Grid text classification method based on DNN neural network

Jutao Huang, Jiesheng Zheng, Shang Gao*, Wenbin Liu, and Jiaxin Lin
Guangdong Power information Technology Co., Ltd. Yuedian Building, 6-8 Shuijun Road, Yuexiu District, Guangzhou, Guangdong, China

Keywords: Natural language processing, Text classification, DNN network.

Abstract. With the rapid development of network technology, the electric power Internet of Things needs to face a large number of electronic texts and a large number of distributed data access and analysis requirements. If the system wants to complete accurate and efficient data analysis and build an existing data and service standard system covering the entire chain of energy and power business on the existing basis, it must implement massive electronic text retrieval, information extraction and classification in the power grid system. In order to achieve this purpose, a DNN neural network classification model is constructed to classify the text information of the power grid, and the effectiveness of the method is verified by experiments based on data from the substations information system.

1 Introduction

With the rapid development of network technology, the power IoT is facing the access of a large number of distributed data, the concurrency of multiple services, the need for joint data analysis, and the sharp increase in electronic text (such as grid customer information, grid business data, etc.). If the system wants to complete accurate and efficient data analysis and build an existing data and service standard system covering the entire chain of energy and power business, expand the data model, and support unified management of all data on the existing basis, it must implement massive electronic text retrieval, information extraction and classification in the power grid system.

Text Classification refers to a technique for classifying a given text object in a fixed category that has been defined based on the characteristics of the text. It is one of the main research issues of Natural Language Processing (NLP). Typical applications include judging spam, automatic web page classification [1], sentiment classification [2], and news personalized recommendation [3]. The initial solution is to rely on the word matching method in the document to classify the document [4], but the algorithm is mainly done by manpower, the efficiency is not high, and the classification result can not meet the requirements. On this basis, people have studied the vector space model and the knowledge engineering method, *Corresponding author: gaoshang8202@126.com*
but there is still a problem of low accuracy. With the development of machine learning algorithms, algorithms such as SVM model [6], Bayesian network [7], and decision tree have also begun to be applied to text classification. Nowadays, the rapid development of artificial intelligence (AI) technology has led to new developments in text categorization, making it an important branch of natural language processing (NLP) in the AI subfield. Neural networks, such as convolutional neural networks (CNN) and deep neural networks (DNN)[10], are also increasingly being applied to text categorization. This paper uses DNN neural network to classify grid text.

2 Preparation

2.1 Natural language processing (NLP)

Natural Language Processing (NLP) is a human-computer interaction method that allows computers to understand the natural language used by humans to achieve functions such as human-computer interaction and language translation [11]. It is an important branch of artificial intelligence, involving three areas of artificial intelligence, linguistics and computer science. From the perspective of linguistics, language can be divided into formal language and natural language. Formal language is a human-created language that can be processed by machines and symbols, such as programming languages and chemical symbols. Naturally evolved languages, such as human language, are natural languages. Compared with formal languages, they lack a fixed format, and there are a large number of ambiguous statements, similar statements, etc., so that they cannot be directly understood by machines. Sentence understanding, expression learning, and choice of context for human language are highly complex for machines. Natural language processing is a discipline that studies how to process natural language to achieve human-computer interaction.

Technologies related to NLP include named entity recognition, part-speech tagging, dependency parsing, text semantic similarity analysis, document analysis, text classification and machine translation, etc. Besides, text classification is the focus of this paper. Since natural language is a language evolved from a large number of people for long conversations, it is an "empirical" language model that can be modeled using statistical-based models. Therefore, by collecting large-scale real language text to format the real language library, and analyzing the language library using statistical techniques, the language text can be classified. Text classification is generally divided into three steps: text preprocessing, text feature extraction and text classification.

2.2 DNN

The full name of DNN is called deep neural network[13]. The neurons of the DNN are fully connected and do not contain convolutional units. The depth of the DNN refers specifically to the number of layers of the neural network. The original neural network had only the input layer, the output layer, and an implicit layer, called the perceptron, which could not perform complex operations. Later, in order to overcome this shortcoming, experts invented a multilayer perceptron with multiple hidden layers. Multilayer perceptrons use functions such as sigmoid to simulate the response of neurons to excitation and use backpropagation algorithms for network training. However, as the number of network layers deepens, the gradient disappearance problem becomes very serious, and the result of the optimization function is more likely to fall into the local optimal solution. In order to solve the problem of local optimal solution, Hinton adopted a pre-training method, which can make the number of
layers of the neural network reach 7 layers[13]. In addition, the use of ReLU and other functions instead of sigmod solves the problem of gradient disappearance, which constitutes the basic form of current DNN. A three-layer DNN network model structure is shown in Figure 1.

![DNN network model structure](image)

**Fig. 1.** DNN network model structure.

$L_1$ is the input layer, $x_1, x_2, x_3$ are the input data, $L_2$ layer is the hidden layer, $l_i^{(k)}$ is the output of the i-th neuron of the $L_k$ layer. Each layer needs an activation function, assuming the activation function is $\sigma(x)$, we use $w_{ij}^{(k)}$ to represent the weight parameter of the j-th neuron in the $L_k$ to the i-th neuron in the $L_{k+1}$, add an $x_0$, let $x_0=1$, then:

$$l_i^{(2)} = \sigma(w_{1i}^{(1)}x_0 + w_{2i}^{(1)}x_1 + w_{3i}^{(1)}x_2 + w_{4i}^{(1)}x_3 + b_i^{(1)})$$

Then $l_i^{(2)}$ is used as the input of the $L_3$ layer into the activation function, let $l_0^{(2)}=1$, then the output of each neuron in the $L_3$ layer is

$$l_i^{(3)} = \sigma(w_{1i}^{(2)}l_0^{(2)} + w_{2i}^{(2)}l_1^{(2)} + w_{3i}^{(2)}l_2^{(2)} + w_{4i}^{(2)}l_3^{(2)} + b_i^{(2)})$$

Therefore, for a neural network of depth $n$, if the number of neurons in each layer is $m_u \ (u = 1, ..., n)$, the output of the i-th neuron in the $L_k$ is

$$l_i^{(k)} = \sigma(\sum_{j=0}^{m_u} w_{ij}^{(k-1)}l_j^{(k-1)} + b_i^{(k-1)})$$ (1)

### 3 Method construction

In this section, a text classification model based on DNN neural network is proposed to solve the problems of text classification in the power grid industry. The model is mainly divided into three parts: preprocessing stage, feature extraction and text classification. Figure 2 shows the three-layer framework of the model.
3.1 Preprocessing stage

In the text classification process, due to the diversified characteristics of grid data, most of the stored data is unstructured data. Faced with this complex data, computers cannot directly process it. This requires pre-processing the text and transforming it into a form that can be recognized by a computer. This paper uses the ICTCLAS Chinese lexical analysis system of the Chinese Academy of Sciences to perform word segmentation preprocessing and uses a vector space model (VSM) to pattern the text.

Assume a certain text \( X \) in the document set \( Y \), where the number of documents of \( Y \) is \( N \). A vector space model is a model that uses vectors to represent data. Through the patterning of vector spaces, it can reduce the difficulty of text classification. For text \( X \), \( X' = \{(x_i, w_i)\}_{i=1}^n \) can be obtained from the vector space model, where \( n \) is the number of words in text \( X \), \( x_i \) is \( i \)-th word in text \( X \), and \( w_i \) is the feature weight corresponding to \( x_i \). The details are shown in the following Equations 2:

\[
w_i = f_i \log_2 \left( \frac{N}{m_{x_i}} + 0.01 \right)
\]

where \( f_i \) is the number of occurrences of \( x_i \) in document \( X \), \( m_{x_i} \) is the total number of texts in which \( x_i \) appears in set \( Y \). Normalize it, then \( w_i \) is shown in Equation 3:

\[
w_i = \frac{f_i \log_2 \left( \frac{N}{m_{x_i}} + 0.01 \right)}{\sqrt{\sum_{i=1}^{n} \left( f_i \log_2 \left( \frac{N}{m_{x_i}} + 0.01 \right) \right)^2}}
\]

3.2 Feature extraction

The text vector space \( X' = \{(x_i, w_i)\}_{i=1}^n \) is obtained after the preprocessing module. Suppose the corresponding category set of document set \( X \) is \( \mathcal{C} = \{c_k\}_{k=1}^l \), where \( l \) is the number of categories. The amount of grid data is very large, so the number of features after data preprocessing is often considerable. If the text is directly classified without any processing, it will not only have a certain impact on the accuracy of the classification model, but also its classification efficiency is not high. For these considerations, we need to extract features from \( X' \), select the feature vectors that are most conducive to classification, and improve efficiency and accuracy for subsequent classification. This article uses improved mutual information (MI) for feature selection and extraction.

Since the mutual information (MI) only considers the relationship between \( x_i \) and text category \( c_k \), this paper considers that the choice of features will also receive the influence of the frequency of \( x_i \) in the entire text set \( Y \) to a certain extent. By improving the MI algorithm, it is shown in Equation 4:

\[
MI_{x_i}^{c_k} = \alpha \sum_{k=1}^{l} P_{c_k} \log \frac{P_{c_k}^{x_i}}{P_{x_i}}
\]

where \( P_{c_k} \) represents the proportion of documents belonging to \( c_k \) in the set \( Y \), \( \alpha \) is the control threshold, and \( P_{c_k}^{x_i} \) is the proportion of text containing the word \( x_i \) belonging to the text category \( c_k \). Its expression is shown in Equation 5 below:

\[
P_{c_k}^{x_i} = \frac{1 + \frac{h_{c_k}}{f_i}}{\sum_{k=1}^{h_{c_k}} f_k}
\]
where $h_{c_k}$ is the number of texts belonging to the category $c_k$, $S_u$ is the total number of words belonging to the category $c_k$, and $F_k$ is the number of all words belonging to the $c_k$ category.

Set a proper feature selection threshold $\beta$, select words with mutual information values higher than the threshold $\beta$, and treat them as text feature values for text classification.

### 3.3 Text Categorization

Assume that the corresponding feature vector of the text $X$ obtained after the above preprocessing and feature extraction is $X'' = \{(x_i, w_i)\}_{i=1}^{w}$, where $w \leq n$. The text classification model is trained by a text training set with known corresponding category labels. This paper uses DNN neural network as a text classification model for classification training. The algorithm pseudo code is as follows: Define the input as text $Y$, and a certain text $X$ is preprocessed and feature extracted to obtain the feature vector $X'' = \{(x_i, w_i)\}_{i=1}^{w}$, as input node of DNN neural network. The output is the classification prediction set $C_Y$ made by the classification model for all text sets $Y$.

#### Text classification algorithm

| Input: Text $Y$ | Output: Classification prediction category $C_Y$ |
|----------------|-----------------------------------------------|
| step 1: $Y'' = \text{Pretreat}(Y)$; //Where $X'' = \{(x_i, w_i)\}_{i=1}^{w}$; |
| step 2: Calculate the mutual information set of each text in the text set $Y''$; |
| step 3: According to the control threshold $\beta$, a model input feature set $Y''$ is obtained; |
| step 4: $C_Y = \text{LSTM}(Y'')$; |

### 4 Experimental verification

The data in the experimental part of this paper comes from the data of the substation information system provided by the State Grid. According to the relevant requirements of the power grid, these data can be specifically divided into grid equipment maintenance operation tickets, information system maintenance schedules, information system maintenance work tickets, information system maintenance operation tickets, and customer service work tickets.

The total number of texts is 3000, with an average of 600 per category. 70% of each class is selected as the text training set for training the model, and the remaining 30% of each class is used as the test set to test the performance of the classification model. After training and testing, the results are as follows: The average rate can reach more than 91%.

#### Table 1. Experimental classification results.

| Grid text category                                      | True value | Properly classified | Actually classified | Accuracy |
|---------------------------------------------------------|------------|---------------------|---------------------|----------|
| Grid equipment maintenance operation ticket             | 600        | 548                 | 596                 | 91.4     |
| Information system maintenance plan                     | 600        | 535                 | 624                 | 89.2     |
| Information system maintenance work ticket              | 600        | 559                 | 578                 | 93.1     |
| Information system maintenance operation ticket         | 600        | 552                 | 586                 | 92.0     |
| Customer Service Ticket                                 | 600        | 534                 | 616                 | 89.7     |

### 5 Conclusion

This article is based on the need for the grid system to build a data and service standard
system covering the entire chain of energy and power business, expand the data model, and support the reality of unified management of all data. In order to retrieve and extract the massive electronic texts in the power grid system, this paper constructs a DNN neural network classification model to classify the grid text information. The validity of the method is verified by experiments based on the data of the substation information system provided by the State Grid.

This research was financially supported by the Self-financing for Guangdong Power Grid Co., Ltd. informatization project under Grants 037800HK42180056.

References

1. Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1):1-47.
2. Wijayanti R, Arisol A. Ensemble approach for sentiment polarity analysis in user-generated Indonesian text[C]. 2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA). IEEE, 2018.
3. X. Han, W. Shang and S. Feng, "The design and implementation of personalized news recommendation system," 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), Las Vegas, NV, 2015, pp. 551-554.
4. P. Long and V. Boonjing, "Longest Matching and Rule-based Techniques for Khmer Word Segmentation," 2018 10th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, 2018, pp. 80-83.
5. Salton G. A vector space model for automatic indexing [J]. Communications of the ACM, 1974, 18(11):613-620.
6. Y. Lin and J. Wang, "Research on text classification based on SVM-KNN," 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, 2014, pp. 842-844.
7. F. S. Nurfikri, M. S. Mubarok and Adiwijaya, "News Topic Classification Using Mutual Information and Bayesian Network," 2018 6th International Conference on Information and Communication Technology (ICOICT), Bandung, 2018, pp. 162-166.
8. S. Bahassine, A. Madani and M. Kissi, "An improved Chi-square feature selection for Arabic text classification using decision tree," 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), Mohammedia, 2016, pp. 1-5.
9. Kim Y. Convolutional neural networks for sentence classification [J]. Eprint Arxiv, 2014.
10. Tong Y, Zhang Y, Jiang Y. Study of sentiment classification for Chinese microblog based on recurrent neural network [J]. Chinese Journal of Electronics, 2016.25(4):601-607.
11. Z. Zong and C. Hong, "On Application of Natural Language Processing in Machine Translation," 2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCEE), Huhhot, 2018, pp. 506-510.
12. A. Barnard, "The Nursing Profession: Implications for AI and Natural Language Processing," 2007 International Conference on Natural Language Processing and Knowledge Engineering, Beijing, 2007, pp. 497-501.
13. Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786):504-507.