Automatic classification model of semi-structured HTML text data based on State Grid cloud architecture

Enjie Zhang*, Zhidong Zhang, Long Yan, Da Li
State Grid Tianjin Binhai Electric Power Supply Company, Tianjin, 300450
*Corresponding author’s e-mail: enjie.zhang@tj.sgcc.com.cn

ABSTRACT: Regarding the construction of the "State Grid Cloud" platform, various businesses of the power grid have their own Web systems. The data between different websites is scattered and the coupling between resources is low, which can easily form the problem of information islands. This paper is oriented to the semi-structured HTML text data in web pages under the State Grid cloud architecture platform, and uses the Python-based Scrapy framework to collect semi-structured power data information from various power business websites. We propose a semi-structured text data classification model based on BiGRU neural network and Bayesian classifier. BiGRU neural network is used to extract text features, Bayesian classifier is used for classification, and the TF-TDF algorithm is used to assign weights to improve the traditional recurrent neural network model with many parameters and long training model time. We use this method to simulate the semi-structured HTML text data of the State Grid, and conduct a comparative experiment with the traditional neural network model. The experimental results show that the classification algorithm can effectively improve the efficiency and accuracy of power semi-structured text data classification.

1. Introduction
With the development of energy internet and smart grid construction[1], the State Grid Corporation of China is actively promoting informatization and smart grid construction. At present, the State Grid cloud architecture is an important part of the intelligent construction of the power grid. One of the important parts of the web page is semi-structured HTML text data. The data types are many and messy, and the information is heterogeneous, which reduces the production management of power grid enterprises effectiveness, which can easily form information islands. Therefore, in view of the problem of information barriers in massive multi-source data under the current State Grid cloud architecture, the automatic classification technology of semi-structured HTML text data in various business web pages in the power grid operation intelligent platform is an important step to realize the commercial value of power grid data assets.

At present, some scholars have carried out research on the classification of text data. As a classic task in natural language processing, text classification mainly includes processes such as text preprocessing [2], text feature extraction and classifier design. Literature [3] uses Bayesian classifier to classify related topics of cultural tourism text. Through comparative experiments, it is found that it is very efficient in processing text classification, but if the amount of data is relatively large, the accuracy of this method is not high. Literature [4] uses the LSTM neural network model to classify financial news text, which improves the accuracy of the text. Literature [5] combines convolutional neural network and GRU neural network, and introduces the attention mechanism to extract text features. Although the feature extraction is better, the classification effect of short text is not good. Literature [6] uses bidirectional GRU neural...
network and attention for feature extraction, but the model complexity is high and the training time is long. Based on the above research, this paper studies the automatic classification technology of semi-structured HTML text data under the State Grid cloud architecture. Through an effective crawler technology, the data in the power cloud platform is collected, and a neural network based on BiGRU is proposed. Compared with traditional neural networks, the automatic classification method based on the BiGRU and Bayesian classifier can fully consider the relationship between text context information and word order, capture more important features in the data, and greatly optimize the process of feature extraction. At the same time, the Bayesian classifier has the characteristics of simple structure and high execution efficiency. The combination of the two improves the efficiency and accuracy of text classification.

2. Semi-structured HTML data capture in the State Grid Cloud Platform

The Python-based Scrapy framework is used to collect unstructured power data information from various power business websites. The basic data processing process of the Scrapy framework is shown in Fig. 1.

![Fig.1 Structure of Scrapy](image)

Based on the Scrapy framework to collect the semi-structured HTML information of the State Grid cloud platform, starting from the power URL web page address, the Scrapy framework engine will hand the web page address to the Scheduler module, and send the processed Request to the Downloader module for download, and the module gets the server The content of the returned webpage will be handed over to Spider for analysis. Spider extracts the data from the webpage. The information transfer Item, Pipeline module, and Pipeline module process and store the data to complete the crawling of power webpage data.

3. Semi-structured HTML text data preprocessing of power web pages

After crawling the data from the State Grid cloud platform, the MongoDb database is used to store the semi-structured HTML text data of power web pages. The MongoDb database is a database based on distributed file storage. It is currently one of the most practical NoSql databases and is a WEB application. Provide scalable high-performance data storage, greatly improving query efficiency.

Preprocessing semi-structured HTML text data includes the following steps:

1) A large amount of information in the State Grid cloud platform is almost concentrated in the information release of electric power enterprises, and the information classification is relatively clear, and the jieba word segmentation tool can be used to efficiently perform word segmentation.

2) Remove stop words such as "at", "have", "is" in the HTML text data of the power web page.

3) Use Word2vec to convert the power text data into word vectors.
4. Semi-structured HTML text data automatic classification model

A semi-structured text data classification model is established based on bidirectional GRU neural network and Bayesian classifier. The text features are extracted by BiGRU neural network, the Bayesian classifier is used to discriminate the classification, and the weight is assigned by the TF-IDF algorithm. This method improves the unidirectional deep learning model with low learning ability, and the complicated structure of the bidirectional deep model takes a longer time. The framework of this article’s automatic classification model is shown in Fig. 2.

4.1. Feature extraction of semi-structured power text data based on BiGRU

According to BiGRU computational logic programming to achieve feature extraction of semi-structured power text data, the calculation formula is as follows:

\[ z_t = \sigma(W_z \cdot [c_{t-1}, x_t]) \]  
\[ r_t = \sigma(W_r \cdot [c_{t-1}, x_t]) \]  
\[ h_t = \tanh(W \cdot [r_t \cdot c_{t-1}, x_t]) \]  
\[ c_t = (1 - z_t) \cdot c_{t-1} + z_t \cdot h_t \]

In the formula, \( \sigma \) is the sigmoid function, and \( W \) is the weight matrix, \( x_t \) is the input text data at time \( t \), \( c_{t-1} \) is the input text data at the previous time, \( r_t \) is the reset gate, \( z_t \) is the update gate, \( h_t \) is the state to be activated, and \( c_t \) is the current moment of output.

The update formula of BiGRU circulating the neural network layer from left to right is:

\[ \overset{\rightarrow}{h}_t = f(\overset{\rightarrow}{W} x_t + \overset{\rightarrow}{V} h_{t-1} + b) \]

The update formula for circulating the neural network layer from right to left is:

\[ \overset{\leftarrow}{h}_t = f(\overset{\leftarrow}{W} x_t + \overset{\leftarrow}{V} h_{t+1} + b) \]

The calculation formula of the final output of the BiGRU neural network is:

\[ y_t = g(U[\overset{\rightarrow}{h}_t; \overset{\leftarrow}{h}_t] + c) \]

In the formula, \( W, V, U \) are weight matrices, \( b, c \) are paranoid matrices, \( x_t \) represents the input text data at time \( t \), and \( h_t \) is the state to be activated.

4.2. Use the TF-IDF algorithm to weight the extracted power information features

TF-IDF [7] is a feature weighting algorithm that assigns weights to the power data features extracted by
BiGRU, and inputs the high-weighted Bayesian classifiers to reduce the data dimension and improve the efficiency of the algorithm.

TF represents the probability of a word appearing in the entire article. The calculation formula is:

\[ TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \]  

(8)

In the formula, \( n_{ij} \) is the number of times the power feature words appear in the semi-structured text, and \( \sum_k n_{kj} \) is the number of all feature words in the text. The result of the calculation is the word frequency of a certain power feature word.

IDF stands for inverse text frequency, which is used to measure the frequency of a word in the entire corpus. The calculation formula is:

\[ IDF_{ij} = \log \frac{N}{1 + N_{power}} \]  

(9)

\( N \) represents the number of all articles in the corpus, and \( 1 + N_{power} \) represents the number of words with obvious characteristics in the article. In order to prevent the word from not existing in the corpus, use \( 1 + N_{power} \) as the denominator. The main purpose of the TF-IDF algorithm is that if a word appears very frequently in the text of a certain power website[8], but appears very low in other texts, it means that the word has obvious characteristics. Then the word can be used as the basis for power information classification. The calculation formula is:

\[ TF - IDF_{ij} = TF_{ij} \cdot IDF_{ij} \]  

(10)

The value of TF-IDF is obtained by the product of term frequency \( TF_{ij} \) and reverse text term frequency \( IDF_{ij} \). The greater the value of TF-IDF, the greater the importance of the power feature word to the text.

4.3. Design a naive Bayes classifier

Text data in the electric power field generally have clear terminology specifications, and the problem of word ambiguity basically does not exist. Therefore, a naive Bayes classifier is adopted. The Naive Bayes algorithm is based on the assumption of attribute independence. For power semi-structured text classification, we assume that each feature word is independent of each other. Therefore, Naive Bayes can be used in text with \( n \) features, namely \( e_1, e_2, \ldots, e_n \). There are \( k \) text categories, \( m_1, m_2, \ldots, m_k \). The final result of the Bayesian classifier:

\[ p(m_i | e_1, e_2, \ldots, e_n) = \frac{p(e_1, e_2, \ldots, e_n | m_i)}{p(e_1, e_2, \ldots, e_n)} \]  

(11)

Find the probability that the text belongs to each category, and use the category with the highest probability as the final classification result.

After completing the text classification task, evaluating the model use three evaluation indicators: accuracy \( P \) (Precision), recall rate \( R \) (Recall) and \( F1 \).

The accuracy rate \( P \) (Precision) is the precision rate of the measurement category, and its formula is as follows:

\[ P = \frac{TP}{Sum} \]  

(12)

Where, \( TP \) represents the number of texts that are classified correctly; \( Sum \) represents the number of texts that are actually classified.

Recall rate \( R \) (Recall) is a measure of the recall rate of a category, and its formula is as follows:

\[ R = \frac{TP}{TS} \]  

(13)

Where, \( TS \) represents the number of texts belonging to the category.
The $F_1$ value is a measure of the combination of recall rate and precision rate, and the degree of bias, so it is calculated from the accuracy rate and the recall rate. The formula is as follows:

$$F_1 = \frac{(R \times P) \times 2}{(R + P)}$$  \hspace{1cm} (14)

5. Experimental simulation analysis

The experiment uses the Scrapy framework to collect the HTML web page data of various industries' electricity and user electricity consumption from a power grid company's metering automation system website. In the experiment, the parameters of each part under the optimal model effect are set as follows: The hidden layers is 3, the nodes in the hidden layer is 100, the learning rate is set to 0.05, the iteration steps for training is 100, and the iterations is 1000. In order to prevent the neural network from overfitting, the Dropout mechanism is adopted and the dropout rate is set to 0.3. At the same time, the model proposed in this paper is compared with the traditional LSTM model, GRU model, and BiGRU model, and each model parameter setting is the same as the model setting in this paper to verify the effectiveness of the model in this paper, and the test results are shown in Fig. 3.

![Fig. 3 Comparison of experimental results of various methods](image)

It can be seen from Fig. 3 that the method proposed in this paper has significantly improved Recall and $F_1$ value compared with the traditional unidirectional neural network. Compared with the bidirectional neural network, although the accuracy is almost the same, our method also has a clear advantage in the time of model training. This shows that the learning ability of the unidirectional deep learning model LSTM or GRU is not enough, resulting in low generalization ability. The bidirectional deep model has a complex structure and it takes longer to train the model, especially when the amount of data is relatively large, the training time of the model tends to be relatively long. Based on BiGRU, Our method uses the TF-IDF algorithm to assign weights to the power data features, and input the high-weight features into the Bayesian classifier to reduce the data dimension, which improves the efficiency of the algorithm, shortens the training time of the algorithm, and has a good classification ability.

6. Conclusion

Aiming at the semi-structured HTML text data in the web pages under the State Grid cloud architecture platform, the main work of this paper is to solve the problems of low accuracy and long training time of traditional neural network models in classification tasks. This paper uses the Python-based Scrapy...
framework to collect semi-structured power data information from various power business websites, and proposes a text classification algorithm combining BiGRU neural network model and Bayesian classifier, with the TF-IDF algorithm to assign weights to the power data features. Compared with the traditional methods, such as LSTM, GRU, and BiGRU, our method effectively extracts text features, shortens the training time, and improves the accuracy of classification.

Acknowledgments
This paper was supported by Science and Technology project of STATE GRID TIANJIN ELECTRIC POWER COMPANY: “Research on Electric Power Data Quality Improvement Technology and Lightweight Machine Learning Method Oriented to Edge Cloud Collaborative Architecture” (KJ21-1-26).

References
[1] Yang Ting, Zhai Feng, Zhao Yingjie. Explanation and Prospect of Ubiquitous Electric Power Internet of Thing[J]. Automation of Electric Systems, 2019, 43(13): 9-20.
[2] Li Yang, Dong Hongbin. Text Sentiment Analysis Based on CNN and BILSTM Feature Fusion [J]. Computer Applications, 2018, 38(11): 3075-3080.
[3] Wang Xiangxiang, Fang Hui, Chen Chongcheng. Research on Cultural Tourism Classification Technology Based on Naive Bayes[J]. Journal of Fuzhou University (Natural Science Edition), 2018, 46(5): 644-649.
[4] ZHANG Guowei, LYU Xueqiang, XIA Hongke. Financial news tendency based on LSTM[J]. Computer Engineering and Design, 2018, 39(11): 3462-3467.
[5] YANG Dong, WANG Yizhi. Text classification based on Attention based C-GRU neural network[J]. Computer and Modernization, 2018, 2(2): 96-100.
[6] LIU Yang. Time series prediction based on GRU neural network[D]. Chengdu: Chengdu University of Technology, 2017.
[7] He Junfei, Zhao Hui, He Xueming. Based on improved TF-IDF suspicious person text representation method [J]. Computer Engineering and Design, 2021, 42(02): 396-401.
[8] Zhao Ruijie, Shi Yong, Zhang Han, Long Jun, Xue Zhi. Webshell file detection based on TF-IDF [J]. Computer Science, 2020, 47(S2): 363-367.
[9] Yang Ting, Zhao Liyuan, Wang Chengshan. Review on Application of Artificial Intelligence in Power System and Integrated Energy System[J]. Automation of Electric Systems, 2019, 43(01): 2-14.