New Energy Power Prediction Optimization Based on Improved TF-IDF Single Machine Information Feature Extraction

Liu Meng¹, Liu Ranjie¹, REN Yidan²,³, Qi Yue¹, BAI Jingjie²,³, Wang Manshuai⁴
¹State Grid Jibei Electric Power Company Limited, Beijing, 100053, China
²Nari Group Corporation/State Grid Electric Power Research Institute
³BeiJing Kedong Electric Power Control System Co., Ltd, Beijing 100192, China
⁴State Grid Jibei Electric Power Co., Ltd. Zhangjiakou Power Supply Company,Hebei,075000,China
*524869920@qq.com

Abstract—In order to lean management of new energy consumption, based on the characteristics of wide range and complex structure of unstructured data including dispatch and maintenance logs in the power grid system, based on the idea of improving TF-IDF, a DF-IDF data reflecting time slots is proposed Feature extraction algorithm, taking wind power in a power grid in the northern region as an example, extracts keyword and part-of-speech features from unstructured power grid dispatch logs and maintenance data to achieve single-machine state data screening and eliminate irrelevant state data. With the help of the effectiveness of the stand-alone information feature extraction algorithm, the management methods of new energy power prediction optimization are then studied. The results show that in the case of highly complex data, the average accuracy of the DF-IDF-processed feature extraction method is 8% and 7% higher than the traditional TF-IDF method and the One-Hot method. At the same time, it also provides pre-processing methods for stand-alone power prediction calculation, accuracy analysis, and summary station power prediction optimization in new energy power prediction and optimization, improving the rationality, accuracy, and effectiveness of data collaborative processing, and laying new energy lean. Contain the foundation of management level.

1. INTRODUCTION
The wind power grid is one of the important features of the new energy power generation. 80% of wind energy is concentrated in the “three north” regions (northeastern China, northern China and northwestern China) [1]. The climate is volatile, and terrain of the region is complex, so its power generation’s volatility and uncontrollability harm the power grid. To resolve the instability, strengthening wind power-forecasting technology is necessary. This is because this technology can enhance the safety and stability of the power system and improve consumption, the economic benefits and utilization rate of wind farm stations, which can provide a reliable basis for the grid dispatching plan [2]. We have related data access processing and new energy power prediction and optimization to achieve lean management of new energy consumption. Due to the complexity of the geographic location and weather information of wind farm
stations, we also need to conduct in-depth research on the accuracy of power forecasting, including the improvement and optimization of short-term power forecasting accuracy based on the maintenance plan.

Although many countries have been rapidly developing in new energy power generation, such as power forecasting methods, its accuracy of new energy wind power is still difficult to meet the standard indicators. In addition, the short-term power forecasting accuracy of wind farms is low, leading to a decrease in grid-connected stability, wind energy utilization of wind farms, and lean consumption analysis. Non-standard management of the maintenance plan for new energy wind turbines also damages the accuracy of new energy power forecasting, and increases the differences in the maintenance plan among new energy stations, and decreases the accuracy of red and green forecasting. Therefore, these factors challenged the safe operation of the grid connection. Taking full advantage of the accurate results of power prediction is significant to orientate the actual service production. Nowadays, we use traditional system acquisition to improve the power prediction of stations and regional power grids, so as to reflect the general status of on-grid and off-grid. Consequently, these do not have detailed equipment status splits and cannot reflect the advantages of lean management.

In the study of power prediction optimization, literature [3] designed a mechanism to improve wind farms' power prediction accuracy. Further, it uses the interval number method to quantify the increased system operating costs for the consumption of new energy. Finally, it designed a fair and non-linear allocation mechanism between wind farms, improving the accuracy of the forecast output reported. Literature [4] established a dual-objective optimization model with the smallest daily system cost and the smallest daily exhaust emissions by studying the optimization of power systems containing photovoltaic power plants. This model improves the BP neural network algorithm, which can effectively improve the power prediction accuracy. Based on the grid dimension, literature [5] used fuzzy cluster analysis, and established a regional wind power forecasting technology framework, cluster forecasting and traditional forecasting methods. Fully considering the characteristics of wind power output and prediction model, literature [6] proposed a segmented support vector machine and neural network prediction algorithm. Moreover, according to the influence of weather characteristics on photovoltaic output, literature [6] proposed a photovoltaic output prediction algorithm based on meteorological characteristics analysis to support the reliability analysis of power prediction. Literature [7] believes that the comprehensive application of multiple refined prediction techniques can significantly improve the accuracy of wind field power prediction. Literature [8] proposed improving the probability prediction model's effectiveness under the context of regional prediction through studying the probability prediction. Literature [9] used feature engineering technology to extract the in-depth features of photovoltaic power plant output in meteorological parameters, building a photovoltaic output prediction model based on fine-grained features and improve the accuracy of power prediction.

To sum up, most of the existing research tends to improve the accuracy of new energy power prediction based on the dimensions of the station and the regional power grid through technical modeling. However, considering the lean of new energy, we lack research on optimizing the prediction of new energy power in stations or regional power grids. At present, we cannot provide lean manufacturing, resulting in insufficient space for new energy consumption and power curtailment.

This paper focuses on wind power resources in northern China and clarifies the research on optimizing new energy power prediction. For example, we use artificial intelligence self-learning methods to improve data features and extraction methods and extract stand-alone state information from massive scheduling technology and maintenance logs. Then, based on the improvement of the stand-alone unit's data quality, the state of the stand-alone wind power is split to eliminate the abnormal state of the stand-alone, ensuring the value of the stand-alone power prediction during the calculation. We have also provided a legal basis for collecting and optimizing the power forecasting of stations and regional power grids.
2. OVERVIEW OF NEW ENERGY POWER PREDICTION AND OPTIMIZATION RESEARCH BASED ON THE IMPROVEMENT IN SINGLE MACHINE INFORMATION FEATURE EXTRACTION

2.1 Framework

This paper's research questions are based on the pre-processing of single machine data, hoping to filter out the device status data impacting the power prediction longitude. This can strengthen data quality and optimize the accuracy of single machine power prediction and establish a data foundation for the operation safety of new energy consumption, achieving lean consumption management. In order to improve lean management's power prediction in the later stage, this article mainly improves the method of data feature extraction. Based on improving TF-IDF with the time slot, we extract the feature weight of the data. Additionally, we label the status of the device according to several status categories. The concept of self-learning is also used to divide the state of single machine data. Finally, the abnormal single machine state is eliminated, and then the power prediction value is optimized. On the other hand, we conducted research on power forecasting and optimization based on single machine wind power. We deepened the single machine state's accurate split based on its data quality, which can enhance the timeliness of the data. Further, we collect the information of a single machine during a specific period, dividing abnormal states. Thereby we can accurately improve the data quality of the single machine and provide a basis for the optimization calculation of power prediction. The overall framework is shown in Figure 1.

2.2 Problem system

1) New energy power prediction and optimization: Most of the existing power prediction methods are aimed at the dimensions of the station and power grid. However, the new energy lean management method is limited to a single machine perspective.

2) Single machine information state splitting: Establish a self-learning model by studying natural language processing technology and based on existing experience state labels. Then we should assess the newly added incremental scene data and divide the status labels of the single machine units, reducing abnormal states.

3) DF-IDF data feature extraction algorithm based on time slot: Given the scheduled day of a specific period and the feature extraction of single machine information data in the maintenance log, we can analyze the continuous state of the single machine and timeliness. Finally, we can determine the label of DF-IDF data according to the characteristic state weight.

4) The foothold of power prediction optimization: Based on the above work in (2) and (3), firstly, the single-machine data is emphasized by the state weight feature of a specific period, and the power prediction value of the single machine will be calculated. We should then compare the accuracy in

![Figure 1 Overall framework of research methods](image-url)
different methods, realizing the research question of the overall new energy power prediction optimization in (1).

3. RESEARCH ON THE DIVISION OF WIND POWER SINGLE MACHINE

According to the dispatch log and maintenance log of Jibei Power Grid, this paper uses a serious and standardized input process, and divides its status, and the results are verified by algorithms.

3.1 Overview of fan data status collection

The state subdivision of wind power single machine data collection is illustrated in Figure 2. The single machine wind power status (12 types) waiting, power generation, unit self-derating, abnormal weather derating, dispatch power limit derating, planned outage, fault outage, abnormal weather outage, dispatch outage, on-site consumption, off-site consumption, and communication interruption. This work can provide a foundation for feature extraction and maintenance for subsequent research and verification of data feature extraction algorithms. According to the existing standard process, dependent upon the dispatch log and maintenance log, we try to extract the electric field name, unit information, power outage reason, shutdown and grid connection time and other information in the log, and then sort out the status information of the single machine and split the abnormal state. Eventually, it can realize the split of each wind power single machine.

3.2 Grid Regulation Data Label Thesaurus and New Energy Characteristic Rules and Regulations

Grid control data labels: Dispatching centers, wind farm, branch center, provincial dispatch, equipment name, high voltage level, low voltage level, gathering station, photovoltaic power station, centralized control station, line, substation, personnel, grid shutdown, grid connection, full-field, fault, Overhaul.
3.3 Recognition rules of new energy-related state features
Based on the background of new energy power prediction optimization, we need to identify the rules of the new energy state characteristics. Subsequently, we can filter the single machine information to access the status.

Step (1): classify the dispatch log and maintenance log: new energy, suspected, and non-new energy. Examples are as follows:
"New energy" = wind farm tag & unit number + photovoltaic power plant tag & unit number.
Step (2): classify dispatch logs and maintenance logs: on-site and off-site. Examples of expressions are as follows:
"Off-site" = substation tag-(wind farm + photovoltaic) tag+(field station tag-(wind farm+photovoltaic) tag)
Step (3): classify dispatch logs and maintenance log logs: shutdown, grid connection.
In terms of the two vocabularies of "shutdown keywords" and "grid-connection keywords," the logs are marked as "1" and "2". The electric field has the same name, and the 95% similar fault (1) and grid-connected (2) logs of the unit number are related to each other, and the uncorrelated mark 0 is output.
Step (4) classify the dispatch log and maintenance log: fault, maintenance.
Examples are as follows:
Fault = fault tag

3.4 Status analysis and feature maintenance of wind power single machine
This article will delete the status information of abnormal faults in a single machine, and analyze the reasons for the shutdown status of the fan and understand the characteristic maintenance information. The unit state split results will be combined and generated according to the tags, see Figure 4.

| Cause identification rule | Coding | Description |
|---------------------------|--------|-------------|
| On-site                   | 0000   | On-site Fan shutdown |
| On-site                   | 0001   | On-site maintenance |
| On-site                   | 0010   | On-site other fault(planned outage) |
| On-site                   | 0011   | On-site Other overhaul |
| Off-site                  | 0120   | Off-site Line fault |
| Off-site                  | 0121   | Off-site Line overhaul |
| Off-site                  | 0130   | Off-site Substation failure |
| Off-site                  | 0131   | Off-site Substation overhaul |

Figure4 Tag state combination

4. DF-IDF DATA FEATURE EXTRACTION ALGORITHM BASED ON IMPROVED TF-IDF WITH TIME SLOT
In order to resolve the problem of new energy consumption and improve the lean management of new energy, the quantification of single-machine status information is a very important premise to determine the extraction of new energy data features [10].

4.1 Algorithm background research
Although we have standardized grid system log records, different dispatchers understand and write log records in different ways. Therefore, we should extract log records efficiently as this is rational dispatch and consumption of new energy.

Problems:
1) It is hard to incorporate production and operation information in data such as scheduling logs, maintenance logs, and business processes into data collection and analysis more effectively.
2) The unstructured data of different units have different characteristics and complexity, so we need to accumulate the data structure characteristic resources allocated by the grid province power company.

According to reading literature, the efficiency of multi-source data feature extraction, data sparsity, semantic expansion, and dynamic data processing issues may become future research trends [11]. In the feature extraction of power grid business data, multiple verification research methods are relatively common. However, we lack the concept of "time" and "timeliness," resulting in poor timeliness and low research value.

In the field of application, we mostly use simple text feature extraction and weight calculation, but there are few studies on data feature extraction that integrates the concept of a time slot. The grid dispatch log and grid maintenance log operate and record grid operation information, an important data source for the power grid operation. The category of natural language is covered in unstructured and institutionalized data [4]. The meaning of natural language may include words, sentences, paragraphs, or entire articles. The category of natural language is covered in unstructured and institutionalized data [4]. The meaning of natural language may include words, sentences, paragraphs, or entire articles. Therefore, extracting the features of words is the most commonly used extraction method [3]. Traditionally, for the feature extraction of power grid data, the traditional word bag model one-hot coding method does not consider the word order, and only assumes that the words are independent of each other, which is idealistic. Moreover, the feature extraction results are discrete and sparse, and the calculation space is complicated.

Finally, the current word in N-gram notation is only related to the previous word, but not related to the latter word.

The TF-IDF feature extraction method is simple and fast, and its results are more realistic, but it only considers the word frequency, ignoring the position between sentences, and the relationship between words, so that it should improve.

Take the feature extraction of wind power stand-alone as an example; This paper adopts natural language word segmentation and other data preprocessing techniques to achieve feature extraction and deletes invalid samples. In combing the grid dispatching and maintenance logs, we describe the sample data of equipment status, faults, and defect maintenance related to the power grid's operation status and analyze the correlation between the fan operation status and the log information. Finally, we use natural language understanding technologies such as natural language word segmentation, part-of-speech analysis, and syntax analysis to extract data features of scheduling and maintenance logs. Finally, we use natural language understanding technologies such as natural language word segmentation, part-of-speech analysis, and syntax analysis to extract data features of scheduling and maintenance logs.

4.2 Significance of DF-IDF algorithm
This section improves the traditional TF-IDF calculation method. Combined with the questions raised by this article, we use time validity and feature extraction to distinguish events, causes, and headlines in different periods. In doing so, you can accurately consult, classify, and search the content you like. Therefore, the keyword vector feature weight DF-IDF of the time slot is an improved method of the algorithm in our research.

According to the scheduling log and maintenance log data, if we use TF-IDF, and a user of a provincial transfer jurisdiction unit A sends a log data about scheduling or maintenance, then there will be ten times the keyword "overhaul management." However, this does not mean the data sent by the user is the key feature data, and this means the high frequency of the occurrence of the keyword.

Therefore, DF-IDF is used in this paper. If users A and B both send several pieces of grid dispatch log data (assuming n). Then if scheduling log data(n) is all keywords such as "overhaul management" during the period from t-1 to t, which means the keywords no matter how many schedules are sent by any unit user during this period, it appears in the maintenance log data. This also emphasizes the keyword characteristics of the data in a certain period. For example, according to the "overhaul management" label management of these data, we can explore the reasons for the occurrence of maintenance-related events, which is more time-sensitive for grid data.
To sum up the two points, this paper proposes the formula of DF-IDF to divide the grid dispatch and maintenance log based on a time slot, see formula 1.

### 4.3 Research and solution of DF-IDF algorithm based on time slot

- **Definition 1:**
  The concept of DF-IDF is to calculate DF-IDF based on time segmentation. DF-IDF is based on the frequency of occurrence of word vectors in grid dispatching and maintenance log data within a certain period of time, and represents the importance of word frequency. The specific definition is shown in Equation 2.

- **Definition 2:**
  The specific definition of the word vector matrix VectorMatrix after word segmentation is shown in Equation 2.

- **Definition 3:**
  The time node in the time slot is \( t \), \( \{t|1 \leq t \leq T\} \).

- **Definition 4:**
  The number of log data where the keyword \( k \) appears is \( \text{Num}_k(t) \).

- **Definition 5:**
  The data quantity of all dispatch logs and maintenance logs from time \( t-1 \) to time \( t \) is \( \text{Num}(t) \).

- **Definition 6:**
  The word vector feature weight DF-IDF expresses the importance of the keyword vector in a certain period of time, see formula 2 for details.

In the solution process, its working principle is based on the word vector that has been constructed in the previous step.

We need to define the time slot. For natural time, the word vectors generated by the scheduling and maintenance log data are divided according to specific time intervals (the time interval is 1 day) to form several time slots.

We find the DF value of the keyword of each word vector in each time slot according to formula 2, and calculate the IDF value of the keyword of each word vector in all time slots. Combining the two, the purpose is to highlight the importance of unstructured data with a keyword for a certain period of time.

Next, we perform feature extraction calculations to reflect the concept of time slot aggregation.

\[
\text{DF - IDF} = \frac{\text{Num}_k(t)}{\text{Num}(t)} \times \log \frac{\sum_{t'=t}^{t+1} \text{Num}(t)}{\sum_{t'=1}^{T} \text{Num}_k(t)}
\]

Among them, \( \text{Num}_k(t) \) is the number of scheduling and maintenance logs that contain the keyword \( k \) in the time period \( t-1 \) to \( t \) for some scheduling and maintenance log data. \( \text{Num}(t) \) is the number of all scheduling and maintenance logs in the period from time \( t-1 \) to time \( t \), on lines 1-2 of the computeDFIDF function. In addition, the beginning of a certain time period represented by \( \log \frac{\sum_{t'=t}^{T} \text{Num}(t)}{\sum_{t'=1}^{T} \text{Num}_k(t)} \) as a whole is the sum of \( n \) times the time slot to the end of a certain time period.

In order to distinguish it from the traditional word frequency, \( (\text{Num}_k(t)/\text{Num}(t)) \) only represents the ratio of the number of scheduling and maintenance log data containing keywords. It is very necessary to design the above formula in the context of grid dispatch log or maintenance log data, because multiple occurrences of the same word are usually associated with the same event in a grid dispatch and maintenance log data.

In order to distinguish it from the traditional word frequency, the difference of \( \log \frac{\sum_{t'=t}^{T} \text{Num}(t)}{\sum_{t'=1}^{T} \text{Num}_k(t)} \) that the set size of the traditional IDF is fixed, and the generation of new data in the power grid is very large. Therefore, the IDF design in the formula makes it possible to accommodate new data.

In order to distinguish it from the traditional word frequency, the difference of \( \log \frac{\sum_{t'=t}^{T} \text{Num}(t)}{\sum_{t'=1}^{T} \text{Num}_k(t)} \) that the set size of the traditional IDF is fixed, and the generation of new data in the power grid is very large. Therefore, the
IDF design in the formula makes it possible to accommodate new data. The scope of the collection does not need to be fixed, because with the development of power grid informatization and artificial intelligence. New words are also generated very quickly. The part to the right of the multiplication sign can accommodate a lot of new words, in the computeDFIDF function computeDFIDF line 15-28.

The existence of DF-IDF in the whole time period makes the use frequency of keywords between t-1 and t higher than other words, then the value of DF-IDF is higher. In other time periods from 1 to T_c, the keyword k rarely occurs, and the value of DF-IDF is low. The process of solving the feature weight of DF-IDF word vector is as follows.

Algorithm computeDFIDF: Calculate the feature weight of word vector

Input: Corresponding word VectorMatrix after word segmentation

Output: WordVector Matrix WeightVectorMatrix with a weight value of DF-IDF

1. VectorTimeSlot ← dividedByTime(VectorMatrix) //VectorMatrixIdf ← null // Initialize the idf value of each keyword in the word vector matrix
2. allVectorNum = length(VectorMatrix)
3. for keyword in allKeywords // Iterate over each keyword and find its idf value
4.    keywordNum = 0
5.    for oneVecotr in VectorMatrix
6.        if(oneVecotr[keyword]>0)
7.            keywordNum += 1
8.        end if
9.    end for
10.   VectorMatrixIdf[keyword] = allVectorNum/keywordNum
11. end for
12. for(t=1;t<=T;t++)
13.   VectorMatrixDf ← null // Initialize the df value of each keyword in the word vector matrix in the time period
14.   currentVectorList = VectorTimeSlot[t] // Get a list of word vectors in that time period
15.   currentVectorNum = length(currentVectorList)
16.   for keyword in allKeywords // Iterate over each keyword
17.      keywordNum = 0
18.      for oneVecotr in currentVectorList
19.         if(oneVecotr[keyword]>0)
20.             keywordNum += 1
21.         end if
22.      VectorMatrixDf[keyword] = keywordNum/currentVectorNum
23.     end for
24.   end for
25. for oneVecotr in currentVectorList
26.     for keyword in allKeywords
27.        WeightVectorMatrix[oneVecotr][keyword]=VectorMatrixDf[keyword]*log(VectorMatrixMatrixDf[keyword])
28.     end for
29. end for
After the original grid dispatching and maintenance log data is represented by the word vector space, it corresponds to the data of each row. After we calculate the list of word vectors in the current time period, we can output the weight vector of word vector features after dimensionality reduction.

5. EXAMPLE VERIFICATION

5.1 Experimental environment
Software: Python 2.7
Hardware: windows 7

5.2 Application background and data conditions
Taking the part of the grid dispatch log of Hebei North Power Grid in 2019 as an example, it contains 6,905 valid records. In addition to the date and time recorded every day, the shortest record contains 17 characters and the longest record contains 465 characters. It can be seen that the length varies greatly. After removing the invalid samples, manually labeling log categories, and word segmentation data preprocessing steps, 1125 pieces of data containing wind power were formed here.

For example, the original log of a grid dispatch log labeled "Wind power generation" and the text after the labeled participle are as follows. The labels are displayed in square brackets.

The part-of-speech word segmentation of the unstructured log data is combined with the grid lexicon and tag library, and the results are as follows.

2019-12-05 12:10:43 [time] Chengde local dispatch [dispatching agency] (Zhang Yong[name]) reported[command] Muhei Line [Line] Two sets of NSR and PCS protection channels on both sides are abnormal [Fault] Device abnormal [fault signal]. Application [instruction] will be [?] Datang [Group] Daheishan Wind Farm [Wind Farm] Station equipment and [referred to] 220kV [Voltage Class] Muhei Line [Line] Transfer to Hot Standby [Action? Synonyms]. Hebei North Dispatch [Dispatching Agency] agrees to [Instruction]. 12:40 [Time] Datang [Group] Daheishan Wind Farm [Wind Farm] Station equipment and [referred to] 220kV [Voltage level] Muhei Line [Line] has been [temporal] transferred to hot standby [action].

5.3 Stand-alone information status labeling
We can combine the existing normative rules studied in Chapter 3 of this article and set the feature extraction labels in the range of new energy as shown in Table 1 in order to facilitate the verification of the effect of the algorithm and to ensure the scientificity of the manual labeling training.

| Name & Example | Dispatch name, central control name, station name, personnel name, equipment name |
|----------------|----------------------------------------------------------------------------------|
| Time           | YYYY-MM-DD hh:mm:ss, hh:mm……                                                      |
| Shutdown term  | Disengage, accompany stop, stop, trip, frost, drop, repair, replace                |
| On-line terminology | Eliminate, connect to the network, complete the deficiencies, send power, pick up |

5.4 Algorithm evaluation criteria
This paper evaluates the DF-IDF algorithm indicators mainly including the effect comparison before and after the algorithm improvement, the recall rate, accuracy rate, accuracy rate, and F value of the status hashtag.

9
• Definition 1: The number of data of the stand-alone status label predicted to be positive, \( B_{\text{Label\_num}} \), indicates the number of data that originally belonged to a real topic status label.

• Definition 2: The number of status labels for a single machine, \( G_{\text{Label\_num}} \), represents the number of original real-world theme labels.

• Definition 3: Stand-alone state label recall rate, \( \text{Label\_Recall} \), can be seen in formula 3 for specific definition.

\[
\text{Label\_Recall} = \frac{B_{\text{Label\_num}}}{G_{\text{Label\_num}}}
\] (3)

• Definition 4: The prediction accuracy of the stand-alone status label, \( \text{Label\_Precision} \), is defined in Equation 4.

\[
\text{Label\_Precision} = \frac{B_{\text{Label\_num}}}{P_{\text{Label\_num}}}
\] (4)

• Definition 5: The single-machine state prediction comprehensive index, \( F\_\text{Measure} \), is defined in Equation 5.

\[
F\_\text{Measure} = \frac{2 \times \text{Label\_Recall} \times \text{Label\_Precision}}{\text{Label\_Recall} + \text{Label\_Precision}}
\] (5)

• Definition 6: The overall stand-alone state prediction accuracy rate, \( \text{Label\_Accuracy} \), is the accuracy rate evaluated for all sample sets (n), as defined in Equation 6.

\[
\text{Label\_Accuracy} = \frac{\sum B_{\text{Label\_num}}}{\sum G_{\text{Label\_num}}}
\] (6)

In short, \( \text{Label\_Recall} \) represents the single-machine state theme recall rate; \( \text{Label\_Precision} \) represents the single-machine state theme precision rate; and \( F \) value is the value obtained by considering the accuracy rate and the recall rate; \( \text{Label\_Accuracy} \) in this article represents the overall algorithm performance measurement index under different single-machine theme states.

5.5 Algorithm evaluation standard
According to the effect of feature extraction of test data, we use accuracy rate, recall rate, precision rate and \( F \) value to comprehensively verify. Based on the stand-alone status topic tag research, we summarized the situation of the 6 major types of grid business data of the day according to the data mentioned, combined with Table 2, and business experience. Next, we label the collected stand-alone data with status thematic tags, see Table 2 for details.

### TABLE II. STANDALONE STATUS LABEL TOPIC DESCRIPTION

| Topic name | Corresponding content stand-alone status label |
|------------|-----------------------------------------------|
| A          | On-site fan maintenance                       |
| B          | On-site downtime                              |
| C          | Unplanned outage of fans in the yard          |
| D          | Off-site lines, substations, faults           |
| E          | Off-site lines, substations, maintenance      |
| F          | Off-site peak shaving shutdown, off-site peak shaving derating |

The hashtags and names of the stand-alone status in this table are clearer for the following experimental analysis.
Figures 5 and 6 show the first three DF-IDFs with word segmentation vectors and the largest weights. Its characteristics represent the state of the topic that has been the most followed in a certain period of time. Figure 5 is the feature extraction effect of the pre-improvement algorithm (TF-IDF), which is roughly divided into three types of status tags. It can be clearly seen that two different labels are extracted in the densest part in the middle, and the distance is very close. The figure does not clearly show the characteristics of high or low similarity, indicating that the effect before improvement is not satisfactory. Fig. 6 is the feature extraction effect of the improved algorithm (DF-IDF). The figure is clearly divided into three types of status labels, and only one label is gathered in the densest part in the middle. This proves that the feature keyword has a high similarity of word vectors after feature extraction, and the similarity between the stand-alone theme keyword vectors is low. They are not confused with other types of stand-alone status labels, which reflects the superiority of the DF-IDF algorithm and the important advantages of feature extraction.

According to the different situations of the number of stand-alone status labels, we compare the overall average accuracy of each algorithm. The accuracy is given by formula (6). The results are shown in Table 3.

### TABLE III. THE AVERAGE ACCURACY OF THE THREE ALGORITHMS UNDER DIFFERENT STAND-ALONE CONDITIONS IN CHINESE TABLE QUESTIONS

| Number of Labels | One-hot | TF-IDF | DF-IDF |
|------------------|---------|--------|--------|
| 2                | 0.71    | 0.71   | 0.73   |
| 3                | 0.63    | 0.66   | 0.68   |
| 4                | 0.61    | 0.64   | 0.69   |
| 5                | 0.57    | 0.60   | 0.65   |
| 6                | 0.58    | 0.59   | 0.66   |
In the above data, TF-IDF represents the traditional word frequency weight feature extraction algorithm. One-hot is a feature extraction algorithm that counts the number of words. DF-IDF represents the time slot-based data feature extraction algorithm proposed in this paper.

In the improved DF-IDF algorithm, regardless of the number of content status tags, the average accuracy of the algorithm is generally higher than that of the original TF-IDF data feature extraction algorithm before improvement.

6. RESEARCH ON POWER PREDICTION AND OPTIMIZATION

6.1 Single-machine status screening and elimination
We use the above data feature extraction method to filter the status information of the single machine. Pre-processing the status data of shutdown, failure, expansion, accompany stop, and interruption. Functionally, this method is used to repair real-time data. The repair result is shown in Figure 7.

![Data repair effect diagram](image)

In addition, we need to preprocess the data of the stand-alone information in the operation data of Hebei North Wind Power. Before the pre-processing, the abnormal rate of data at the device level such as stand-alone is high, exceeding 10%. After pre-processing, the data anomaly rate of the single-machine and other device layers is reduced to about 8%.

6.2 Lean power prediction and optimization

1) Stand-alone power prediction calculation
In terms of the stand-alone information and digital weather forecast information, we calculate the stand-alone power prediction value. In addition, according to the improved DF-IDF, we screen out the uncertain state data of single machine maintenance, failure, and expansion, as shown in the process in Figure 10. By calculating the stand-alone power prediction value, we compare the short-term and ultra-short-term power prediction accuracy before and after using the DF-IDF algorithm. The steps are as follows.

a) First, we select the three factors that have the greatest correlation with the output of a single machine from the five factors of wind speed, wind direction, air pressure, humidity, and temperature through the correlation coefficient.

b) In addition, we need to set the maximum three-factor threshold, according to the correlation value of the factors from high to low, and iteratively filter the historical data to obtain the corresponding results. If the result set is empty after iterative filtering of a factor, we need to add the factor field length and iterate again until the result set is not empty.

c) Finally, after iteratively screening the three factors, we need to take the average of the top 40% of the power in the result set as the theoretical power value.

The result of single-machine power optimization is shown in Figure 8. The comparison results of the predicted fan power and actual output data are as follows.
Through the above results, we found that the predicted value is basically consistent with the true value. We have analyzed that the lean management method for power forecasting is mainly aimed at the single-machine short-term power forecasting and ultra-short-term power forecasting data, and calculates its reporting rate, pass rate, accuracy rate, and average absolute error. This helps to improve the station management level and optimize the new energy management methods.

6.3 Aggregate power prediction and optimization

Regarding the accuracy of the stand-alone power prediction, the overall accuracy of the short-term power prediction and ultra-short-term power prediction of the total station or the entire network is optimized to more than 95%.

7. Conclusion

Based on the contradiction of new energy consumption, this paper studies the optimization problem of new energy power prediction. Starting from the stand-alone information, we have integrated the precise concept of time slot. In addition, based on the idea of improving TF-IDF, we propose the feature extraction method of DF-IDF in order to realize the extraction of business data features of unstructured natural language data of dispatch logs and maintenance logs, so as to realize the extraction of stand-alone information features and remove anomalies status. This means that it has certain theoretical and practical significance in the category of artificial intelligence research [14]. In order to verify the effectiveness of the algorithm, we combine the state labels of feature extraction to perform self-learning classification. The abnormal state of the single machine is proposed from various states, which is helpful for the prediction and optimization of new energy power. Finally, we conduct prediction accuracy analysis for single-machine power prediction calculations, so as to summarize and optimize the power forecast lean management methods in the dimension of the station or power grid. The purpose of this paper is to promote the power grid to solve practical problems and improve the consumption space and operation management level.

REFERENCES

[1] Chen Guoping, Liang Zhifeng, Dong Yu. Analysis and Thinking of the Construction of Electricity Market with Chinese Characteristics Based on Energy Transformation [J]. Chinese Journal of Electrical Engineering, 2020, 40 (2).
[2] Xue Yusheng, Yu Chen, Zhao Junhua, et al. Comment on short-term and ultra-short-term wind power prediction [J]. Automation of Electric Power Systems, 2015 (6): 141-151.
[3] Zhong Haiwang, Xia Qing, Zhang Jian, et al. Mechanism design to encourage wind farms to improve the accuracy of power prediction[J]. Automation of Electric Power Systems, 2015, 39 (5): 47-53.
[4] Yang Quxia, Liu Tongxin, Gao Chen, et al. Power system optimization analysis considering power prediction of PV power station% power system optimization analysis considering photovoltaic power station power prediction [J]. Power System Protection and Control, 2018, 046 (002): 117-123.
[5] Wang Bo, Liu Chun, Feng Shuanglei, et al.% Prediction Method for Short-term Wind Power Based on Wind Farm Clusters [J]. High Voltage Technology, 2018, 044 (004): 1254-1260.
[6] Shi Jie, Liu Xiaofei. Study on optimization of new energy power prediction algorithm [J]. Generation Technology, 2019, 40 (01): 78-82.

[7] Qiao Ying, Lu Zongxiang, Min Yong. Methods to improve the prediction accuracy of wind power [J]. Power System Technology, 2017, 041 (010): 3261-3268, Intermediate 3.

[8] Wang Zhao, Wang Bo, Feng Shuanglei, Wang Weisheng. Probability prediction method of quantile regression for regional multi-wind farm power [J]. Grid Technology, 2020, 44 (04): 1368-1375.

[9] Xie Congzhen, Wang Jiangchu, Xie Xinhao, Liu Zhijian, Bai Jianfeng. BOA-GBDT photovoltaic output prediction based on fine-grained features [J]. Grid Technology, 2020, 44 (02): 689-696.

[10] Cui Yang, Qu Yu, Zhong Wuzhi, et al. Trend-based Quantitative Method for Wind Power Convergence Trend Based on Improved Shapley Value [J]. Power System Technology, 2019, 43 (6).

[11] Liao Ruijin, Wang Youyuan, Liu Hang, et al. Research status of transmission and transformation equipment condition assessment methods [J]. High Voltage Technology, 2018, 44 (11).

[12] Du Ruopeng, Xian Guojian, Kou Yuantao. Text feature extraction of agricultural science and technology literature based on improved TF-IDF-CHI algorithm [J]. Digital Library Forum, 2019 (8).

[13] Dong Ruifang, Liu Changan, Yang Guotian. A SLAM loop detection algorithm based on improved TF-IDF [J]. Journal of Southeast University (Natural Science Edition), 2019, 49 (02): 48-55.

[14] Ma Lihong, Qin Dan, Lin Qiang, et al. Intelligent extraction of key indicators of distribution network projects based on NLP method [J]. Automation and Instrumentation, 2019 (4).

[15] Podgorsak A R, Rava R A, Bhurwani M M S, et al. Automatic radiomic feature extraction using deep learning for angiographic parametric imaging of intracranial aneurysms [J]. Journal of Neurointerventional Surgery, 2019: neurintsurg-2019-015214.

[16] Chouigui A, Khiroun O B, Elayeb B. A TF-IDF and Co-occurrence Based Approach for Events Extraction from Arabic News Corpus[C]// 23rd International Conference on Natural Language & Information Systems (NLDB 2018). Springer, Cham, 2018.

[17] Kadhim A I. Statistical computation and term weighting for feature extraction on Twitter[C]// 2018 International Conference on Advance of Sustainable Engineering and its Application (ICASEA). IEEE, 2018.