, respectively, while the proposed TextRank optimization hybrid model achieves an accuracy of 94.37% in information extraction. Chinese news text is classified as the test object. The ROUGE Scores, integrity, consistency, and grammar correctness of the intelligent extraction model for key electronic information are analyzed. Among them, ROUGE Scores are mainly used to evaluate the quality of automatically generated abstracts or translations. Integrity measures whether the key information extraction covers all important parts of the document. Consistency measures the consistency between the extracted key information and the manually marked key information. Grammar correctness assesses the generated grammar level of the text, ensuring good readability and no grammar errors. The specific results are shown in Fig. 10.

Fig. 10(a) and 10(b) show the ROUGE Scores, integrity, consistency, and grammar correctness scores of the model for keyword extraction and document summarization generation on the THUCNews data set and Sogou news data set, respectively. From Fig. 10, the model had high ROUGE Scores, integrity, consistency, and grammar correctness scores on both datasets. When the sample size was 6000, the ROUGE Scores were 89.6 and 88.2, with integrity, of 86.2 and 84.1. The consistency was 96.3 and 94.7, respectively. The correctness of the method was 92.1 and 92.2, respectively. The model constructed by the research method can accurately and completely extract key electronic information, and ensure the accuracy of grammar.

TABLE II. COMPARISON OF TRAINING EXPERIMENT RESULTS OF DIFFERENT MODELS

Contrast model	CSL database			NLPCC2017 database		
	Accuracy	Recall	F value	Accuracy	Recall	F value
E-SVM	77.47	80.52	79.21	70.32	70.23	70.44
Meta AI	82.25	81.49	81.71	80.07	79.32	80.11
GCN-FCN	84.36	85.32	84.38	83.25	82.43	83.48
MLS-ITM	86.43	87.25	88.12	87.12	89.24	88.75
LSTM-SAttentionModel	91.22	90.32	92.23	90.47	91.16	90.52

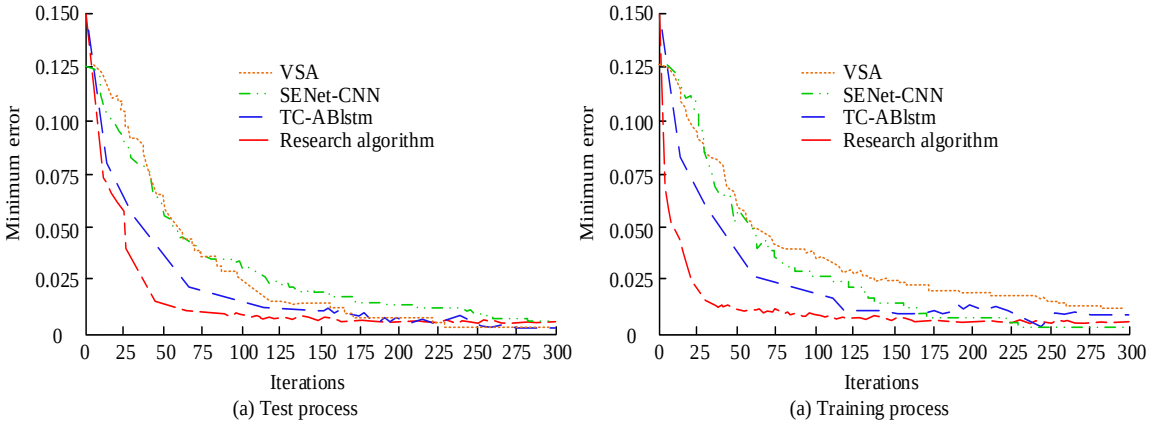

Fig. 8. Comparison of error results during testing and training.

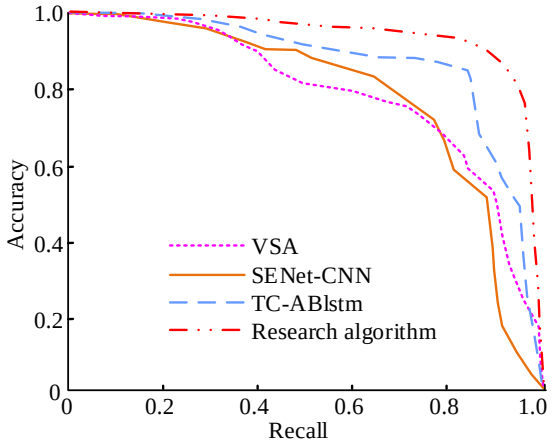

Fig. 9. Accuracy results of information extraction using different algorithms.

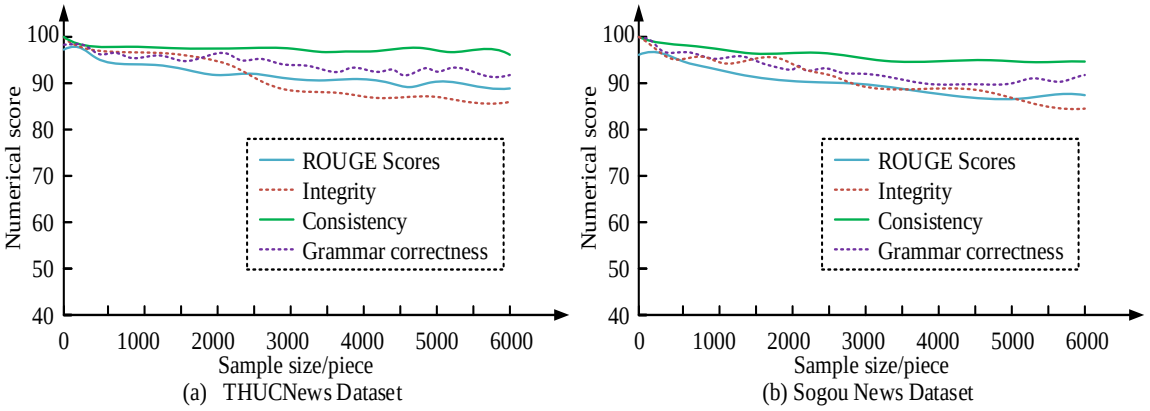

Fig. 10. Test results of intelligent extraction model for key electronic information.

V. Discussion

Performance analysis and case testing were conducted on the proposed electronic information extraction algorithm, and compared with different algorithms. The results showed that when the window size was 2, the optimized TextRank achieved F1 values of 0.59 and 0.56 for keyword extraction on the CSL dataset and NLPCC2017 dataset, respectively. When the extraction time was 10 seconds, the extraction accuracy remained stable at over 92%. The LSTM-SAttention model showed higher accuracy and stability on both datasets. The reason is that the LSTM attention model mainly combines context to extract key electronic information, which can grasp the correlation between different information, so information extraction has high credibility. Introducing a self attention mechanism enables the model to focus more on the information rich parts when processing sequential data, thereby making feature extraction more accurate and efficient. Other algorithms, such as VSA, SENet CNN, and TC Ablsm, have also been improved, but there is still a gap in error convergence speed and final stability. The introduction of self attention mechanism can help the model capture long-range dependencies in text data and provide more refined feature representations for the model. The GCN-FCN model is difficult to consider the time series characteristics of text data, and the E-SVM model has significant dependence on structure and cannot simultaneously consider the spatiotemporal differences of text information. The Meta AI language model performs well in data processing, but it heavily relies on the level of data training, while the MLS-ITM model struggles to meet different information extraction needs, such as keyword localization and relationship recognition, with good integrity and consistency in information extraction on both datasets. When the sample size is 6000, the ROUGE score, completeness, consistency, and grammatical correctness score of the TextRank optimized hybrid model exceed 88, 84, and 94 points, respectively, which can better ensure the semantic integrity and consistency of the text. Compared with literature [8] fusion convolutional neural network, the TextRank optimized hybrid model can effectively ensure the integrity of semantic information. Compared with the automatic extraction of text information by literature [9], the TextRank optimized hybrid model has better information expansion ability. Compared with the information extraction method proposed by literature [11], the TextRank optimized hybrid model can not only extract key information, but also better grasp the correlation between sentence information before and after.

The proposed model has good information extraction performance and efficiency, but there are still issues that need improvement, such as the introduction of self attention mechanism, which will increase the time and cost to a certain extent. Therefore, further optimization of neural network parameters is needed in the future. At the same time, the proposed method will be combined with other intelligent algorithms to improve cross modal information processing and feature fusion, and enhance the good adaptability of data characteristics. It is expected to provide reference for the field of electronic information processing and the improvement of information extraction services.

VI. Conclusion

With the explosive growth of electronic information, how to automatically extract key information from a large amount of text data has become increasingly crucial for information processing and analysis. In this context, the intelligent extraction method of key electronic information based on neural networks has become a hot and challenging research topic. For the key electronic information extraction, the TextRank algorithm is used to construct a key information extraction model and optimize it. A key electronic information extraction model based on optimized TextRank is obtained. Combined with attention mechanism and LSTM network, it facilitates context connection and improves the accuracy of extracting key electronic information. From the research results, with the accumulation of extraction time, the extraction accuracy of the optimized TextRank algorithm increase. The accuracy on the CSL data set exceeded on the NLPCC2017 data set, with a stable extraction accuracy of over 92%. The research method showed high accuracy and stability on both datasets. The keyword extraction and document summary generation effects of the research method were tested. The ROUGE Scores on the THUCNews data set and Sogou news data set were 89.6 and 88.2, respectively. The integrity was 86.2 and 84.1, the consistency was 96.3 and 94.7, and the grammatical correctness was 92.1 and 92.2, respectively. This indicates that the research method has a good effect on extracting key electronic information.

Acknowledgment

The research is supported by: The 2023 Chongqing University Research Project "Construction and Key Technology Application Research of Orange Honey Quality Traceability System Based on Blockchain Service Network (BSN)" (KJZD-K202303502);

The 2022 Chongqing Vocational Education Teaching Reform Research Project "Three dimensional Integration, Five party Collaboration: Exploration and Practice of the Long Term" Green Manufacturing "Talent Integration Training Model" (GZ223113).

References

[1] Zhang M, Bo X U, Xiaoyun L I, Dong FU, Liu J, Baojian WU, Qiu K. Artificial Neural Network-Based QoT Estimation for Lightpath Provisioning in Optical Networks. IEICE Transactions on Communications, 2019, E102.B(11):2104- 2112.

[2] Gao L, Li X, Liu D, Wang L, Yu Z.A Bidirectional Deep Neural Network for Accurate Silicon Color Design. Advanced Materials, 2019, 31(51):1905467.1-1905467.7.

[3] Wang K, Liu M. A feature, ptimized Faster regional convolutional neural network for complex background objects detection. IET Image Processing, 2021,15(2):378-392.

[4] Chi L, Hu L. ISKE: An unsupervised automatic keyphrase extraction approach using the iterated sentences based on graph method. Knowledge-Based Systems, 2021, 223(6):107014.1-107014.12.

[5] Li T, Hu L, Li H, Sun C, Li S, Chi L. Towards unsupervised keyphrase extraction via an autoregressive approach. Knowledge- based systems, 2023,274(Aug.15): 1.1-1.10.

[6] Singh Y, Kaur L. Effective key-frame extraction approach using TSTBTC–BBA. IET Image Processing, 2020, 14(4):638- 647.

[7] Sachan M, Dubey A, Hovy E H, Mitchell TM, Xing EP. Discourse in Multimedia: A Case Study in Extracting Geometry Knowledge from Textbooks. Computational Linguistics, 2019, 45(8):1-35.

[8] Zheng J, Shen S, Jiang T, Zhu W. Deep neural networks design and analysis for automatic phase pickers from three-component microseismic recordings. Geophysical Journal International, 2020, 220(1):323-334.

[9] Sonntag D, Profitlich H J. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. Artificial intelligence in medicine, 2019, 93(JAN.):13-28.

[10] Nasar Z, Jaffry S W, Malik M K. Named Entity Recognition and Relation Extraction: State-of-the-Art. ACM computing surveys, 2022,54(1):20.1-20.39.

[11] Zhang X, Ma L, Peng K. A novel key performance indicator oriented process monitoring method based on multiple information extraction and support vector data description. The Canadian Journal of Chemical Engineering, 2022,100(5):1013- 1025.

[12] Hayat S, Kun S, Shahzad S, Suwansrikham P, Mateen M, Yu Y. Entropy information-based heterogeneous deep selective fused features using deep convolutional neural network for sketch recognition. IET Computer Vision, 2021,15(3):165-180.

[13] Hamouda M, Ettabaa K S, Bouhlel M S. Smart Feature Extraction and Classification of Hyperspectral Images based on Convolutional Neural Networks. IET Image Processing, 2020, 14(10):1999-2005.

[14] Chunhao D, Peng C, Kang O. Enhanced high-order information extraction for multiphase batch process fault monitoring. The Canadian Journal of Chemical Engineering, 2020, 98(10):2187-2204.

[15] A L Y, B W G, D L J C. Keyword guessing attacks on a public key encryption with keyword search scheme without random oracle and its improvement. Information Sciences, 2019, 479:270-276.

[16] Zhu E, Sheng Q, Yang H, Li J. A unified framework of medical information annotation and extraction for Chinese clinical text. Artificial intelligence in medicine, 2023,142(Aug.):1.1-1.12.

[17] Wong R L, Sagar M, Hoffman J, Huang C, Gore JL. Clinical accuracy of information extracted from prostate needle biopsy pathology reports using natural language processing. Journal of Clinical Oncology, 2021, 39(15_suppl):1557- 1557.

[18] Xingchen Z, Yan G, Xian-Min M, Cao Y, Chen X, Chen Z, Du W, Fu L, Luo Z. Extracting photometric redshift from galaxy flux and image data using neural networks in the CSST survey. Monthly Notices of the Royal Astronomical Society, 2022,512(3):4593- 4603.

[19] Luo N, Yu H, You Z, Li Y, Zhou T, Jiao Y, Han N, Liu C, Jiang Z, Qiao S. Fuzzy logic and neural network-based risk assessment model for import and export enterprises: A review. Journal of Data Science and Intelligent Systems, 2023, 1(1): 2-11.

[20] Xiong Y, Chen Y, Chen C, Wei X，Wang P. An Odor Recognition Algorithm of Electronic Noses Based on Convolutional Spiking Neural Network for Spoiled Food Identification. Journal of The Electrochemical Society, 2021, 168(7):1-9.