



Received January 25, 2021, accepted February 22, 2021, date of publication March 2, 2021, date of current version March 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063354

Entity Extraction of Electrical Equipment Malfunction Text by a Hybrid Natural Language Processing Algorithm

ZHE KONG¹, CHANGXI YUE², YING SHI¹, JICHENG YU¹², (Member, IEEE),
CHANGJUN XIE¹¹, (Member, IEEE), AND LINGYUN XIE³

¹School of Automation, Wuhan University of Technology, Wuhan 430070, China

²China Electric Power Research Institute, Wuhan 430070, China

³Fiftieth Research Institute, China Electronic Technology Group Corporation, Shanghai 200331, China

Corresponding authors: Ying Shi (yingsh@whut.edu.cn) and Jicheng Yu (yujicheng@epri.sgcc.com.cn)


This work was supported by the State Grid Corporation of China Headquarter Science and Technology Project under Grant 5600-201918181A-0-0-00.

ABSTRACT Many electrical equipment malfunction text messages are collected during power system operation and maintenance procedures. These texts usually contain crucial information for maintenance and condition monitoring. Because these power system malfunction texts are characterized by multidomain vocabularies, complex-syntactic structures, and long sentences, it is challenging to for automated systems to capture their semantic meaning and essential information. To address this issue, we propose a hybrid natural language processing (hybrid-NLP) algorithm to extract entities that represent electrical equipment. This algorithm is composed of a dictionary-based method, a language technology platform (LTP) tool, and the bidirectional encoder representations from a transformers-conditional random field (BERT-CRF) model. Significantly, the softmax output layer of the bidirectional encoder representations from the transformers (BERT) model is replaced by the conditional random field (CRF) algorithm to strengthen the contextual relationships between words and thus solve the local optimization of the preferred word label. The effectiveness of the proposed hybrid-NLP method is verified on a realistic dataset. Moreover, a statistical analysis is conducted to provide a reference for the operation and maintenance of power systems.

INDEX TERMS Electrical equipment malfunction text, natural language processing, entity extraction, BERT-CRF model.

I. INTRODUCTION

A large amount of malfunction data is collected during power system operation and maintenance procedures [1]. These data consist of semistructured and unstructured texts [2] that refer to a large proportion of modern power system equipment [3] and contain critical information highly relevant to power grid safety and security. However, only a small number of these texts can be utilized at present [4]. By applying natural language processing (NLP) technology, these malfunction texts can be analyzed effectively to obtain more useful information, such as the names of equipment and their corresponding failure rates. Such information is vital for maintenance and for monitoring the condition of power systems.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang .

Named entity recognition (NER) is a core NLP technology employed to identify entities such as persons, locations, organizations, and dates in texts [5]. NER is the basis of entity-relationship extraction [6], knowledge graphs [7], and automatic question answering systems [8], which have been widely applied in many other fields [9], [10]. Therefore, NER is a potentially practical approach for mining critical information from electrical facility malfunction texts.

However, the text of the power system maintenance field is unstructured data, which has no definite form and lacks machine-understandable semantics [4], consequently, transforming it into structured data is necessary before further mining. Furthermore, as power systems expand, some entity types continue to grow, such as the number of lines and manufacturers' names [11], [12]. This introduces a demand for vast text corpora and eventually results in problems extracting entities limited by the effectiveness and scalability of these

corpora. It is difficult to address this issue using a single NER method. At present, no acceptable NER framework exists for performing comprehensive entity extraction from electrical equipment malfunction texts of due to the abovementioned complexities [13].

Regarding different languages, it is necessary to consider the text processing differences between Chinese and English. A natural space always exists between words in English texts, making it easy to segment words in English texts. However, Chinese texts lack a natural separator and have complex forms (entities consist of multiple characters), making named entity recognition in Chinese texts more difficult [14], [15]. Furthermore, Chinese texts are annotated based on characters, while English texts are annotated based on words [16]. Thus, compared with English text, processing Chinese text is more complex and challenging. Therefore, it is necessary to propose a practical NLP framework to extract entities from electrical equipment malfunction text. In addition, mining useful data and providing a useful reference for system operation and maintenance is of critical importance. The abovementioned aspects form the motivation for our work.

This paper focuses on Chinese texts. According to the characteristics of different equipment malfunction entity categories, a hybrid-NLP framework for various types of entity extractions of electrical equipment malfunction information is proposed to address the issues mentioned above. First, a dictionary-based method is adopted to extract proper nouns belonging to the electric power field, while a language technology platform (LTP) tool is employed to extract time entities. Specifically, we develop the BERT-CRF algorithm to extract strongly extensible entities, including line names and manufacturer names. Based on the entity extraction, we can then perform statistical analyses to obtain prior knowledge of malfunctions from historic maintenance records, allowing more attention to be paid to equipment maintenance and the components most prone to failure to prevent latent power system faults in advance. The proposed method is suitable for the analysis of malfunction texts in the electric power field.

The remainder of this paper is organized as follows. Section II reviews the related works on Chinese NER. Section III provides the definitions of malfunction entities and the structure of the proposed entity extraction framework. Section IV presents the preprocessing process for malfunction text data. Sections V and VI describe the extraction processes for malfunction text entities. Finally, Section VII presents the experiment and reports the results of the statistical analysis. Conclusions are drawn in Section VIII.

II. RELATED WORK

The NER method mainly involves approaches based on rules, statistical machine learning, and deep learning [17]. For example, Sun *et al.* [19] applied a meta-character template generalization slot-based syntax rule template to extract critical factors of power systems, such as equipment locations, involved ontologies, corresponding indexes, and index values from nonstructured preplan texts.

In addition, Wang *et al.* [20] adopted a rule-based method to perform NER for power equipment operation and maintenance. The rule-based method primarily adopted machine learning models such as support vector machines as classifiers to conduct text mining from power system defect logs, such as power equipment crucial defect information. Finally it constructed a knowledge graph based on the resulting graph. Wang *et al.* [21] also applied a rule-based approach that involved an ontology dictionary and a semantic framework to extract defect components and attributes for further research. The rule-based methods achieved good performance for extracting specific entities; however, they are difficult to adapt to complex language texts due to their poor generalizability [21].

Machine learning provides many effective solutions to address this issue. Ji *et al.* [13] proposed a named entity recognition algorithm that extracted power system entities based on conditional random fields (CRF) and bidirectional long short-term memory (BLSTM) models. The CRF model achieved an accuracy of 83%, indicating that it can identify power entities better than the BLSTM model. In addition, Chen *et al.* [23] counted word frequencies in power system text through the TF-IDF algorithm. By ranking the TF results, they were able to extract critical TF height information. However, the above methods are trained through labeled datasets, and the experimental results are highly dependent on the quality of the training dataset [23]. Moreover, the experimental performance is also impacted by different text features, and the universality and accuracy of the algorithm need to be improved [24].

To address this issue, deep-learning-based methods were introduced. Xie *et al.* [26] proposed a deep learning algorithm based on an RNN-LSTM model for performing NER on fault inspection report records to extract defect category entities. The experiment verified the excellent performance of the RNN-LSTM model. Investigating similar deep learning approaches, Wei *et al.* [27] proposed a deep BP neural network matching model that could extract target objects from operating instructions with an accuracy of 95%. Furthermore, Jin *et al.* [28] introduced an attention mechanism into the LSTM model, whose corresponding accuracy rate, recall rate and F1-value were 93.71%, 92.46%, 93.08% on Chinese NER. Based on the attention mechanism, Google developed a new network model, i.e., bidirectional encoder representations from transformers (BERT), which has achieved good performances on NER tasks [28]. Subsequently, Yu *et al.* [29] used the BERT model to extract entity relationships and conducted experiments on public data sets that achieved good results. Such deep learning-based methods are more widely applicable because appropriate models can be obtained by merely retraining on a new dataset to address new tasks. However, the deep learning methods have high computer hardware requirements when applied to large datasets, resulting in high training costs.

NER has achieved remarkable performances in various fields. Indeed, it has been verified to be a possible way

to extract critical entity information related to equipment defects from electrical equipment malfunction texts. However, power system defect text consists of unstructured data that must first be converted into structured data. The expansion of entities in the power field also introduces difficulties to entity extraction. Finally, Chinese text is more complex and challenging for processing methods compared with English text. At present, Chinese NER is still in its infancy in electricity applications, and a complete process for addressing malfunction texts in this field is lacking. This paper focuses on the abovementioned issues, proposes an effective method for mining power system defect text, and provides a reference for effective power system operation and maintenance.

III. ENTITY EXTRACTION FRAMEWORK FOR ELECTRICAL EQUIPMENT MALFUNCTION AND DEFECT TEXT

A. ENTITY DEFINITION OF ELECTRICAL EQUIPMENT MALFUNCTION AND DEFECT TEXT

By analyzing electrical equipment malfunction and defect texts provided by a power grid company and referring to the Evaluation Criteria for the Reliability of Power Transmission and Transformation Facilities [30], we classified texts containing electrical equipment malfunction and defects into 8 types of entities: equipment name, component name, malfunction type, voltage level, production/operation time, malfunction level, line names and manufacturer names.

Then, we further grouped the entities into three categories to process similar catalogs of entities using the same approach. Specifically, Class I entities represent proper nouns in the electric power field, including equipment name, component name, malfunction type, malfunction level, voltage level, and limited entity types. For instance, the malfunction level also has limited catalogs (i.e., emergency defects and significant defects). For proper nouns, a dictionary-based approach for entity extraction is efficient and straightforward.

Similarly, production time and operational time both belong to time entities; therefore, they are classified into Class II entities that can be processed appropriately by the LTP tool.

Class III entities are strongly extensible and include line and manufacturer names. As the power grid expands and only the fittest companies survive, equipment line names and manufacturer names cover an increasing range, making it impossible to use a dictionary matching method to represent all the line and manufacturer names. Therefore, the extraction of line names and manufacturer names is the focus of this paper. We propose a BERT-CRF model to solve this issue.

B. ELECTRICAL EQUIPMENT MALFUNCTION TEXT ENTITY EXTRACTION FRAMEWORK

According to the entity classification system mentioned above, the proposed hybrid-NLP method framework for the

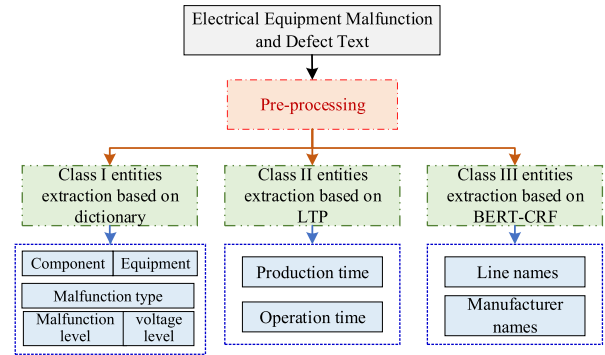


FIGURE 1. The entity extraction framework.

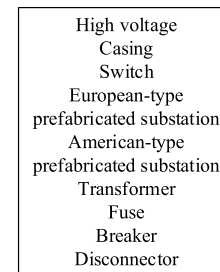


FIGURE 2. Example terms from the power field.

entity extraction of electrical equipment malfunction text is illustrated in Fig. 1. First, preprocessing operations such as word segmentation are performed on the text data; then, the entities are extracted using the dictionary, LTP tool, and BERT-CRF model. The hybrid algorithm executes these three algorithms in parallel and combines their results. It can automatically extract entity A for Class I, entity B for Class II, and entity C for Class III.

IV. TEXT DATA PREPROCESSING OF ELECTRICAL EQUIPMENT MALFUNCTION AND DEFECT INFORMATION

A. WORD SEGMENTATION OF ELECTRICAL EQUIPMENT MALFUNCTION AND DEFECT TEXT

Data preprocessing is a critical aspect of entity extraction. An LTP-customized dictionary approach is proposed to segment electrical equipment malfunction and defect texts. Some terms in the dictionary are depicted in Fig. 2.

For the sentence “配电运维小组检测到 10kV 龙岭线上的欧式箱变有明显破损(The distribution operation and maintenance team has detected obvious damage to the European-type prefabricated substation on the 10 kV Longling line)”, the results obtained by LTP-based word segmentation and the results of the LTP-customized dictionary word segmentation are shown in Fig. 3.

By comparing the above results of LTP-based word segmentation and LTP-customized dictionary word segmentation, it is apparent that after the LTP-customized dictionary word segmentation, “欧式箱变”(European-type prefabricated substation) is correctly segmented. That is,

Original defect text:
 配电运维小组检测到10kv龙岭线上的欧式箱变有明显破损
 LTP based word segmentation:
 配电 运维 小组 检测 到 10kv 龙岭 线上 的 欧式箱变 有 明显 破损
 LTP-customized dictionary based word segmentation:
 配电 运维 小组 检测 到 10kv 龙岭 线上 的 欧式箱变 有 明显 破损

FIGURE 3. Results of LTP word segmentation and word segmentation after the introduction of a customized dictionary.

the customized dictionary word segmentation results are a significant improvement, which helps ensure the reliability of subsequent tasks.

B. STANDARDIZATION OF ELECTRICAL EQUIPMENT MALFUNCTIONS AND DEFECT TEXT

Due to the irregularity of records, electrical equipment malfunction and defect texts may contain symbols such as “%”, “&”, and so forth. Because some special symbols do not express useful information, they are included in a stop word list. Stop words are removed to standardize the texts [31]. Note that symbols with specific meanings do not appear in the stop word list.

The text standardization steps are formulated as follows:

Step 1: Import the segmented electrical equipment malfunction and defect text and create a new data file to store the standardized text.

Step 2: Traverse each word in the current text, determine whether the word contains special symbols in the stop list, and if so, remove them.

Step 3: Store the words composed of Chinese characters in the data file.

For instance, we take the “配电运维小组\$#检测到 10kV 茶南线上的@变压器有明显破损 (The distribution, operation, and maintenance team \$# has detected obvious damage to @ transformer on the 10kV Chanan line)” as an example. According to the above procedure, some standardized results are shown in Fig. 4.

The special symbols \$, #, and @ have been removed from the text after standardization. Note that the processed text is more concise, which is conducive to improving the accuracy and effectiveness of entity extraction.

V. EXTRACTING ENTITY CLASSES I AND II FROM EQUIPMENT MALFUNCTION TEXTS

After word segmentation and standardization of electrical equipment malfunction and defect text, entity extraction is conducted as follows.

A. CLASS I DICTIONARY-BASED ENTITY EXTRACTION

For Class I entities, a dictionary-based method is proposed for entity extraction. The specific algorithmic procedure is illustrated in Fig. 5.

Five entity types are first stored in the dictionary. Each entity category has a corresponding dictionary with a category label. For each preprocessed electrical equipment malfunction and defect text, the words in each dictionary are

LTP-customized dictionary based word segmentation:
 配电 运维 \$ # 小组 检测 到 10kv 茶南线 上 的 @ 变压器 有 明显 破损
 Standardized result:
 配电运维小组检测到10kv茶南线上的变压器有明显破损

FIGURE 4. Standardized results from the example text.

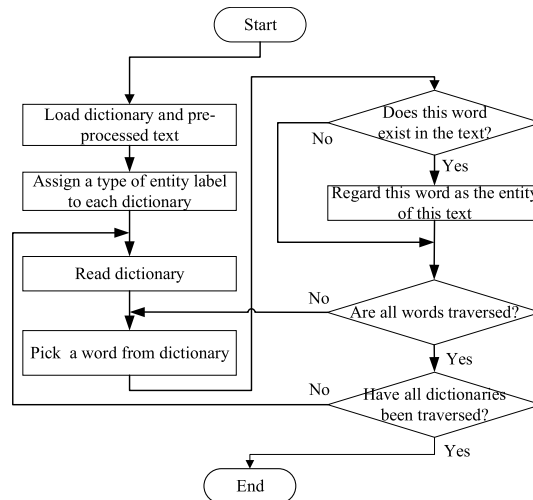


FIGURE 5. The dictionary-based entity extraction process.

traversed in turn to determine whether the word appears in the text record. When a matching word is found, the category of the word in the text is determined by the category in the dictionary to which the word belongs. This process is conducted iteratively. After traversing all the dictionaries, the five types of Class I entities in the electrical equipment malfunction and defect texts will have been extracted.

B. CLASS II LTP-BASED ENTITY EXTRACTION

Note that Class II entities are time entities. The LTP tool is employed to address this issue because it has been verified to achieve good extraction of such entities. The specific processing procedure is presented in Fig. 6.

First, word segmentation of each electrical equipment malfunction and defect text is conducted by LTP to perform part of speech (POS) tagging for each word. This process traverses all the words and tags them with their part of speech. On one hand, if the word is a time noun, the next word is also evaluated to determine whether it is a time noun. This process is conducted iteratively until a word that is not a time noun appears; then, all the time nouns are connected to form a time entity. Otherwise, the operation is repeated until all the texts have been traversed.

VI. CLASS III ENTITY EXTRACTION FROM EQUIPMENT MALFUNCTION AND DEFECT TEXT

Because the number of Class III entities is updated continuously and cannot be extracted using the preceding methods, this paper employs a BERT model as the backbone and further proposes a BERT-CRF model to extract such entities.

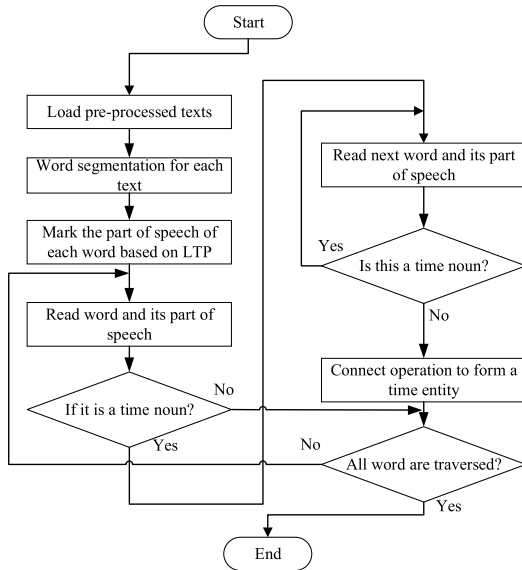


FIGURE 6. Tool-based time entity extraction.

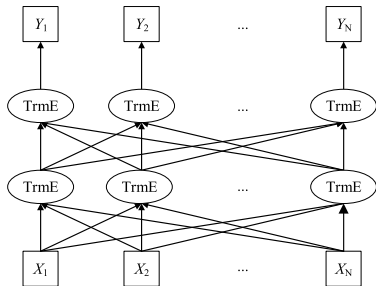


FIGURE 7. Structure of the BERT model.

A. WORD VECTOR REPRESENTATION BASED ON BERT

The transformer model [32] is a new architecture for a text sequence network, and BERT is a multilayer bidirectional transformer encoder based on fine-tuning. The structure of the BERT model is demonstrated in Fig. 7.

In Fig. 7, $Y_1, Y_2 \dots Y_N$ are the model outputs. For the entity extraction task, these outputs are a label corresponding to each character. The TrmE items represent the transformer encoder structure, which is the core part of the BERT model. In addition, $X_1, X_2 \dots X_N$ represent the model inputs, which include token, segment, and position. For instance, the input representations of “设备 (equipment)” and “开关 (switch)” are shown in Fig. 8.

In more detail, the character [CLS] denotes the start mark of the input information, while [SEP] is the end mark. On this basis, the three mentioned components are introduced as follows. First, the Token is the text input sequence, while the Segment is regarded as clause information, in which E_A represents the first sentence (denoted by “1”) and E_B is the second sentence (denoted by “0”). Furthermore, Position is positional information that represents the index position of each input character.

Note that the BERT model has 12 layers of the TrmE network, and each layer of the TrmE network is composed

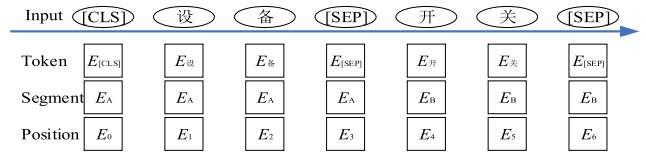


FIGURE 8. Structure of BERT model input.

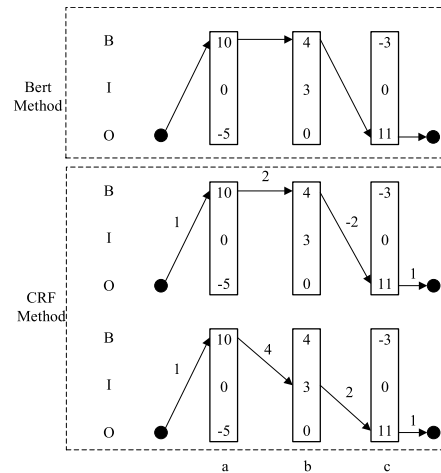


FIGURE 9. CRF sequence probability calculation.

of 6 encoder layers. In detail, encoder layer consists of a self-attention mechanism and a feed-forward neural network. Compared with traditional methods of expressing word vectors, the word vector obtained based on self-attention better reflects the textual meaning. The BERT model based on self-attention uses a multihead attention mechanism; the number of heads is set to 12 in this paper.

B. CLASS III ENTITY EXTRACTION BASED ON THE BERT-CRF METHOD

The conventional BERT model further extracts word vector features based on acquiring the value of the sequence word vector labels. On this basis, it also selects the label corresponding to the maximum value of each word through the softmax layer to complete the word vector feature classification. However, during the classification, the BERT model considers only local information; thus, it is prone to fall into a local optimum. However, both the word and its label have contextual relationships with neighboring words. Regarding this issue, CRF was developed to consider the context of the sequence information by selecting the globally optimal label to complete word vector classification.

A comparison of the BERT and CRF calculation sequence labeling processes is shown in Fig. 9. In this study, the BIO marking method is adopted, where “B” represents the beginning character of the entity. In addition, “I” denotes an internal character, which means that the character is part of the entity but not at the starting position. In addition, “O” denotes out, which means that the character does not belong to the entity to be extracted.

The three characters (a, b, and c) shown in Fig. 9 are the label sequences to be recognized. The three characters correspond to the values of the three labels “B”, “I” and “O”, respectively; thus, the actual label of a, b, and c sequence is “BIO”. When calculating the label sequence with BERT, each character is assigned the label with the highest value, that is, the predicted label of sequence a, b, and c is “BBO”, which is inconsistent with the actual situation.

It can be observed that the BERT model selects only the label with the highest value for each character, which results in an error when predicting the label of the sequence. In contrast, CRF considers the total value of the entire sequence $C(y_1, \dots, y_m)$, including the starting value, transfer value, and character value. The CRF calculation formula is

$$C(y_1, \dots, y_m) = b(y_1) + \sum_{t=1}^m s_t(y_t) + \sum_{t=1}^{m-1} T(y_t, y_{t+1}) + e(y_m), \quad (1)$$

where $b(y_1)$ and $e(y_m)$ are the values of the initial state and the end state, respectively. In addition, $s_t(y_t)$ is the value when the label is y_t , and $T(y_t, y_{t+1})$ is the value of the label when the y_t state is transferred to the y_{t+1} state.

As shown in Fig. 11, for the sequence a, b and c, the label “BBO” has a value of $1 + 10 + 2 + 4 - 2 + 11 + 1 = 27$, while the label “BIO” has a value of $1 + 10 + 4 + 3 + 2 + 11 + 1 = 32$. On this basis, CRF enumerates all the possible sequences and selects the label sequence with the highest value based on the maximum tag sequence probability. The label sequence probability is

$$P(y|w) = e^{C_i} / \sum_{i=0}^n e^{C_i}, \quad (2)$$

where w is the input character sequence and y is the output label sequence. C_i is the value of each sequence, and n is the number of all possible sequence labels. Considering both methods, a combination of BERT and CRF can solve the issue correctly.

In this study, we employ a CRF model to replace the softmax output layer of the BERT model to extract two types of entities (line names and manufacturer names) from electrical equipment malfunction defect text, as illustrated in Fig. 10.

The BERT-CRF model first converts the input electrical equipment malfunction and defect text into the standard processing format for the BERT model to obtain the Token, Segment, and Position information for each character. Then, each layer of the encoder network transfers the learned features to the next layer of the Encoder to learn the information. The final feature information is input to the CRF model iteratively. Finally, the CRF model calculates the sequence label with the maximum probability based on the feature information to obtain a label corresponding to each character.

When the BERT-CRF model is used to extract the two types of entities (i.e., line and manufacturer names), it is first necessary to manually label electrical equipment malfunction and defect texts to provide the label corresponding to

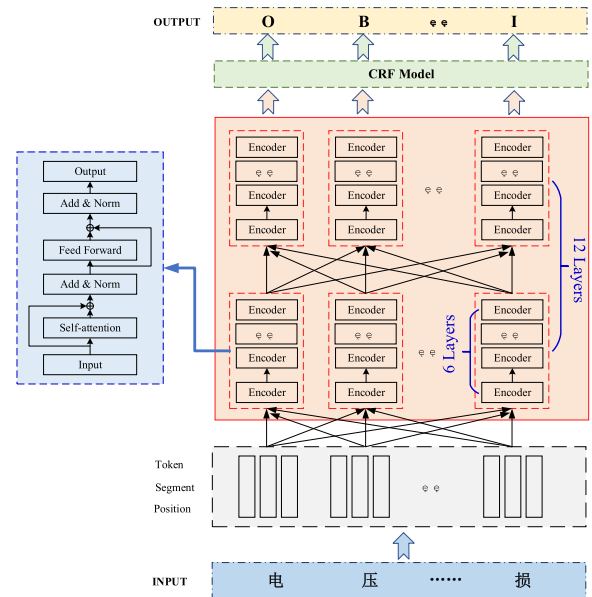


FIGURE 10. Structure of the BERT-CRF-based Class III entity extraction algorithm.

each character. Then, by learning the features in the text, BERT-CRF can predict the label for each character and identify line and manufacturer names in unseen electrical equipment malfunction and defect texts. A simple labeling example is shown in Fig. 11.

In Fig. 11, the line name is represented by LINE, and the manufacturer name is represented by ORG. In detail, the two entities “红路线 (Honglu line)” and “某一变压器厂 (A transformer factory)” to be extracted are marked as “红B-LINE路I-LINE线I-LINE” and “某B-ORG—I-ORG变I-ORG压I-ORG器I-ORG厂I-ORG”, and the remaining nonentities are denoted by “O”.

By training using these labeled texts, BERT-CRF learns the text features and finds the labeling rules for each character. Then, the labels corresponding to each character are predicted from the test text based on the learned labeling rules. Finally, the entity labels are extracted from all the labels and compared with the labeling situation.

VII. SIMULATION AND RESULTS ANALYSIS

A. EXPERIMENTAL PARAMETER CONFIGURATION AND EVALUATION INDICATORS

In this study, 2,000 electrical equipment malfunction and defect text records are selected to test the entity extraction results. To test the line and manufacturer name extraction, we used 1,600 text records as the training set for the model to learn text features and used 400 as the test set to test the model performance.

The extracted entities are grouped into correct entities TP , lost entities TN and incorrect entities FN , where TP denotes the entities correctly identified by the algorithm, TN denotes entities unrecognized by the algorithm, and FN denotes entities incorrectly identified by the algorithm. On this basis,

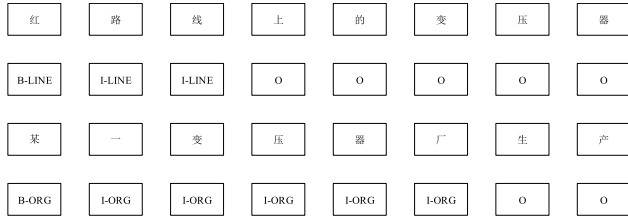


FIGURE 11. Entity tag.

three indexes of entity extraction involving accuracy rate P , recall rate R , and F1-value can be defined as follows.

$$P = TP / (TP + TN) \times 100\%, \quad (3)$$

$$R = TP / (TP + FN) \times 100\%, \quad (4)$$

$$F1 = 2P \cdot R / (P + R) \times 100\%. \quad (5)$$

The entity extraction performance is analyzed by comprehensively considering the scores of the above three indicators.

The proposed BERT-CRF mode uses Adam for algorithm optimization. The Python language is used for programming, and the algorithm is implemented on the TensorFlow framework. The initial learning rate is $\eta = 5e - 5$, the dropout rate is 0.1, and the sizes of the training and test batches are 32 and 8, respectively. In addition, the number of multihead attention tasks in the BERT model is 12. The text feature vector dimension is 768, and the number of fully connected layers connected to the CRF is 7. The specific model parameters are listed in TABLE 1.

B. ANALYSIS OF THE EXPERIMENTAL RESULTS OF THE EXTRACTION ALGORITHM FOR CLASS III ENTITIES

The BERT-CRF algorithm for Class III entity extraction process is divided into word vector representation and feature classification. The results of these two aspects are analyzed in the following subsections.

1) ANALYSIS OF TEXT VECTORIZATION RESULTS

The most critical step in the text processing task is text vectorization. The quality of the downstream tasks is generally determined by the word vectors obtained by the BERT model training text. Word vectors represent the semantic meaning contained in the text and can be evaluated with semantic-related tasks. In this paper, a word similarity task is applied to calculate the similarity between two words. The similarity score is used as a measurement indicator whose calculation formula is

$$\begin{aligned} \text{similarity} = \cos \theta &= \frac{A \cdot B}{\|A\| \|B\|} \\ &= \frac{\left[\sum_{i=1}^n A_i \cdot B_i \right]}{\left[\sqrt{\sum_{i=1}^n (A_i)^2} \cdot \sqrt{\sum_{i=1}^n (B_i)^2} \right]}, \end{aligned} \quad (6)$$

TABLE 1. Model parameters.

Parameter	Parameter value
Programming language	Python
Algorithm framework	TensorFlow
Optimizer	Adam
Initial learning rate	5e-5
Discard rate	0.1
Training batch	303
Test batch	8
Number of iterations	30
Multihead-attention	12
Vector dimension	768
CRF fully connected layers	7

TABLE 2. Word similarity.

—	茶南线 (Chanan line)	茶北线 (Chabei line)	生产厂家甲 (Manufacturer A)	生产厂家乙 (Manufacturer B)
茶南线 (Chanan line)	1.00	0.98	0.74	0.76
茶北线 (Chabei line)	0.98	1.00	0.75	0.78
生产厂家甲 (Manufacturer A)	0.74	0.75	1.00	0.92
生产厂家乙 (Manufacturer B)	0.76	0.78	0.92	1.00

where A and B are two-word vectors obtained by NLP model training, n is the word vector dimension, and A_i and B_i are the values corresponding to each dimension.

The line names “茶南线 (Chanan line)” and “茶北线 (Chabei line)” as well as the manufacturer names “生产厂家甲 (Manufacturer A)” and “生产厂家乙 (Manufacturer B)” are used to test the similarity and analyze the performance of the BERT-CRF model for extracting the electrical equipment malfunction text features. In particular, the manufacturer name that appears in this paper refers to the name of the specific manufacturer but with significant differences. The cross-validation experimental results are listed in TABLE 2.

TABLE 2 shows that the similarity between “茶南线 (Chanan line)” and “茶北线 (Chabei line)” reaches 0.98, indicating that the line names are highly correlated and are likely to refer to the same line. Furthermore, the similarity between the manufacturer names “生产厂家甲 (Manufacturer A)” and “生产厂家乙 (Manufacturer B)” is 0.92. However, the similarity between the line names and the manufacturer names is below 0.8, indicating a weaker correlation between these two entities. The experimental results show high similarity between entity words in the same category and low similarity between entity words in different categories. These results are consistent with the actual situation and indicates that the BERT model is effective.

2) ANALYSIS OF THE CLASS III ENTITY RECOGNITION RESULTS

Based on the BERT model, this paper replaces the softmax output layer of the original BERT model with a machine learning model, CRF, and proposes the combined BERT-CRF

TABLE 3. Comparison of experimental results.

Model	Class III entities	Accuracy	Recall rate	F1-value
CRF	Line names	93.57%	90.86%	92.20%
	Manufacturer names	90.75%	88.87%	89.80%
BERT	Line names	95.29%	94.45%	94.88%
	Manufacturer names	94.18%	93.63%	93.91%
BERT-CRF	Line names	95.64%	94.88%	95.26%
	Manufacturer names	94.85%	94.08%	94.46%

model to extract two types of entities (line names and manufacturer names) from texts containing electrical equipment malfunction descriptions. To verify the effectiveness of this method, we compare the results obtained by the CRF, BERT, and BERT-CRF methods. The experimental results of these three models are listed in TABLE 3.

Compared with the CRF model and the BERT model, the BERT-CRF model improves the experimental results for both line names and manufacturer names to a certain extent. Specifically, the F1-value of the BERT-CRF model for line name recognition reaches 95.26%, 3.06% higher than that of the CRF model, and 0.35% higher than that of the BERT model. In addition, the F1-value for manufacturer name recognition can be as high as 94.46%, which is 4.66% higher than that of the CRF model and 0.55% higher than that of the BERT model. The experimental results show that the BERT-CRF model proposed in this paper effectively extracts line names and manufacturer names from electrical equipment malfunction and defect texts. In addition, CRF, BERT, and BERT-CRF all achieve better results (accuracy rate, recall rate, and F1-value) when recognizing line names than when recognizing manufacturer names. Specifically, for the CRF model, the accuracy rate, recall rate, and F1-value of line name recognition are 2.82%, 1.99%, and 2.40% higher than the respective values for manufacturer name recognition (for the BERT model, they are 1.11%, 0.82%, and 0.98% higher, respectively, and for the BERT-CRF model, they are 0.79%, 0.80%, and 0.80% higher, respectively). By analyzing the electrical equipment malfunction and defect texts, we found that the text characteristics of line names are more evident than those of the manufacturer names. Therefore, the model is more likely to learn the line name characteristics, making the line name recognition effect is better than that for manufacturer names.

3) ANALYSIS OF ENTITY EXTRACTION RESULTS

By analyzing the electrical equipment malfunction and defect texts and the characteristics of various entities, this paper proposes three methods based on dictionaries, LTP tools, and BERT-CRF models to extract entities with different characteristics. The dictionary-based method extracts five types of entities: equipment name, component name, malfunction type, malfunction level, and voltage level. The LTP

2018年11月25日配电运维小组检测到10kV黄茅咀线上的配电变压器的配变套管有明显裂缝、损伤,已经影响安全运行,可初步判断设备损坏是由外部影响造成,属紧急级别。调查后得知此设备由生产厂家丙于2010年12月10日生产,在2011年02月19日投入使用。

FIGURE 12. Electrical equipment malfunction and defect text records.

TABLE 4. Entity results extracted from the sample text.

Entity category	Entity
Equipment name	变压器 (Transformer)
Component name	配变套管 (Distribution transformer casing)
Malfunction type	外部影响 (External influence)
Malfunction level	紧急 (urgent)
Voltage level	10kV
Time	2018年11月25日 (November 25, 2018)
	2010年12月10日 (December 10, 2010)
	2011年02月19日 (February 19, 2011)
Line names	黄茅咀线 (Huangmaozu line)
Manufacturer names	生产厂家丙 (Manufacturer C)

TABLE 5. Extraction results for eight types of entities.

Extraction method	Entity category	Accuracy	Recall rate	F1-value
Dictionary matching	Equipment name	100%	100%	100%
	Component name	100%	100%	100%
	Malfunction type	100%	100%	100%
	Malfunction level	100%	100%	100%
	Voltage level	100%	100%	100%
LTP tool	Time	99.91%	97.84%	98.86%
	Line names	95.64%	94.88%	95.26%
BERT-CRF	Manufacturer names	94.85%	94.08%	94.46%

tool-based method extracts time entities, and the BERT-CRF model extracts two types of entities: line names and manufacturer names. Taking the record of electrical equipment malfunction and defect texts shown in Fig. 12 as an example, the results of the 8 types of entities extracted by these three methods from the text records are shown in TABLE 4. The runtime performances of the three employed methods are listed in TABLE 5.

The dictionary-based entity extraction achieves a good extraction result because of the small vocabulary. The accuracy rate, recall rate, and F1-value for the five types of extracted entities (equipment name, component name, malfunction type, malfunction level, and voltage level) all reached 100%. In addition, subsequent expansion of this method is straightforward because terms can be directly added to the dictionary. LTP is a mature time entity extraction tool that achieves excellent performance. For electrical equipment malfunction and defect texts, the accuracy rate, recall rate, and F1-value of time entity extraction can be as high as 99.91%, 97.84%, and 98.86%, respectively, indicating that the tool is appropriate for this application.

Specifically, the proposed BERT-CRF model combines the advantages of the BERT and CRF models. The accuracy

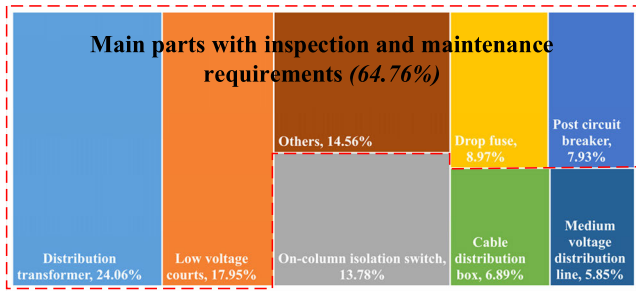


FIGURE 13. Frequency diagram of electrical equipment malfunction.

rate, recall rate, and F1-value for the line name entity extraction are 95.64%, 94.88%, and 95.26%, respectively. For manufacturer names, these metrics are 94.85%, 94.08%, and 94.46%, respectively. Compared with the CRF algorithm or BERT algorithm alone, the BERT-CRF model achieves comprehensive improvements in accuracy, recall rate, and F1-value, as discussed in TABLE 3, and could meet the use requirements.

In general, the cascaded entity extraction method designed in this paper is sufficient to enable the processing of electrical equipment malfunction texts and provides an automate extraction solution to solve the problems caused by various entity categories and large feature differences.

C. STATISTICAL ANALYSIS OF ELECTRICAL EQUIPMENT MALFUNCTIONS AND DEFECTS

Based on the entity extraction results, we further analyzed the causes of electrical equipment malfunctions to provide prior knowledge for electrical equipment operation and maintenance.

1) EQUIPMENT NAME

First, among the wide varieties of equipment, we investigated the most vulnerable equipment. The experimental results are shown in Fig. 13.

Figure 13 shows that “配电变压器 (distribution transformer)”, “低压台区 (low-voltage transformer area)” and “柱上隔离开关 (the pole disconnecting switch)” are the equipment types most prone to breakdowns.

2) COMPONENT NAME

The malfunction frequencies of various components are illustrated in Fig. 14. It can be observed that electrical equipment components associated with “本体 (ontology)”, “接头 (linker)”, “导线 (traverse)” and “构架 (framework)” are more prone to failure.

3) MANUFACTURER NAMES

Malfunctions of power equipment and components are closely related to families of defects. If the electrical equipment and components are not up to standard when produced, they will often fail after being deployed. Fig. 15 shows the manufacturers whose equipment is most prone to malfunction. The results show that equipment produced by

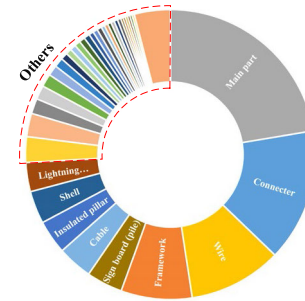


FIGURE 14. Component malfunction frequency diagram.

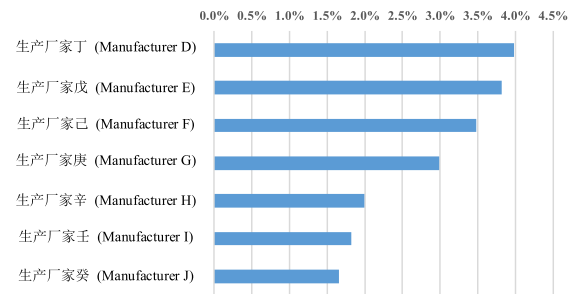


FIGURE 15. Frequencies of equipment malfunctions from specific manufacturers.

“生产厂家丁 (Manufacturer D)” and “生产厂家戊 (Manufacturer E)” is the most vulnerable to malfunctions. However, all the electrical equipment and components show a low and acceptable malfunction rate.

4) MALFUNCTION TYPES

Various factors can cause equipment malfunctions during operation. The causes of malfunctions can generally be grouped into 7 categories. As shown in Fig. 16, nearly half of equipment malfunctions are the result of aging equipment (i.e., 设备老化 (equipment aging)); indicating that timely equipment replacement is essential. External influences are the second-largest factor. Equipment malfunctions are inseparable from the local geographical environment and temperature. We found that equipment associated with the “龙山线 (Longshan line)”, “双江线 (Shuangjiang line)”, and “石湖线 (Shihu line)” are most prone to malfunctions.

5) RELATIONSHIP BETWEEN SERVICE TIME AND MALFUNCTION FREQUENCY

At the initial stage of an operation, problems may appear in each component that can lead to further equipment malfunction. The equipment gradually becomes stable after a run-in period, except for occasional malfunctions that occur due to weather conditions, human operations, etc. As the service time increases, the number of equipment malfunctions induced by aging will also increase, as shown in Fig. 17. It is important to emphasize that the time scale covered by the dataset used in this study does not include the depletion period. During the first 3 years of operation, the frequency

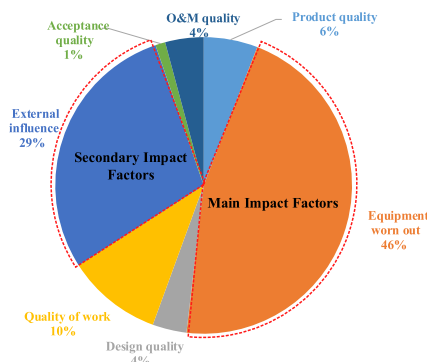


FIGURE 16. Formation of malfunction types.

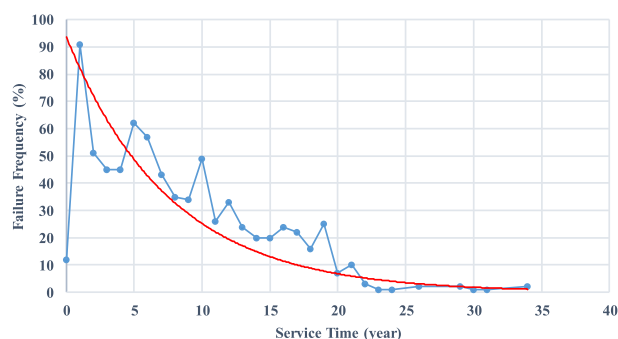


FIGURE 17. Relationship between service time and malfunction frequency.

of equipment malfunctions is high. Nevertheless, as time increases, the frequency of malfunctions tends to stabilize.

VIII. CONCLUSION

In this paper, we proposed a hybrid NLP algorithm to perform named entity extraction from electrical equipment malfunction texts; then, we evaluated malfunction texts based on entity extraction. First, word segmentation and standardization were applied to electrical equipment malfunction and defect texts. Then, three types of entities with different features were extracted using dictionary-based matching, the LTP tool, and the proposed BERT-CRF model. In particular, this paper concentrates on the extraction process for line and manufacturer names. The F1-values for the two types of entities reached 95.26% and 94.46%, respectively. In addition, the F1-value for the line and manufacturer names obtained by the proposed combined BERT-CRF model represented increases of 3.06% and 4.66% over the CRF model alone and increases of 0.35% and 0.55% over the BERT model alone. In addition, the experimental results show that the cascaded entity extraction method designed in this paper achieves good performance on electrical equipment malfunction texts. It also provides some reference results that can be useful for the operation and maintenance of power facilities.

REFERENCES

[1] J. Zhan, J. Huang, L. Niu, X. Peng, D. Deng, and S. Cheng, "Key technologies of electric power big data and its application prospects in smart grid," *Proc. CSEE*, vol. 35, no. 3, pp. 503–511, 2015, doi: 10.13334/j.0258-8013.pcsee.2015.03.001.

[2] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *Proc. IREP Symp. Bulk Power Syst. Dyn. Control-IX Optim., Secur. Control Emerg. Power Grid*, Rethymno, Greece, Aug. 2013, pp. 1–9, doi: 10.1109/IREP.2013.6629368.

[3] J. van den Hoven, "Information resource management: Foundation for knowledge management," *Inf. Syst. Manage.*, vol. 18, no. 2, pp. 80–83, Mar. 2001.

[4] H. Wang, J. Cao, and L. Luo, "Current status and challenges of power text data mining," *Zhejiang Electr. Power*, vol. 38, no. 3, pp. 1–7, 2019, doi: 10.19585/j.zjdl.201903001.

[5] S. Chen and X. Ouyang, "Overview of named entity recognition technology," *Radio Commun. Technol.*, vol. 46, no. 3, pp. 251–260, 2020, doi: 10.3969/j.issn.1003-3114.2020.03.001.

[6] M. Maggini, G. Marra, S. Melacci, and A. Zugarini, "Learning in text streams: Discovery and disambiguation of entity and relation instances," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4475–4486, Nov. 2020, doi: 10.1109/TNNLS.2019.2955597.

[7] D. Song, F. Schilder, S. Hertz, G. Saltini, C. Smiley, P. Nivarthi, O. Hazai, D. Landau, M. Zaharkin, T. Zielund, H. Molina-Salgado, C. Brew, and D. Bennett, "Building and querying an enterprise knowledge graph," *IEEE Trans. Services Comput.*, vol. 12, no. 3, pp. 356–369, May 2019, doi: 10.1109/TSC.2017.2711600.

[8] Y. Lan, S. Wang, and J. Jiang, "Knowledge base question answering with a matching-aggregation model and question-specific contextual relations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1629–1638, Oct. 2019, doi: 10.1109/TASLP.2019.2926125.

[9] J. Xu, H. He, X. Sun, X. Ren, and S. Li, "Cross-domain and semi-supervised named entity recognition in chinese social media: A unified model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2142–2152, Nov. 2018, doi: 10.1109/TASLP.2018.2856625.

[10] J. Kim, Y. Ko, and J. Seo, "A bootstrapping approach with CRF and deep learning models for improving the biomedical named entity recognition in multi-domains," *IEEE Access*, vol. 7, pp. 70308–70318, 2019, doi: 10.1109/ACCESS.2019.2914168.

[11] H. Yu, Y. Cao, G. Cheng, P. Xie, Y. Yang, and P. Yu, "Relation extraction with BERT-based pre-trained model," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Limassol, Cyprus, Jun. 2020, pp. 1382–1387, doi: 10.1109/IWCMC48107.2020.9148384.

[12] A. Tan, "Text mining: Promises and challenges," in *Proc. South East Asia Res. Comput. Confederation (SEARCC)*, Singapore, 1999. [Online]. Available: https://www.researchgate.net/publication/2408427_Text_Mining_Promises_And_Challenges

[13] Z. Ji, X. Wang, C. Cai, and H. Sun, "Power entity recognition based on bidirectional long short-term memory and conditional random fields," *Global Energy Interconnection*, vol. 3, no. 2, pp. 186–192, 2020, doi: 10.1016/j.gloi.2020.05.010.

[14] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.

[15] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: 10.1109/TKDE.2020.2981314.

[16] Y. Liu and W. Che, "Exploring segment representations for neural semi-Markov conditional random fields," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 813–824, 2020, doi: 10.1109/TASLP.2020.2964960.

[17] X. Yin, Y. Huang, B. Zhou, A. Li, L. Lan, and Y. Jia, "Deep entity linking via eliminating semantic ambiguity with BERT," *IEEE Access*, vol. 7, pp. 169434–169445, 2019, doi: 10.1109/ACCESS.2019.2955498.

[18] L. Liu and D. Wang, "A review on named entity recognition," *J. China Soc. Sci. Tech. Inf.*, vol. 37, no. 3, pp. 329–340, Mar. 2018, doi: 10.3772/j.issn.1000-0135.2018.03.010.

[19] S. Sun, Z. Dai, X. Xi, X. Shan, and B. Wang, "Power fault preplan text information extraction based on NLP," in *Proc. IEEE Int. Conf. Saf. Produce Informatization (HICSPI)*, Chongqing, China, Dec. 2018, pp. 617–621, doi: 10.1109/HICSPI.2018.8690379.

[20] H. Wang, Z. Liu, Y. Xu, X. Wei, and L. Wang, "Short text mining framework with specific design for operation and maintenance of power equipment," *CSEE J. Power Energy Syst.*, early access, May 21, 2020, doi: 10.17775/CSEEJPES.2019.01120.

[21] H. Wang, J. Cao, and D. Lin, "Deep analysis of power equipment defect based on semantic framework text mining technology," *CSEE J. Power Energy Syst.*, early access, Oct. 7, 2019, doi: 10.17775/CSEEJPES.2019.00210.

[22] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," *Proc. 26th AAAI Conf. Artif. Intell.*, 2016, pp. 2659–2665.

- [23] P. C. Piaoran Chen, J. Ponocko, N. Milosevic, G. Nenadic, and J. V. Milanovic, "Towards application of text mining for enhanced power network data analytics—Part I: Retrieval and ranking of textual data from the Internet," in *Proc. Medit. Conf. Power Gener., Transmiss., Distrib. Energy Convers. (MedPower)*, Belgrade, Serbia, 2016, pp. 1–8, doi: [10.1049/cp.2016.1076](https://doi.org/10.1049/cp.2016.1076).
- [24] J. Ni and R. Florian, "Improving multilingual named entity recognition with wikipedia entity type mapping," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1275–1284, doi: [10.18653/v1/D16-1135](https://doi.org/10.18653/v1/D16-1135).
- [25] H. Tang, H. Wang, Z. Zhang, and X. Wang, "Extracting names of historical events based on Chinese character tags," *Data Anal. Knowl. Discovery*, vol. 2, no. 7, pp. 89–100, 2018, doi: [10.11925/infotech.2096-3467.2018.0057](https://doi.org/10.11925/infotech.2096-3467.2018.0057).
- [26] C. Xie, G. Zou, H. Wang, and Y. Jin, "A new condition assessment method for distribution transformers based on operation data and record text mining technique," in *Proc. China Int. Conf. Electr. Distribution (CICED)*, Aug. 2016, pp. 1–7, doi: [10.1109/CICED.2016.7576179](https://doi.org/10.1109/CICED.2016.7576179).
- [27] B. Wei, S. Liqin, L. Chengyu, L. Huafeng, X. Xiaohua, and R. Lixiang, "Power system text information matching research based on deep learning," in *Proc. IEEE Innov. Smart Grid Technol.-Asia (ISGT Asia)*, Chengdu, China, May 2019, pp. 1181–1186, doi: [10.1109/ISGT-Asia.2019.8881815](https://doi.org/10.1109/ISGT-Asia.2019.8881815).
- [28] Y. Jin, J. Xie, W. Guo, C. Luo, D. Wu, and R. Wang, "LSTM-CRF neural network with gated self attention for chinese NER," *IEEE Access*, vol. 7, pp. 136694–136703, 2019, doi: [10.1109/ACCESS.2019.2942433](https://doi.org/10.1109/ACCESS.2019.2942433).
- [29] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423/>
- [30] *Reliability Evaluation Code for Transmission and Distribution Installation*, document DL/T 837, 2012.
- [31] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2733–2742, Oct. 2020, doi: [10.1109/JBHI.2020.3001216](https://doi.org/10.1109/JBHI.2020.3001216).
- [32] A. Vaswani, N. Shazeer, N. Parmar, and J. Uszkoreit, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.



ZHE KONG received the B.E. degree in automation from the Wuhan University of Technology, Hubei, China, where she is currently pursuing the M.A. degree in control science and engineering. Her research interests include big data systems and semantic segmentation.



CHANGXI YUE received the B.S. and M.S. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2006, respectively.

He currently leads the High-Voltage and High-Current Test Technique Group, China Electric Power Research Institute, Wuhan, China. He is also a Senior Research Engineer experienced in high-voltage and high-current tests and measurements.



YING SHI received the Ph.D. degree in marine engineering from the Wuhan University of Technology, Wuhan, Hubei, China.

She is currently a Professor of artificial intelligence with the Wuhan University of Technology. Her current research interests include big data systems, grid security, machine learning, and deep learning.



JICHENG YU (Member, IEEE) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the M.S. and Ph.D. degrees from Arizona State University, Tempe, USA, in 2013 and 2017, respectively, all in electrical engineering.

He is currently a Research Engineer with China Electric Power Research Institute, Wuhan. His research interests include sensors, smart meters, and big data analytics in the power systems.



CHANGJUN XIE (Member, IEEE) received the Ph.D. degree in vehicle engineering from the Wuhan University of Technology (WHUT), Wuhan, China, in 2009. He is currently a Professor with the School of Automation, WHUT. His research interests include battery management systems, control strategies for intelligent and connected vehicles and vehicle control, and optimization of new energy vehicles.



LINGYUN XIE received the B.S. and M.S. degrees in automation from the Wuhan University of Technology, Hubei, China, in 2017 and 2020, respectively. He is currently with the Fiftieth Research Institute, China Electronics Technology Group Corporation.

...