Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210463

Application of Text Mining Method for Classification of Work Order in Power Grid Production

Yadi ZHAO[1], Zhifeng WEI, Bingqiang GAO and Shuo ZHANG
State Grid Information & Telecommunication Accenture Information Technology Co., Ltd.-Beijing, China

Abstract. With the completion of the State Grid Corporation's maintenance system, the number of substations has increased dramatically, the grid structure has become increasingly complex, and there have been internal and external reasons such as the contingency of emergencies, and equipment failures have occurred from time to time. This paper aims to explore the potential value of massive data, show the laws of business data, and further give full play to the comprehensive support of data for enterprise operation and production management, and promote the realization of intelligent and lean power grid core business. This paper uses power system data to provide reliable data support for equipment defect full cycle management and equipment state analysis through ANOVA and neural network statistical analysis. At the same time, we use Term Frequency-Inverse Document Frequency(TF-IDF)Algorithm to calculate the importance of keywords and construct the power keyword library. By constructing Bayesian text classification model, we can classify the defect parts, defect categories and defect causes automatically. This method can be applied to the construction of power grid production work order text analysis system, improve the data quality and system automation level, help the business department to improve work efficiency and provide the basis for power grid business analysis. This method is applied to the data cleaning of the primary production equipment of power grid enterprises, and the accuracy of data error correction for equipment defects with voltages above 110kV is between 93% and 95%, and good results have been achieved.

Keywords. Power system data, statistical analysis, text mining

1. Introduction

Under the influence of the development of modern enterprises and the macro-economy, as the operators of the main power network business and the main trading body of the regional power market, Power grid enterprises have higher demand for the timely and effective enterprise management operation data[1]. With the continuous development of network technology and information technology, the digital degree of power grid enterprises is gradually improved, the application of equipment (asset) operation and maintenance lean management system (PMS2.0) is deepening, while the continuous increase of information system, the enterprise data is also growing greatly[2]. Extension

[1] Corresponding Author, Yadi ZHAO, State Grid Information & Telecommunication Accenture Information Technology Co., Ltd.-Beijing, China; E-mail: 15510266633@163.com

of digital information enables them to realise on an increased efficiency of the electrical grids, and further control the network dynamically to improve the security and reliability[3]. With the increasing increase of substation scale, the grid structure is increasingly complex, equipment failure occurs, the traditional equipment failure prediction lacks the systematic scientific support and effective early warning and decision-making mechanism. Meanwhile, most of the business data reporting of power grid enterprises in China still adopts manual method, which is easily affected by personnel quality and department management mode, that makes the data appear inconsistent or inaccurate in the process of creating, resulting in serious data crossover and redundancy[4].

How to better tap the potential value in these massive data, further give play to the comprehensive support of data for enterprise operation and production management, and promote the intelligent and lean realization of the core business of the power grid has become an important topic facing enterprises in the era of big data. State Grid Corporation proposes a new strategy for developing leading international energy Internet companies with Chinese characteristics, accelerates the construction of a new power systems. It requires greater use of corporate data value, big data intelligent analysis and multi-source big data governance, active enabling business management and external business services[5]. As an important foundation, big data method has been applied to many specialties in the field of electrician to improve data quality and realize reasonable operation of data[6-7]. Chen Yongjun analyzed the defect data of power network operation analysis decision support system and proposed a time series method for equipment defect prediction[8]. Tarik provides a text classification, which is a process of automatically assigning sets of documents into class labels depending on their data contents[9]. From the defect data of secondary equipment, Zhang Yanxu et al proposed a defect data mining method based on Apriori algorithm, which can effectively analyze the weak links of secondary equipment and provide reference for equipment control[10]. At the same time, we also used big data methods to carry out electricity recovery risk prediction and electricity potential sensitive customer prediction, which achieved relevant results[11-12].

For descriptive text data, text mining method can be used to organize and solidify the data based on the low efficiency and incomplete data[13-15]. This paper makes full use of the data of PMS2.0 and other systems, through text mining[16] and data analysis of equipment defects and other data to provide reliable data support for the whole cycle management of equipment defects and equipment status analysis, providing accurate and effective auxiliary decision-making suggestions and improving the level of lean transportation and accurate maintenance. Through this study, we have realized the correct prediction of the probability of defects of the main network equipment and treatment in advance, reduce the risk of power grid operation, ensure that the management personnel have more energy to study and judge the management process, making lean scientific management possible.

This paper aims to explore the potential value of massive data, show the laws of business data, and further give full play to the comprehensive support of data for enterprise operation and production management, and promote the realization of intelligent and lean power grid core business. The paper first introduces some theoretical methods, such as Pearson correlation analysis, analysis of variance, text mining, logistic regression algorithm and neural network algorithm. And then, we apply these theoretical methods for equipment fault prediction and equipment defect management. Finally, we make the conclusion and propose the content of future work.

2. Theoretical methods

2.1. Pearson correlation analysis

Correlation analysis is the process of describing the degree of close relationship between objective things with appropriate statistical indicators. The degree of correlation between the two variables is expressed by the correlation coefficient r. When two variables are positively correlated, the r values are between 0 and 1, indicating an increase as the other variable increases, and when two variables are negatively correlated, the r values are between-1 and 0, indicating that the other variable decreases as one variable increases. The closer the absolute value of r to 1, the stronger the association of the two variables and the weaker vice versa.

2.2. Analysis of variance

The ANOVA was investigated by analyzing the contribution magnitude of variants of different sources to the total variants, thus determining the magnitude of the influence of controlled factors on the study results. The rationale for ANOVA is that there are two basic sources of differences between means across treatment groups, and we used F values to infer whether each sample came from the same population.

- Experimental conditions: differences caused by different treatments are called intergroup differences, it is expressed as the sum of square of the deviation between the mean in eac group.(Msb)
- Random error: differences caused by measurement error or differences between individuals, called within-group differences, are expressed as the sum of the square sum of the deviation of the variable mean in each group from the values of the variables within that group.(Msw)
- The MSb/MSw ratio constitutes the F distribution.

2.3. Text MINING

Text Mining refers to the classification of a text into a known text category, mainly including text preprocessing and classifier model construction. The text classification process is shown in Figure 1.

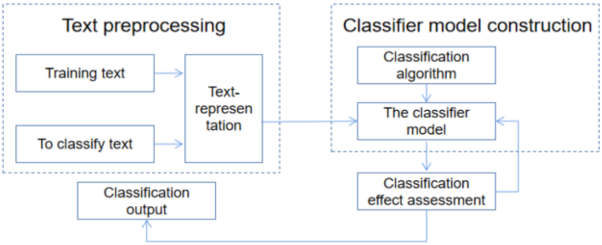

Figure 1. Text classification process

Text preprocessing refers to through Chinese text segmentation, text representation, text feature selection, etc procedures, change the text data table that expresses the natural language into the machine language data that the computer can handle. Text representations are often represented by a vector space model: $d_i = (t_{i1} : w_{i1}, t_{i2} : w_{i2}, \cdots, t_{in} : w_{in})$, d_i represents the text, t_{ij} represents the word j in the i text, w_{ij} represents the weight of the t_{ij} in the text d_i .The classifier model construction process is to generate the classifier model based on the training set data after text pretreatment through relevant algorithms, and use the classifier model to automatically classify the new text data.

In this paper, text preprocessing is implemented by jieba word segmentation and the term frequency-inverse document frequency method(TF-IDF), the classifier construction takes a Bayesian classification algorithm.

2.3.1. Text Segmentation Algorithm

Chinese word segmentation can be roughly divided into two categories: dictionary-based word segmentation and statistics-based word segmentation methods. The word segmentation method used in this paper is jieba word segmentation, which is a combination of the two before and after, that is, based on the prefix dictionary to scan the word graph to form a directed acyclic graph of all possible word segmentation results, and dynamic programming to find the path of maximum probability, thus solving the problem of double understanding word combination.

In statistics-based word segmentation method, the statistical sample content comes from some standard corpora. If there is a sentence M ,it has m word segmentation options:

$$\{A_{11}A_{12}\cdots A_{1n_1}, A_{21}A_{22}\cdots A_{2n_2}, \ldots, A_{m1}A_{m2}\cdots A_{mn_m}\} \tag{1}$$

where the subscript n_i represents the number of words in the i th participle. If we choose the best segmentation method r ,the statistical distribution probability corresponding to this word segmentation method should be the largest, which is: $r = \arg\max_i P(A_{i1}, A_{i2}, \ldots, A_{in_i})$. In order to simplify calculations, Markov hypothesis is usually used, that is, assuming that the probability of each participle is only related to the previous participle, it can be written as:

$$P\left(A_{ij} \middle| A_{i1}, A_{i2}, \ldots, A_{i(j-1)}\right) = P\left(A_{ij} \middle| A_{i(j-1)}\right) \tag{2}$$

Then the joint distribution can be obtained as:

$$P(A_{i1}, A_{i2}, \ldots, A_{in_i}) = P(A_{i1})P(A_{i2}|A_{i1})P(A_{i3}|A_{i2})..P\left(A_{in_i} \middle| A_{i(n_i-1)}\right) \tag{3}$$

Through the standard corpus, the binary conditional probability between any two word segments can be calculated approximately. For example, any two words a_1, a_2 , the conditional probability distribution can be approximately expressed as:

$$\begin{aligned} P(a_2|a_1) &= \frac{P(a_1, a_2)}{P(a_1)} \approx \frac{\text{freq}(a_1, a_2)}{\text{freq}(a_1)} \\ P(a_1|a_2) &= \frac{P(a_2, a_1)}{P(a_2)} \approx \frac{\text{freq}(a_1, a_2)}{\text{freq}(a_2)} \end{aligned} \tag{4}$$

where $\text{freq}(a_1, a_2)$ indicates the number of times two words appear next to each other in the corpus, $\text{freq}(a_1)$、$\text{freq}(a_2)$ respectively indicates the statistical times that each word appears in the corpus. Then for a new sentence, using the corpus to establish statistical probability, the word segmentation method corresponding to the maximum probability can be found by calculating the joint distribution probability corresponding to various word segmentation methods, which is the optimal word segmentation.

2.3.2. TF-IDF algorithm

Using Jieba word segmentation can divide the document, but not every word is meaningful (the word contributes less to the content of the document.), so we use TF-IDF method to judge the importance of words to the document.

The main idea of TF-IDF is: if a word or phrase appears in a certain category with a high frequency of TF and rarely appears in other categories, it is considered that this word or phrase has good classification ability, suitable for classification, and can be marked as a keyword glossary. Suppose that d is the classification category, f is a feature word based on the content of the ticket, $TF_{f,d}$ represents the number of times a given word appears in the category.

Document frequency is represented by DF, it represents the number of all categories containing a vocabulary in a sample set, it is mapped to a smaller value range and expressed as inverse document frequency (IDF) as follows:

$$IDF_f = \log(N/(DF_f + 1)) \tag{5}$$

in which N represents the total number of categories in the sample collection, DF_f represents the number of categories that contain the vocabulary f .The significance is that if the fewer categories include the word, the larger the IDF, which means that the vocabulary has a good classification ability. The weight of the term is represented by $TF-IDF$, and the calculation formula is as follows:

$$TF-IDF = (TF_{f,d}) * (IDF_f) \tag{6}$$

This method can calculate the importance of a certain vocabulary in a certain classification, identify whether the vocabulary is a keyword vocabulary, realize the construction of classification rules, and accurately locate the basic content of the text description, so as to provide a basis for subsequent classification analysis.

2.3.3. Bayesian Classification Algorithm

In the naive Bayes classifier, it is determined based on each feature that the label should be assigned to the given input value. Besides, the Naive Bayes classifier determines the label prior probability by calculating the frequency of each label on the training set, and the contribution of each feature is combined with its prior probability to obtain the likelihood estimate of each label. The label with the highest likelihood estimate will be assigned to the input value. Let each data sample describe the value of the n properties with a n-dimensional eigenvector $X = \{x_1, x_2, \dots x_n\}$, we assume m categories which are represented by $\{c_1, c_2, \dots c_m\}$, respectively. Given an unknown data sample X (there is no class number), if we assign an unknown sample X to the class c_i, it must be $P(c_i|X) > P(c_j|X), 1 \le j \le m, j \ne i$. Basis Bayes Theorem, because of $P(X)$ is the known constant, the maxizing posterior probability $P(c_i|X)$ can be converted to maximizing prior probability $P(X|c_i)P(c_i)$. Assuming that the values of each property are independent of each other, therefore the prior probabilities $P(x_1|c_i), P(x_2|c_i), \dots P(x_n|c_i)$

can be obtained from the training dataset. The classification principle is shown in Figure 2.

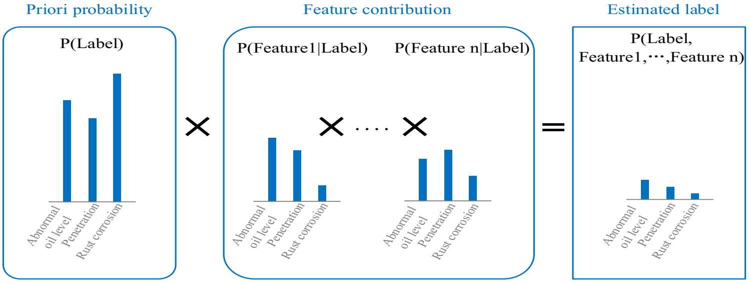

Figure 2. Classification principle description

Suppose that $label$ is the output label, $features$ represent the n eigenvalues of the input. The calculation process of the algorithm is as follows:

$$\text{Calculate } P(features) = \sum\nolimits_{label \in labels} P(features, label) \tag{7}$$

The likelihood label can be expanded as the probability of the label multiplied by the probability of a given label feature, in the case of feature independence, we have:

$$\begin{aligned} P(features, label) &= P(label) \times P(features \mid label) \\ &= P(label) \times \prod\nolimits_{f \in features} P(f \mid label) \end{aligned} \tag{8}$$

where $P(label)$ is the prior probability of a given label, each $P(f \mid label)$ is the contribution of a single feature to the likelihood of the label, it can be calculated as:

$$P(f \mid label) = count(f, label) / count(label). \tag{9}$$

Calculate $P(label \mid features) = P(features, label) / P(features)$, select the label corresponding to the maximum probability value as the label result for the new input value.

When the training set has features that never appear with a given label, the $P(f \mid label)$ calculated value is 0, which would cause the input to never be assigned to this label, resulting in a reduced classification accuracy. To avoid such situations, we typically apply the "Laplace corrections" for "smooth" processing when estimating probability values. Let N represent the number of possible categories in the training set D, N_i represents the possible number of values for the i'th attribute, D_c represents a collection of class c samples composed in the training set D, $D_{c_i x_i}$ represents a collection of samples with value x_i on the i'th attribute in D_c, the correction calculation is as follows:

$$\begin{aligned} count(label) &= \frac{|D_c| + 1}{|D| + N} \\ count(f, label) &= \frac{|D_{c_i x_i}| + 1}{|D_c| + N_i} \end{aligned} \tag{10}$$

2.4. Logistic Regression Algorithm

Logistic regression is a study of dichotomous variable Y and a series of influencing factors X_n ,the multivariate analysis method of the relationship is further developed on the basis of the linear model.Its general form is: $p=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n$, $P=\frac{1}{1+e^{-p}}$. In the formula, P is a probability occurrence of the variable Y , between 0-1.Logistic regression model has fast calculation speed, obvious results and good fitting effect. It is widely used in big data, machine learning, economics and other fields.

2.5. Neural network algorithm

The BP neural network is a multi-layer feedforward network trained for propagation from the direction of error. The BP neural network maps the input variable to the output variable via the excitation function and the constant weight threshold for adjusting each layer connection. To align the output variables of the network with the expectation, the learning of BP neural networks generally requires repeated training multiple times so that the error value tends to zero and eventually reaches zero.

The implementation process of the BP neural network algorithm is as follows: The input matrix of the sample is assumed to be $X=(x_{ij}), i=1,2,...,n, j=1,2,...,p$. Each row of data represents a set of input samples, and each set corresponds to a set of output samples, and then the actual output samples corresponding to all input samples are $Y=[y_1,y_2,...,y_n]^T$. We regard each column as an indicator of the sample, and then the input sample I_1 of the input layer is X.

If the implied layer of the network contains m neurons, weights matrix $W=(w_{ij}), i=1,2,...,m, j=1,2,...,p$, threshold matrix $B=[b_1,...,b_m]^T$,then the input of the implied layer is:

$$I_2=W_{m\times p}X'_{p\times n}+Bones_{1\times n} \quad [17] \tag{11}$$

In which $ones_{1\times n}$ represents a matrix with all elements of 1. The expression of the implied layer incentive function is:

$$f(x)=\frac{1}{1+e^{-x}} \tag{12}$$

Thus we can get the output of the implied layer as $O_2=f(I_2)$, the input for the output layer is $I_3=X_{jk}O_2+B_{jk}+ones_{1\times n}$. Since the transfer function is a class of linear functions, the $O_3=I_3$ can be considered for the output of the output layer.

The initial weight threshold selected by the BP algorithm can determine the weight threshold from the second to the n adjustment weight threshold, so the accuracy of the BP network training is determined by the selection of the initial weight threshold. If the initial weight threshold is inappropriate, it is likely to make the network slow convergence and easily fall into the local optimal solution, so the selection of the initial power threshold has a great relationship to the quality of the BP network training results.

Genetic Algorithm (GA) is a heuristic global search algorithm (Three-dimensional coordinate measurement algorithm by optimizing BP neural network based on GA) inspired by the evolutionary processes of living organisms. Genetic algorithm simulates the problems to be solved into a process of biological evolution, generating the next generation of solutions through selection, crossover and variation. By calculating the fitness of individuals, the individuals with low fitness are eliminated, the individuals with high fitness are increased, and the optimal fitness function value solution is found after n iterations. The characteristics of genetic algorithms can compensate for the defect that BP neural networks can easily fall into local optimal solutions, so usually we use genetic algorithms to optimize BP neural networks to build correlation models.

3. Scene application

3.1. Equipment fault prediction

We use Pearson correlation analysis, single-factor variance analysis and other methods to conduct correlation analysis of various types of data and determine the main factors affecting the equipment failure. Then, we establish a Logistic regression fault prediction identification model to achieve an advance warning of the device state. Finally, we establish a classified neural network model to judge the possible fault types of the early warning equipment, supporting the differentiated operation and maintenance of the substation equipment.

We get the data from SCADA system, PMS2.0 system, intelligent inspection robot system, equipment online monitoring system, weather forecast system, etc. We take data from a prefecture-level company from January 2019 to August 2020, obtaining data from five dimensions, including operating conditions, equipment attributes, system mode, external environment, and seasonal characteristics, with a total of 135 fields and 179311 data. The specific data sources are listed in the Table 1.

Table 1. Data required for equipment failure prediction

Data category	Data source	Data field
operating conditions	SCADA system	Equipment current, Voltage, Oil temperature, Switch knife gate position, Load rate, Load value,etc.(Class 34 data)
equipment attributes	PMS2.0 system	Equipment name, Equipment coding, Equipment type, Equipment model, Manufacturer, Leave the factory date, Date of delivery, Voltage grade,etc.(Class 39 data)
system mode	intelligent inspection robot system	Appearance of equipment, Equipment separate/merged status, Electric energy meter measure instruction, Infrared temperature measurement, Noise,etc.(Class 28 data)
external environment	equipment online monitoring system	Transformer, Online monitoring system for the lightning arrester, Online temperature and humidity monitoring system, Infrared imaging online analysis system,etc.(Class 22 data)
seasonal characteristics	weather forecast system	Regional ID, Maximum forecast temperature, actual highest temperature, Minimum forecast temperature, Actual minimum temperature, Temperature and humidity, Thunderstorm, Hail,etc.(Class 12 data)

Equipment fault influencing data can be divided into continuous, classification and unstructured types.According to different data types, Pearson correlation analysis and single-factor variance analysis etc. are used to analyze the fault influencing factors.

The operating influence factors of equipment failure mainly extracted "the average monthly load rate" as the characteristic variables. We calculate the correlation coefficient value of 0.9317 through the correlation analysis, so there is a strong positive correlation between the device failure rate and the load rate. At the same time, we analyzed a strong positive correlation between the equipment failure rate and the operation life, temperature and humidity, with the correlation coefficient values of 0.7484,0.617 and 0.602, respectively. The experimental results are shown in Figure 3, Figure 4 and Figure 5.

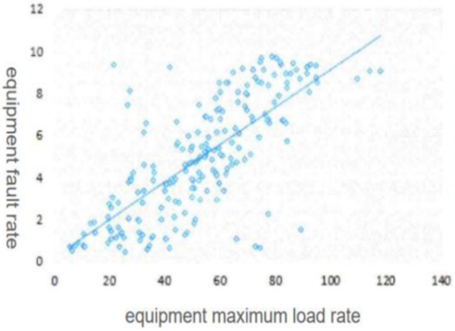

Figure 3. Monthly average load rate and fault rate scatter plot

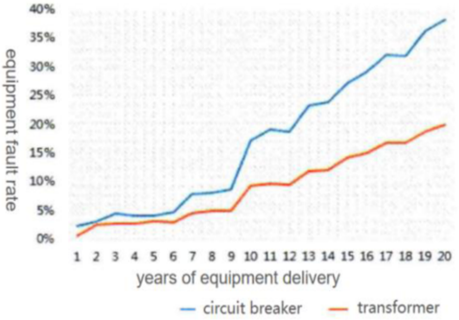

Figure 4. Correlation between equipment years of operation and failure rate

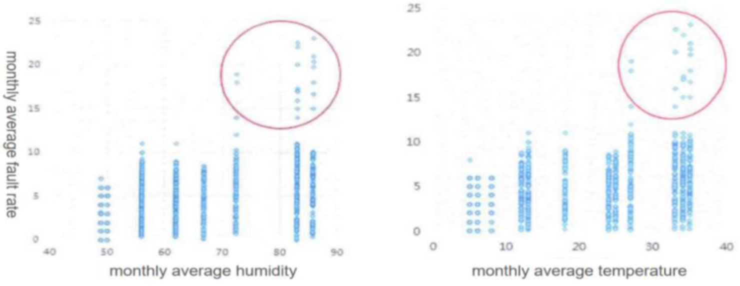

Figure 5. Relationship between equipment failure rate and temperature/humidity

We perform a variance analysis for the "region where the device belongs" variables. According to the variance analysis table, calculating P=0.0387<0.05, we can assume that the "region where the device belongs" has a significant impact on the equipment failure rate. Using the same method, we analyzed that the equipment model number has a significant influence on the equipment failure rate as shown in Table 2.

Table 2. ANOVA of the area where the equipment belongs

Error source	Quadratic sum	Free degree	Mean square error	F-value	P-value	F-critical value
Intergroup	5826.434784	12	485.536	4.0938	0.03876	3.9705
Within the group	20497992.24	143819	142.526	-	-	-
Total	20503818.68	143831	-	-	-	-

Based on the "time (month)", "manufacturer", "precipitation", "thunderstorm", "gale" and other classification or unstructural factors, we use the statistical comparison method, the examples are as follows. It can be seen in Figure 6 and Figure 7 that different manufacturers and weather conditions will affect the improvement of equipment failure rate.

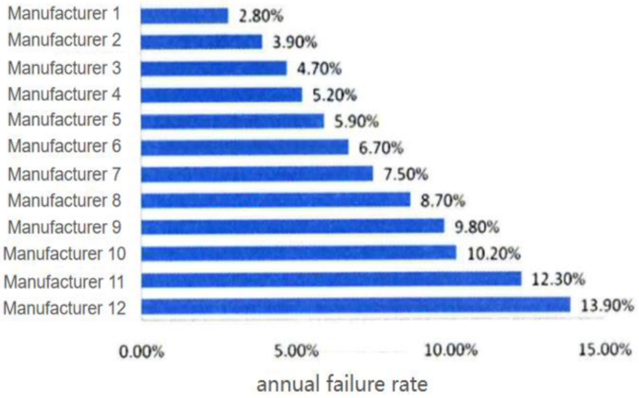

Figure 6. Equipment failure rate of different equipment manufacturers

We included 143,832 records for full year 2019 as training samples and the first quarter 2020 as test data. We used the Logistic regression model to obtain the training results with different probabilities. From the Table 3, it works best when the classification critical probability is 0.5.

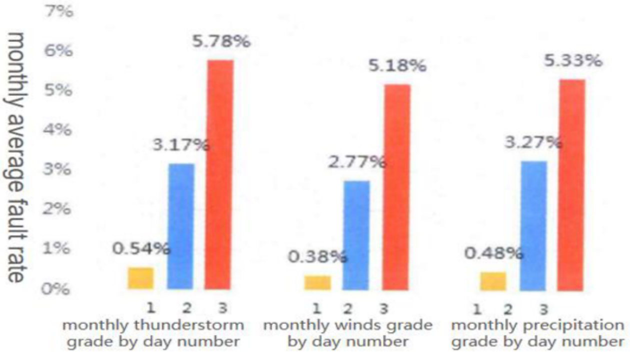

Figure 7. The relationship between substationfailure and thunderstorm gale, precipitation, etc.

Table 3. Training results at different probabilities p

Classification critical probability-p	**0.3**	**0.4**	**0.45**	**0.5**	**0.55**	**0.6**	**0.7**	**0.8**	**0.9**
Identification accuracy	77.86%	82.39%	90.11%	94.23%	89.55%	82.17%	79.40%	77.22 %	69.93%
Fault shooting rate	64.31%	73.42%	76.36%	82.53%	84.21%	85.39%	87.28%	88.01 %	87.37%

Based on the regression model constructed above, the samples from April-May 2020 were predicted, as shown in the Table 4. The total prediction accuracy was approximately 90.45% with high accuracy.

Table 4. Analysis of the fault prediction error of the substation

Forecast month	Actual number of no failure has occurred	Actual number of failure has occurred	0-0	0-1	1-0	1-1	Predictive accuracy
April, 2020	9813	549	9054	759	52	497	92.17%
May, 2020	10519	609	9351	1168	96	513	88.64%

We performed fault identification through a neural network method with an identification accuracy of 93.16% and the results shown in the Table 5. At the same time, we can calculate that variables such as load rate, operation years of equipment, ttemperature and humidity, equipment model play an important role in the prediction process.

Table 5. Identify results

Transformer No.	Forecast Fault	Actual Failure	Whether identification is accurate
Transformer 1	The oil temperature is too high	The oil temperature is too high	Yes
Transformer 2	overload	overload	Yes
Transformer 3	The gas content of the oil exceeds the standard	The gas content of the oil exceeds the standard	Yes
Transformer 4	The oil level is abnormal	The oil level is abnormal	Yes
...	...	...	...

In internal management, we often use the influencing factors of equipment failure prediction to carry out equipment defects management, and the purpose is to reduce the time cost of labor input and increase work efficiency through intelligent means.

3.2. Equipment defect management

The main classification algorithm is used in the construction of the power grid production work order text analysis system, and its functional framework is shown in Figure 8. The keyword library management module extracts production keywords based on defect grading standards, realizes the original keyword mining function, and provides keyword query at the same time. The basic data classification standard library and the original work order library are based on the keyword library, and the classification rules are composed of keywords, and training is performed based on the rules and tags, forming an automatic classification after the work order is entered. The classification knowledge base and statistical analysis report function are based on the keyword extraction and classification functions to establish classification rules and various production business relationships to support more business applications. The classified work orders are used as basic data according to different businesses. Perform report statistics to realize different functional applications based on grid business.

As one of the important contents of power equipment operation and maintenance, defect management has always been a work of great concern for power enterprise managers and production and maintenance personnel. In actual work, there are irregularities in defect management, especially the non-standard description and handling of defects, which affect the level of defect management. The description content of the equipment defect is the closest to the site situation and can reflect the specific situation of the site. Therefore, the word segmentation processing of the defect description text can obtain the keywords of the specific situation of the site. Perform word frequency analysis based on the processed keywords and classification criteria, use the TF-IDF algorithm to calculate the frequency of words in this category and the frequency of words in other categories, generate work order keywords, and build a keyword library. The principle is shown in Figure 9. Based on the keyword database, a sample data set of classification rules is established, and the classification rules are obtained by means of machine learning and Bayesian model training.

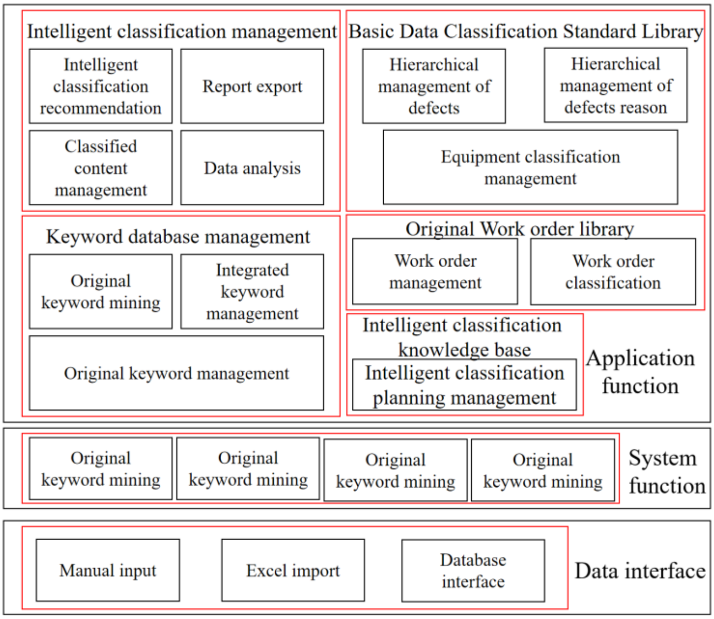
Intelligent classification management
Intelligent classification recommendation
Report export
Classified content management
Data analysis
Basic Data Classification Standard Library
Hierarchical management of defects
Hierarchical management of defects reason
Equipment classification management
Keyword database management
Original keyword mining
Integrated keyword management
Original keyword management
Original Work order library
Work order management
Work order classification
Intelligent classification knowledge base
Intelligent classification planning management
Application function
Original keyword mining
Original keyword mining
Original keyword mining
Original keyword mining
System function
Manual input
Excel import
Database interface
Data interface

Figure 8. The functional framework of the text analysis system for power grid production work orders

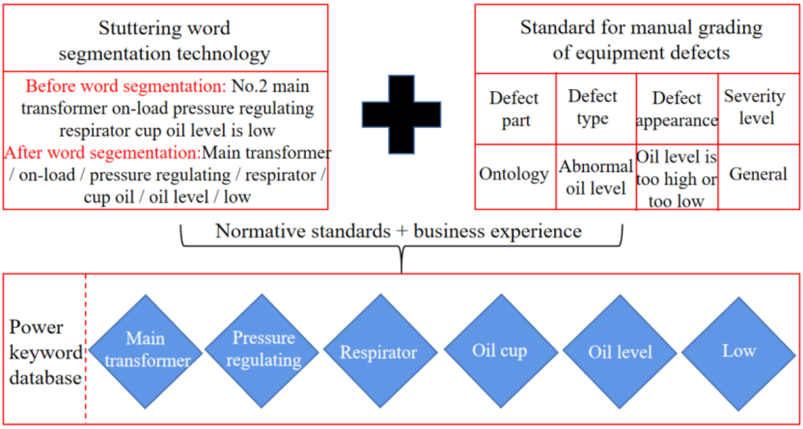
Stuttering word segmentation technology
Before word segmentation: No.2 main transformer on-load pressure regulating respirator cup oil level is low
After word segementation:Main transformer / on-load / pressure regulating / respirator / cup oil / oil level / low
Standard for manual grading of equipment defects
Defect part
Defect type
Defect appearance
Severity level
Ontology
Abnormal oil level
Oil level is too high or too low
General
Normative standards + business experience
Power keyword database
Main transformer
Pressure regulating
Respirator
Oil cup
Oil level
Low

Figure 9. Keyword library construction principle

Based on the defect data, the defect location, defect category, and defect cause are classified respectively, where the defect locations are divided into first, second, and third levels, the defect causes are divided into first and second levels, and algorithms can be used for iterative calculations. Figure 10 shows an example of the whole process of automatic classification. The defect type is used as the target classification content. The input value is "The oil level of the No. 2 main transformer on-load pressure regulator respirator is low", and the output type value is "Abnormal oil level".

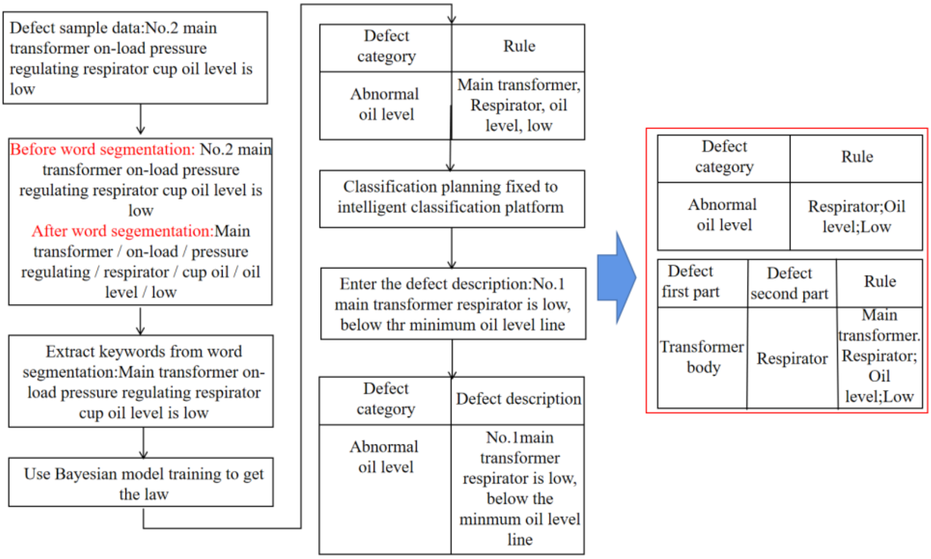
Defect sample data:No.2 main transformer on-load pressure regulating respirator cup oil level is low
Before word segmentation: No.2 main transformer on-load pressure regulating respirator cup oil level is low
After word segementation:Main transformer / on-load / pressure regulating / respirator / cup oil / oil level / low
Extract keywords from word segmentation:Main transformer on-load pressure regulating respirator cup oil level is low
Use Bayesian model training to get the law
Defect category
Rule
Abnormal oil level
Main transformer, Respirator, oil level, low
Classification planning fixed to intelligent classification platform
Enter the defect description:No.1 main transformer respirator is low, below thr minimum oil level line
Defect category
Defect description
Abnormal oil level
No.1main transformer respirator is low, below the minmum oil level line
Defect category
Rule
Abnormal oil level
Respirator;Oil level;Low
Defect first part
Defect second part
Rule
Transformer body
Respirator
Main transformer. Respirator; Oil level;Low

Figure 10. Example of the whole process of classification

4. Conclusion

The power grid production work order text analysis system applies text word segmentation, keyword mining, and automatic classification technology to each module to solidify the keyword database and cleaning rules for cleaning the production data of the primary equipment of the power grid. This method was applied to the data cleaning of primary production equipment for a power grid enterprise in the past seven years, and the total number of defects of equipment with a voltage of 110kV or above was 36255. The classification results are shown in Table 6, where there are 21,220 pieces of error correction data, accounting for 58.52%; 31,857 pieces of error correction data, accounting for 87.87%; data correction accuracy rates are between 93%-95%, achieving good results.

Table 6. Intelligent classification effect

	Number before correction/piece	Accuracy	Number after correction/piece	Accuracy
Defects	4398	12.13%	31857	93.53%
Defect category	15035	41.48%	21220	94.61%

The text analysis system of power grid production work orders ranges from a general data quality management platform that can only find problems to the system from discovery to problem solving. It combines classification rules and standard formulation to provide data problem solutions.

At present, the power industry has not established a standardized power word database, which is in its infancy. Therefore, the establishment of this system provides a basis and development direction for grid companies to establish a set of standardized power word databases in the future. On the basis of realizing the automatic data cleaning function, the system has improved data quality and clear efficiency. It has functions such as intelligent classification of work orders, construction and optimization of keyword databases, and statistical analysis assistance. It can be used for system application construction departments, equipment management departments and various business research departments.

In the future, it is planned to apply more algorithms to the system to optimize the functions of each module, and at the same time increase the business expansion of the application part, such as increasing the research on the distribution network data, the intelligent classification of equipment accident event data, etc., to support more and higher power grid business analysis to promote the high-quality development of ubiquitous power Internet of Things business.

References

[1] Bird, S. & Klein, E. & Loper,E. Python natural language processing. POSTS & TELECOM PRESS, 2009.

[2] Chen, Y. J. Prediction of equipment defects in Power System. Zhejiang University, Hang Zhou, Zhejiang Province, China, 2003.

[3] Kaneriya, S., Tanwar, S., Nayyar, A., Verma, J. P., Tyagi, S., Kumar, N., ... & Rodrigues, J. J.. Data consumption-aware load forecasting scheme for smart grid systems. IEEE Globecom Workshops (GC Wkshps) , 2018, 1-6.

[4] Dang, F. F. Research on business data quality control technology of Power Grid Enterprises. North China Electric Power University, Beijing, China, 2014.

[5] Du, X. M. & Qin, J. F. & Guo, S. Y. Text Mining of typical fault cases of power equipment. High Voltage Engineering, 2018, 44(4): 1078-1084.

[6] Ju, X. L. Power marketing analysis system based on data warehouse technology.Digital Technology and Application, 2012(3): 68-69.

[7] Li, Y. H. & Wang, J. X. & Wang, X. L. Power system network loss evaluation method based on hybrid cluster analysis.Automation of Electric Power Systems, 2016(1): 60-65.

[8] Ma, R. & Zhou, X. & Peng, Z. Data Mining of correlation chatacteristics of load characteristics statistical indexes considering temperature factors .Proceedings of The Chinese Society for Electrical Engineering, 2015(1).

[9] Tarik A. Rashid , Arazo M. Mustafa and Ari M. Saeed. A Robust Categorization System for Kurdish Sorani Text Documents. Information technology Journal, 2017(1): 27-34.

[10] Wang, X. Research on Power Internet of things Architecture for Smart Grid Construction . Power Systems and Big Data, 2018, 21(10): 34-37.

[11] Zhao Y. D., Wu Z., Chen X. F. etal. Application of the big data method for electricity bill recovery risk prediction.Telecommunication Science, 2019, 35(02):125-133.

[12] Chen X. F., Zhao Y.D., Zhang L. P. etal. Application of big data methods for power potential sensitive customers.Telecommunication Science, 2020,35(11):117-124.

[13] Wu, G. Y. & Zhang, Q. B. & Wu, H. C. Text Mining Analysis of power customer complaints based on natural language processing technology.Power Systems and Big Data, 2018(10).

[14] Zhang, Y. X. & Hu, C. C. & Hang, S. Data Mining and Analysis method of secondary equipment defects based on Apriori Algorithm. Automation of Electric Power Systems, 2017(19).

[15] Zou, B. P. & Zhou, L. Research on data service of Power Grid Enterprises. Electric Power Information and Communication Technology, 2010(9): 30-32.

[16] Zou, Y. F. & He, W. M. & Zhao, H. Y. Application of Text Mining Technology in power work order data analysis . Modern Electronics Technique, 2016,39(17): 149-152.

[17] Hang, L. Statistical learning methods. TSINGHUA PRESS, 2012.