IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 December 2022, accepted 3 January 2023, date of publication 27 January 2023, date of current version 13 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240162

APPLIED RESEARCH

Electric Power Audit Text Classification With Multi-Grained Pre-Trained Language Model

QINGLIN MENG[1], (Member, IEEE), YAN SONG[1], JIAN MU[2], YUANXU LV[3], JIACHEN YANG[4], (Senior Member, IEEE), LIANG XU[5], JIN ZHAO[6], JUNWEI MA[7], WEI YAO[8], RUI WANG[9], MAOXIANG XIAO[10], AND QINGYU MENG[11]

[1]Comprehensive Service Center, State Grid Tianjin Electric Power Company, Tianjin 300010, China
[2]State Grid Tianjin Electric Power Company, Tianjin 300010, China
[3]State Grid Corporation of China, Beijing 100031, China
[4]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[5]Tianjin Tianyuan Power Engineering Company Ltd., Baodi Power Supply Branch, State Grid Tianjin Electric Power Company, Tianjin 301800, China
[6]Electric Power Research Institute, State Grid Shanxi Electric Power Company, Taiyuan 003001, China
[7]Information and Communication Branch, State Grid Shanxi Electric Power Company, Taiyuan 030012, China
[8]Taiyuan Power Supply Company, State Grid Shanxi Electric Power Company, Taiyuan, Shanxi 003000, China
[9]Chengxi Power Supply Branch, State Grid Tianjin Electric Power Company, Tianjin 300190, China
[10]Ningdongshengyuan Electric Power Engineering Company Ltd., Ninghe Power Supply Branch, State Grid Tianjin Electric Power Company, Tianjin 301500, China
[11]Zhangjiakou Wanquan District Power Supply Branch, State Grid Jibei Electric Power Company Ltd., Zhangjiakou, Heibei 076261, China

Corresponding author: Jiachen Yang (yangjiachen@tju.edu.cn)

ABSTRACT Electric power audit text classification is one of the important research problem in electric power systems. Recently, kinds of automatic classification methods for these texts based on machine learning or deep learning models have been applied. At present, the development of computing technology makes ''pre-training and fine-tuning'' the newest paradigm of text classification, which achieves better results than previous fully-supervised models. Based on pre-training theory, domain-related pre-training tasks can enhance the performance of downstream tasks in the specific domain. However, existing pre-training models usually use general corpus for pre-training, and do not use texts related to the field of electric power, especially electric power audit texts. This results in that the model does not learn too much electric-power-related morphology or semantics in the pre-training stage, so that less information can be used in the fine-tuning stage. Based on the research status, in this paper, we propose EPAT-BERT, a BERT-based model pre-trained by two-granularity pre-training tasks: word-level masked language model and entity-level masked language model. These two tasks predict word and entity in electric-power-related texts to learn abundant morphology and semantics about electric power. We then fine-tune EPAT-BERT for electric power audit text classification task. The experimental results show that, compared with fully supervised machine learning models, neural network models, and general pre-trained language models, EPAT-BERT can significantly outperform existing models in a variety of evaluation metrics. Therefore, EPAT-BERT can be further applied to electric power audit text classification. We also conduct ablation studies to prove the effectiveness of each component in EPAT-BERT to further illustrate our motivations.

INDEX TERMS Pre-trained language model, text classification, electric power audit text, natural language processing, masked language model.

I. INTRODUCTION

Text classification has been widely applied in electric power information processing [1], [2], [3], [4]. In researches of electric power audit information processing, the automatic and standardized classification of electric power audit texts in the form of natural language is a key problem to be solved. **On the one hand**, with the acceleration of the digitization, a large number of audit texts have been accumulated in the audit process of electric power enterprises, including the audit

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

Audit Problem Description (Used as Electric Power Audit Text in this paper)	Issue Category
The project settlement approved the completion surveying and mapping fee of 76,230 yuan (71,915.09 yuan excluding tax), approved 76,230 yuan (74,009.71 yuan excluding tax), miscalculated the contract value-added tax rate from 3% to 6%, and the settlement audit was not strict.	Project settlement management problem
In these companies, 180 project accounting adjustments were not timely, and the project accounting adjustments were not completed within 30 days after the final accounts were approved, involving 1,233 accounting adjustments and an amount of 136,696,400 Yuan.	Project final account management problem
(cd-marketing 18-04) For 3 projects including the new charging pile group in Xiaodian Town, District BC, the estimated design cost is calculated according to the "overhead line and cable line project", and the actual should be based on the "distribution station, switch station and charging (Replacement) Power Plant Project" , an artificially high 236,600 yuan.	Project design management problem
From 2016 to 2020, three companies jz, nh, and wq have completed agricultural power grid transformation and upgrading projects. Among them, the estimated budget of 10 projects has changed by more than 40% compared with the settlement, and the preliminary design depth of the project is insufficient, involving a difference of 10.4606 million yuan.	Bidding management

FIGURE 1. Several examples of provided audit problem descriptions (used as electric power audit texts in this paper) and their corresponding categories.

problem description, problem category, referenced provision, and audit opinions recorded manually by auditors. These texts are obviously unstructured and ambiguous. Influenced by the auditors' personalized language expression and subjective judgment, manual classification will leads to low efficiency and insufficient accuracy. Therefore, how to efficiently classify these texts with high accuracy is the practical demand of power enterprises to improve the efficiency of audit information processing. **On the other hand**, the audit texts of electric power enterprises are short texts in specific fields, which have distinct industry characteristics such as high text similarity and fuzzy classification boundary. They are different from general languages. Therefore, the direct application of existing text classification models can not consider the features of the domain-specific electric power audit texts. Existing models should be further improved to adapt to these features and improve the classification effectiveness.

The electric power audit text classification is a standard text multi-classification task. Given the audit problem description, we need to predict a corresponding category. Figure 2 shows several samples of electric power audit texts, each of which includes an audit problem description and a corresponding category. As can be seen from Figure 2, different audit problem description corresponds to different issue categories. For general audit departments, there are usually dozens of categories in total, each category corresponds to a large number of audit problem descriptions recorded by the audit department previously, so that it is unnecessary to consider the problem of sample imbalance, such as few-shot learning, dataset sampling, etc.

Text classification based on machine learning and neural network algorithms has been paid attention [5]. Many text classification models including RNN [6], LSTM [7], and FastText [29] etc. have also been gradually applied to **the processing of texts related to electric power audit**. For example, Chen et al. [8] used category mixed embedding method to classify power texts hierarchically. Zhao et al. [9] used classic TF-IDF and word vector technology to classify the power audit text. Chen et al. [10] introduced a professional dictionary for the audit field and classified the audit text using the bi-directional recurrent neural network BiLSTM. Feng et al. [11] further introduced the attention mechanism on the basis of BiLSTM to mine the defect text of power equipment. The development of these related work illustrates that, first, text classification is developing from machine learning to **deep learning** [9], [11]. Second, **domain expertise** is very important for text information mining [10] and should be further integrated into the deep learning model to improve the performance of downstream tasks.

In recent years, ''pre-training and fine-tuning'' paradigm has gradually become the latest research direction of text classification. Compared with the previous fully-supervised neural network models, pre-trained models can achieve better results in various natural language processing tasks [12]. However, existing pre-trained models such as BERT [13] and ERNIE [14] are all pre-trained using common corpora, such as Wikipedia data, and do not use texts related to the electric power field, especially the electric power audit field, for pre-training. Intuitively, we deem that corpus related to electric power domain is closer to the semantic domain of the power audit text classification task. From the perspective of pre-training theory [24], [25], **domain-related pre-training tasks can enhance the performance of domain-related downstream tasks**. Therefore, in this paper, we aim to improve electric power audit text classification task by modifying pre-training tasks of a Pre-trained Language Model (PLM). To achieve this, we propose **two granularity of power audit text pre-training tasks**: word-granularity masked language model (WMLM) and entity-granularity masked language model (EMLM). These two pre-training tasks use large-scale power text as training corpus, and let the model complete word-granularity prediction and entity-granularity prediction, so as to leverage the morphology, grammar and related knowledge in the power text. Based on these two pre-training tasks, we proposes a BERT-based model EPAT-BERT (Electric Power Audit Text-BERT) for power audit text classification. We evaluate and compare EPAT-BERT with strong baseline models to prove its effectiveness in electric power audit text classification task, and then conduct ablation studies to illustrate the effectiveness of each component of EPAT-BERT, including the influence of two pre-training tasks, and the order of them.

II. ELECTRIC POWER AUDIT TEXT CLASSIFICATION WITH PRE-TRAINED LANGUAGE MODEL

A. ELECTRIC POWER AUDIT TEXT CLASSIFICATION

Text classification is one of the basic tasks in natural language processing. As a kind of natural language text, electric power audit text is recorded by the auditors of electric power

enterprises, which is of great significance for the enterprises to conduct audit works. Audit texts usually contain audit contents and methods, audit concerns, audit findings, referenced provision, audit opinions, problem classification, and other information manually recorded by the auditors. Several common audit texts are shown in Figure 2. Each audit text requires auditors to manually mark a classification labels to achieve the classification of audit texts. However, large-scale labeling of classification labels manually means consuming manual labour and material resources, and is inefficient and prone to be error. Therefore, efficient automatic classification of electric power audit texts has become an urgent problem to be solved.

Before the emergence of ''pre-training and fine-tuning'' paradigm, traditional natural language processing tasks usually required a large-scale fully-supervised training dataset to achieve end-to-end neural network training. Given the data $x = \{w1, w2, \ldots, w_n\}$ and the corresponding category label y, the model needs to learn the conditional probability distribution $p(y|x)$ from the known data. In general, the model may be a machine learning classifier or a deep neural network which maps the input (a piece of natural language) into the output (a class label). These methods achieve satisfying results in many basic classification tasks, like sentiment analysis, spam detection, and face recognition.

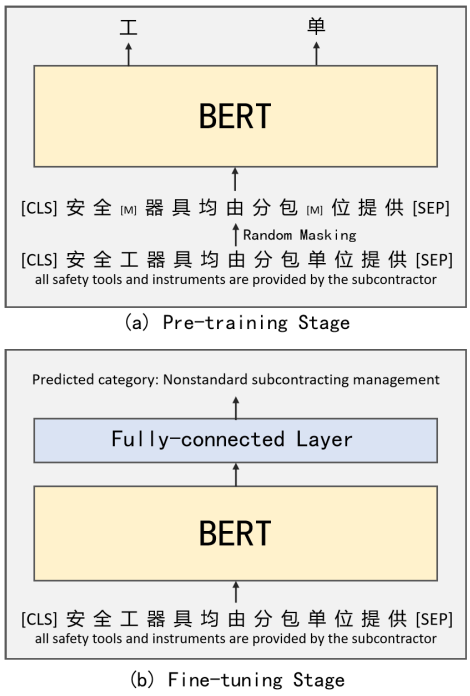


FIGURE 2. Two Stages of the Pre-trained Language Model BERT. The meaning of Chinese input in the figure is "all safety tools and instruments are provided by the subcontractor".

B. PRE-TRAINING

With the advance of natural language processing model BERT [13], computer vision model MAE [15] and cross modal retrieval model CLIP [16], **pre-trained language model (PLM)** and **fine-tuning** have become one of the important research fields in all kinds of research fields. The meaning of pre-training is to design a training task which is not directly related to the downstream task but can learn internal information of a language from large-scale general corpus. The meaning of fine-tuning is to use the pre-trained model to train the downstream tasks again. The earliest pre-training models focused on obtaining the semantics of a single word and obtaining its word embedding [5], [17]. Later, the emergence of models such as CoVe [18] and ELMo [19] made it possible to extract contextual features. With the emergence of Transformer network [20], emerging models such as BERT [13] and GPT [21], [22] have made ''pre-training and fine-tuning'' a new paradigm of solving natural language processing tasks [23]. One advantage of this model is that since the model has learned a large amount of morphology and semantic information in the pre-training stage, only a small amount of fully-supervised data is required for re-training the model in the fine-tuning stage, and it is experimentally proved that PLMs can achieve better results than non-pre-trained neural model [13], [21], [22].

BERT [13] model is a classic PLM, which uses the encoder of Transformer network [20] as the basic structure, as shown in Figure 2 (a). The BERT model takes a sentence as input, for example, assuming that the input sentence is''all safety tools and instruments are provided by the subcontractor''. The model will automatically add a special token ''[CLS]'' before this sentence to indicate the beginning of this sentence, and add an ''[SEP]'' token after this sentence to indicate the end of this sentence. Then, the model converts the input into an ID sequence, obtains the sequence of corresponding word vectors, and then encodes the word vector sequence to obtain the contextual output corresponding to each word. The original BERT model designs two pre-training tasks: masked language model (MLM) and next sentence prediction (NSP). As shown in Fig. 2 (a), the MLM task masks part of tokens in the input sentence ([M] for a special token ''[MASK]''), and then lets the model predict which token should be filled in the masked position. The NSP task combines the two sentences A and B, and lets the model judge whether A is followed by B. **After pre-training, the BERT model can be seen as a text encoder, which maps semantically similar texts to similar feature spaces, while texts with large semantic difference will be far away in the feature space after BERT encoding**.

C. FINE-TUNING

Taking BERT as an example, its pre-training task MLM is defined as predicting a masked token in the input sequence, which is totally different from downstream tasks like text classification, sentence similarity calculation, and part-of-speech tagging. However, it is deemed that the pre-trained model can still learn general language structure, such as Chinese morphology and grammar in the pre-training stage. When the model uses additional data of downstream tasks for further training, the parameters in the network will change

slightly on the original basis of pre-trained parameters. This process is called "fine-tuning". As shown in Fig. 2 (b), for different downstream tasks, such as text classification, text generation and reading comprehension in natural language processing, a learnable neural network layer can be added downstream of the pre-training model, so that the pre-trained parameters and the newly introduced parameters of the newly-introduced layer can be trained together. In this paper, for the electric power audit text classification task, the input is a description of the audit problem. The sentence is encoded by BERT, then the BERT outputs a vector representation of the sentence. Finally, the vector is then transformed by a full-connected layer to predict the corresponding category label. This is a typical downstream task of text classification. **After fine-tuning, original parameters in BERT and newly introduced parameters of the fully-connected layer can be trained together, so that the fine-tuned BERT becomes a powerful text classifier.**

III. EPAT-BERT: A MULTI-GRANULARITY PRE-TRAINED LANGUAGE MODEL FOR ELECTRIC POWER AUDIT TEXT CLASSIFICATION

Existing pre-trained language models such as BERT can be further fine-tuned to complete the text classification task. However, for the field of electric power audit, there is no suitable and universal pre-trained language model and pre-training task. As a result, there is no domain-specific pre-trained model in electric power audit text, so there is still much room for improvement in the task of electric power audit text classification. Recently, several relevant studies have shown that, domain-specific pre-training tasks can improve the down-streaming tasks. For example, LawFormer [26] is a LongFormer [27] encoder that is pre-trained with Chinese legal provisions, SciBERT [28] is a BERT model trained with scientific texts. In this paper, we introduce relevant texts in the electric power field to improve the effectiveness of downstream electric power audit text classification. Therefore, for this situation, we propose two pre-training tasks related to electric power audit texts, and proposes a pre-trained language model **EPAT-BERT**, for electric power audit text classification.

A. DESIGN OF PRE-TRAINING TASKS

As the downstream task, electric power audit text classification has been clarified. On the other hand, how to design robust and reasonable pre-training tasks will be the key to improve the classification ability. In recent years, For electric power audit texts, first of all, this paper uses **word-granularity masked language model** of the original BERT [13] model as one of the pre-training tasks, but the pre-training text should be adjusted from the Chinese Wikipedia used by BERT to electric-power-related text collected from the Internet. We do this so that the model can learn more morphology and semantics related to electric power contents and is closer to the downstream audit text classification task.

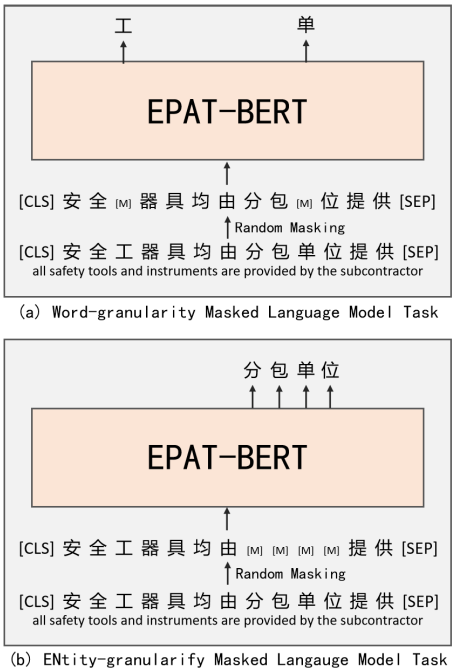


FIGURE 3. Two pre-training tasks of EPAT-BERT. The meaning of Chinese input in the figure is "all safety tools and instruments are provided by the subcontractor".

In addition, compared with general texts, electric power texts will contain more professional terms, concepts, and representations, which always needs to be presented more accurately than general texts. Only using word-granularity masked language model is inaccurate [14]. Therefore, the **entity-granularity masked language model** is designed in this paper. The model not only predicts the masked *words* in the pre-training stage, but also masks the *entities* composed of multiple words or phrases, and then predicts them. This process allows the model to learn entity-granularity knowledge that related to electric power audit, like some long concepts that rarely occur in general texts, not just limited to general morphology and semantics. In the following subsections, we will introduce the two-granularity pre-training tasks in detail, then explain how to construct, train, and evaluate the EPAT-BERT model.

B. WORD-GRANULARITY MASKED LANGUAGE MODEL

As shown in Figure 3 (a), staying consistent with BERT model, the **word-granularity masked language model** task randomly selects 20% of the Chinese characters in a paragraph of text to mask, and then uses the output vectors corresponding to the mask positions to let the model predict the Chinese characters. In Figure 3 (a), "[M]" represents a special mask token "[MASK]". Since the pre-training corpus is changed from general Chinese texts to electric-power-related texts, the model can learn the vocabulary and grammar information more relevant to electric power in the pre-training stage, so that it can theoretically achieve better results in the downstream tasks related to electric power texts.

In order to train the word-granularity masked language model, it is necessary to set a word-granularity loss function at the corresponding position of each masked Chinese character for optimization. Consistent with the BERT [13] model, the *cross entropy* loss function is selected in this paper.

C. ENTITY-GRANULARITY MASKED LANGUAGE MODEL

Electric power audit text are usually highly professional short texts, in which the entities and knowledge related to electric power audit industry often appear, yet these entities and knowledge do not appear frequently in general texts. The existing research [14] shows that, for this kind of text, there will be severe inaccuracy problem in the training stage of single word-granularity masked language model task. For example, for the masked sentence ''the second largest city in China is the mask in [MASK] [MASK]'', it is prone to predict incorrect cities, because the content to be predicted in this sentence is knowledge related, while the word-granularity masked language model task pays more attention to lexical information when prediction, and sometimes ignores these knowledge information.

In order to make up for the defects of the word-granularity masked language model, in this paper, we propose an **entity-granularity masked language model** task. Specifically, as shown in Figure 3 (b), unlike the word-granularity random mask, EPAT-BERT first identifies the entity part in the sentence according to a professional vocabulary and syntax analysis toolkit in the electric power field. Then these entities are masked one by one randomly. When the amount of mask exceeds 20% of the total length of the sentence, the mask process will be stopped. Through this entity-granularity masking method, the content that the model needs to predict during pre-training process not only contain morphology or semantics contained in words, but also to learn the corresponding facts or knowledge in the texts. This is helpful for the model to further understand the text in a higher perspective, especially for electric power audit text, which is highly integrated with professional knowledge.

D. MODEL CONSTRUCTION AND TRAINING

1) MODEL TRAINING IN THE PRE-TRAINING STAGE

In the pre-training stage, the input vector representation of EPAT-BERT model is consistent with that of BERT model. The input vector at the corresponding position of each word w is composed of three parts: (1) The vector of the **word representation** W_w: that is, the initial word vector of the word, which is used to distinguish different Chinese characters. In this paper, we use the Word2Vec toolkit to obtain the original word vectors. (2) **Position representation** P_w of the word: we use absolute position coding [20] to incorporate sequence position information into input data. (3) **Segment representation** S_w: when the input contains multiple sentences or parts, different segments should be represented with different codes, while the input of EPAT-BERT has only one part, so the segment representation is unique. Finally, the vector representation V_w of each word w is the summation of the representations of three parts:

$$V_w = W_w + P_w + S_w \tag{1}$$

For the two pre-training tasks, cross-entropy loss function with L2 regular term is used to measure the difference between predicted values and real values, and the loss function is optimized using AdamW learner with a learning rate of 5e-5. In the pre-training stage, the training data is used to optimize parameters in the model. The batch size is set to be 8. After every 8000 training rounds, the loss function is calculated on the validation set with a 5-fold cross validation. When the loss function does not fall in the 8000 training rounds, the pre-training process will be stopped, so as to avoid overfitting. The model is built using transformers and PyTorch libraries. Since EPAT-BERT needs to be pre-trained from scratch, its model parameters are all initialized randomly. After pre-training, its all parameters are stored to be fine-tuned later.

In order to realize the random mask of the input data, OwnThink knowledge graph (www.ownthink.com) is introduced to mark the entities contained in the input text. Then, each word in the corresponding entity that should be masked is replaced with a special mask token ''[MASK]''. After transformed by EPAT-BERT, the position of each ''[MASK]'' in the input will be transformed again by a hidden layer vector. By connecting a fully-connected layer, the word at the corresponding position of each ''[MASK]'' can be predicted, so as to carry out end-to-end training. We believe that by introducing the entity-granularity masked language model task, the model can learn more content related to domain knowledge on the basis of the word-granularity language model task, so as to more accurately understand the text related to the power field and improve the performance of the downstream classification task.

2) MODEL TRAINING IN THE FINE-TUNING STAGE

The input vector representation of EPAT-BERT in the fine-tuning stage is the same as that in the pre-training stage, which is also composed of the vector representation of words, position coding of words and segmented representation of words. In the fine-tuning stage, according to existing work, in order to complete text classification task, we need to add a special mark ''[CLS]'' at the beginning of the input text. The output vector corresponding to the ''[CLS]'' token can then be seen as the vector representation of the entire input text. After that, a fully-connected layer can be added on the upper layer of EPAT-BERT, whose number of neurons is the total number of categories of the audit text that need to be classified. So far, the whole EPAT-BERT has formed an end-to-end neural architecture. In the training stage, the loss function with L2 regularization is used for optimization. In the test stage, we select the category corresponding to the neuron with the highest output probability as the prediction category to achieve the purpose of automatic classification of audit text.

IV. EXPERIMENTS AND RESULT ANALYSIS

A. EXPERIMENTAL SETUPS

This experiment runs on a GPU cloud server. The specific configuration is as follows: the CPU uses Intel (R) Xeon (R) silver 4114 CPU@2.20GHz, GPU is four NVIDIA Titan V, each with 12GB VRAM. The server memory is 256GB and the hard disk capacity is 2T. The software packages and frameworks required for the experiment include pytorch 1.7.1, transformers 4.7.0, scikit learn 0.24.2, numpy 1.19.5, pandas 1.1.5 and matplotlib 3.3.4. For the BERT and our proposed EPAT-BERT, we use the pre-trained Chinese BERT-base as the original model.[1] The neural network structure of Chinese BERT is totally the same with English BERT model. The only difference between them is the way of tokenization of the text: the English BERT-base applies BPE tokenization, while the Chinese BERT-base tokenizes Chinese texts in units of Chinese characters.[2]

B. DATASET

In order to obtain electric power texts, the professional vocabulary in electric power field is first sorted into a vocabulary list V, and then, we search web pages that contain one or more vocabularies in V. We use the web pages candidate set provided by Yahoo. We then record the web pages set as W. Using the extraction algorithm based on regular expression, texts in set W is extracted as the pre-training corpus in this paper, which is recorded as C. The C contains 1.5M pieces of texts. We then divide C into training set (95%) and validation set(%5). We do not use much validation data like 20% or 30% because the scale of C is large enough, and 5% of the data from C is enough for validation. In the pre-training stage, when the loss of each round of the model on the validation set does not continue to decline, the training will be stopped. In the fine-tuning stage, we select 1,500 electric power audit texts from the daily audit records of a electric power company to form a dataset T, which is divided into training set with 1,000 pieces of data and test set with 500 pieces of data. There are 24 categories in total, and we sample the data categories evenly, so that the amount of data for each category is approximately equal. The process achieves fairness for each category. We also split the training set into 800 pieces of training data and 200 pieces of validation data. The training set is used to optimize the model. After each training round, the F1 score is calculated on the validation set. When the F1 score on the validation set does not drop anymore compared with the last training round, the training is stopped. Then, the evaluation metrics are calculated on the test set. F1 score is adopted as the basis for early stop because this metric is a synthesis of other metrics, which has representative significance. By dividing the model into training set, validation set and test set, it can ensure that the model can achieve the best generalization ability, which is better than the case of dividing the model into training set and test set only.

[1]The Chinese BERT-base: https://huggingface.co/bert-base-chinese
[2]https://github.com/electricAudit/auditTextClassification/

C. BASELINE MODELS AND EVALUATION METRICS

In order to illustrate the effectiveness of EPAT-BERT proposed in this paper, we design several groups of baseline models to be compared. First, we implement several main-stream traditional machine learning algorithms, including Naive Bayes, SVM, GBDT, AdaBoost, and XGBoost. In these algorithms, the input text is represented as a bag-of-words vector, then the models classify the vector in different manners.

In addition, two deep learning models commonly used for text classification are selected: text convolution neural network (TextCNN): the word vector sequence corresponding to the text is regarded as a matrix, and the convolution neural network is used to extract spacial features of the matrix and conduct end-to-end learning for text classification. Long short-term memory network (LSTM): the word vector sequence corresponding to the text is sequentially sent to the LSTM, and then an end-to-end learning process is performed.

Finally, in order to demonstrate the effectiveness of the electric power text pre-training task, the general pre-trained BERT model is selected for comparison. Compared with our proposed EPAT-BERT, the original BERT model does not contain entity-level masked language model task, and it is pre-trained with general texts, instead of electric-power-related texts. Therefore, we deem that EPAT-BERT can achieve better performance than BERT, which is also the main emphasis of this paper.

We use classification accuracy, precision, recall, and F1 score, four commonly-used evaluation metrics, to evaluate our proposed model, and compare with existing text classification baselines. For the detailed implementation of these four metrics, we recommend to read [6].

D. EXPERIMENTAL RESULTS

The evaluation metrics calculated in different models on the test set are shown in Table 2. All the results are calculated with 5-fold cross validation. Based on the experimental results, we can obtain the following conclusions: (1) Compared with traditional machine learning models (Naive Bayes and SVM), deep learning models like TextCNN and LSTM based on neural networks can achieve better results in four evaluation metrics, which proves that the models based on neural network is superior to traditional machine learning models based on statistical learning. (2) Compared with deep learning models, pre-trained language model BERT has further improved the experimental results in four evaluation metrics. (3) The text classification model EPAT-BERT proposed in this paper is significantly better than the general pre-trained model BERT, which confirms the effectiveness of the two granularity pre-training tasks proposed in this paper and the promotion of the field related pre-training to the field downstream tasks. We further apply the t-test to illustrate the significance of our proposed EPAT-BERT compared with existing baselines. The t-test results are also shown in Table 2. It can be seen that, except for the Precision, other three metrics (Accuracy, Recall, and F1-score) all pass the t-test

TABLE 1. The experimental results of electric power Audit text classification on four evaluation metrics. The Bold Indicates the Best Results Compared with Other Models. The "†" denotes outperforming the best baseline (BERT) in t-test with $p < 0.05$.

Algorithm	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.6042	0.6217	0.5950	0.6081
SVM	0.6203	0.6329	0.6300	0.6314
GBDT	0.6415	0.6532	0.6628	0.6579
AdaBoost	0.6636	0.6577	0.6698	0.6640
XGBoost	0.6670	0.6672	0.6704	0.6688
TextCNN	0.7165	0.7427	0.6901	0.7156
LSTM	0.7278	0.7439	0.7050	0.7239
BERT	0.7791	0.7823	0.7794	0.7808
EPAT-BERT	**0.8196**†	**0.8079**	**0.8162**†	**0.8120**†

TABLE 2. Experimental results of Ablation studies.

Algorithm	Accuracy	Precision	Recall	F1-score
EPAT-BERT	**0.8196**	**0.8079**	**0.8162**	**0.8134**
EPAT-BERT w/o. W	0.7926	0.7931	0.7870	0.7870
EPAT-BERT w/o. E	0.7839	0.7902	0.7866	0.7983
EPAT-BERT-WE	0.8009	0.7935	0.8037	0.8012
EPAT-BERT-EW	0.8043	0.7944	0.8059	0.8028

with $p < 0.05$, showing that the improvements over all evaluation metrics are significant.

E. RESULT ANALYSIS

EPAT-BERT model focuses on two pre-training tasks: word-granularity masked language model and entity-granularity masked language model. Therefore, it is important to explore the impact of these two pre-training tasks on the experimental results. To achieve this, two groups of ablation experiments are further designed in this paper. In each ablation study, we remove a single module of the EPAT-BERT model, or change the settings of these modules, then compare the ablation results with original EPAT-BERT model. The experimental results are shown in Table 2.

In the first group of experiments, the pre-training tasks of word-granularity and entity-granularity masked language model in EPAT-BERT are removed and recorded as EPAT-BERT w / o. W and EPAT-BERT w / o. E respectively. The experimental results show that, when the two pre-training tasks in the model are removed, the model decreases in the four classification evaluation metrics, which proves that the two granularity pre-training tasks both have an important role in further improving the classification effect of electric power audit text. In addition, the effect of entity-granularity pre-training on downstream tasks is more significant than word-granularity pre-training. In the second group of experiments, the effects of the training order of the two pre-training tasks in EPAT-BERT on the experimental results are explored. In the experiment, "-WE" means that word-granularity is performed first, and then entity-granularity masked language model training is performed. "-EW" is the opposite, which indicates that we first pre-train EPAT-BERT with entity-granularity masked language model, followed by word-level masked language model. The experimental results show that, compared with completing two pre-training tasks sequentially ("-WE" and "-EW"), the fusion of the two tasks is better, which is used in our proposed EPAT-BERT model, and the order of the two tasks has no significant effect on the results.

V. CONCLUSION

With the development of computer science and machine learning technology, electric power industry has accumulated a large number of audit texts. It has become a key problem for electric power enterprises to classify these audit texts quickly and automatically with high accuracy. The "pre-training and fine tuning" paradigm has greatly improved the effectiveness of various natural language processing tasks. This paper integrates this paradigm into the text classification task of power audit, and proposes two-granularity pre-training tasks: word-granularity and entity-granularity masked language model. The experimental results show that, compared with traditional machine learning models, fully-supervised models based on deep neural networks and general pre-training language models, our proposed EPAT-BERT model can significantly exceed existing models in accuracy, precision, recall, and F1 score of the classification of electric power audit texts, and can be applied to audit work of electric power industry to improve the efficiency and accuracy of audit analysis.

EPAT-BERT model has strong expansibility and easy popularization. Since it uses texts related to electric power for pre-training, it can be expanded and complete any downstream tasks related to electric power text by modifying down-streaming neural network layers of the pre-training module. For example, it can be applied to the classification of other related texts related to electric power, automatic generation or retrieval of audit opinions, project type annotation, etc. EPAT-BERT is based on the classical pre-training language model BERT proposed in 2019. Theoretically, the model can also be based on other pre-training model frameworks, but it does not belong to the scope of this paper, and can be further studied as one of the future research directions.

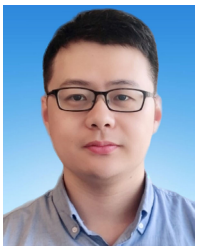
REFERENCES

[1] X. Yu, Z. Zhao, Y. Ma, R. Zheng, Z. Xi, and C. Ma, "Multi-label text classification for power ICT custom service system based on binary relevance and gradient boosting decision tree," *J. Autom. Electr. Power Syst.*, vol. 45, no. 11, pp. 144–151, 2021, doi: 10.7500/AEPS20200511001.
[2] Y. Ding, X. Shang, and W. Mi, "Deep learning based knowledge extraction method for text of power grid dispatch and control," *Automat. Electr. Power Syst.*, vol. 44, no. 24, pp. 161–168, 2020, doi: 10.7500/AEPS20191228004.
[3] S. Guanyu, W. Huifang, W. Xianghong, L. Jinlong, L. Jianhong, and H. Benteng, "Precise information identification method of power equipment defect text based on dependency parsing," *Automat. Electr. Power Syst.*, vol. 44, no. 12, pp. 178–185, 2020, doi: 10.7500/AEPS20190401001.
[4] J. Qiu, H. Wang, G. Ying, B. Zhang, G. Zou, and B. He, "Text mining technique and application of lifecycle condition assessment for circuit breaker," *Automat. Electr. Power Syst.*, vol. 40, no. 6, pp. 107–112, 2016, doi: 10.7500/AEPS20150812003.
[5] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in *Proc. IEEE 14th Int. Conf. Cognit. Informat. Cogn. Comput. (ICCI CC)*, Jul. 2015, pp. 136–140, doi: 10.1109/ICCI-CC.2015.7259377.
[6] K. Kowsari, J. Meimandi, J. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.
[7] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/j.neucom.2019.01.078.
[8] X. Chen, P. Gao, Y. Liang, and Y. Ma, "A category hybrid embedding based approach for power text hierarchical classification," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 58, no. 1, pp. 77–82, 2022, doi: 10.13209/j.0479-8023.2021.104.

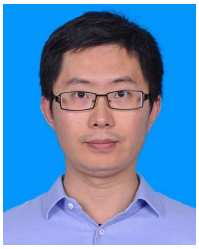
[9] Z. Yaxin, Z. Minghong, S. Linxin, X. Fei, J. Jinyang, and Y. Xin, "A two-phase short-text classification method for classifying audit problems in power grid companies," *J. Southwest Univ., Natural Sci.*, vol. 42, no. 10, pp. 1–7, 2020, doi: 10.13718/j.cnki.xdzk.2020.10.001.
[10] C. Ping, K. Yao, H. Jingyi, W. Xiangyang, and C. Jing, "Text categorization method with enhanced domain features in power audit field," *J. Comput. Appl.*, vol. 40, no. S1, pp. 109–112, 2020, doi: 10.11772/j.issn.1001-9081.2019111973.
[11] B. Feng, Y. Zhang, X. Tang, C. Guo, J. Wang, Q. Yang, and H. Wang, "Power equipment defect record text mining based on BiLSTM-attention neural network," *Proc. CSEE*, vol. 40, no. S1, pp. 1–10, 2020, doi: 10.13334/j.0258-8013.pcsee.200530.
[12] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, Jan. 2020, doi: 10.1162/tacl_a_00300.
[13] J. D. M. W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[14] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451, doi: 10.18653/v1/P19-1139.
[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.
[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[17] W. Ling, C. Dyer, A. W. Black, and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 1299–1304, doi: 10.3115/v1/N15-1142.
[18] B. McCann, J. Bradbury, C. Xiong, and R. Socher., "Learned in translation: Contextualized word vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
[19] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 55–65.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[21] P. Budzianowski and I. Vulić, "Hello, it's GPT-2—How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems," in *Proc. 3rd Workshop Neural Gener. Transl.*, 2019, pp. 15–22.
[22] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 583–592, doi: 10.18653/v1/2020.acl-main.54.
[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 328–339.
[24] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Mar. 2010, pp. 201-208. Accessed: Nov. 25, 2022. [Online]. Available: http://proceedings.mlr.press/v9/erhan10a.html
[25] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 2712–2721. Accessed: Nov. 25, 2022. [Online]. Available: https://proceedings.mlr.press/v97/hendrycks19a.html
[26] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, Jan. 2021, doi: 10.1016/j.aiopen.2021.06.003.
[27] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.
[28] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3615–3620, doi: 10.18653/v1/d19-1371.
[29] A. Alessa, M. Faezipour, and Z. Alhassan, "Text classification of flu-related tweets using FastText with sentiment and keyword features," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 366–367, doi: 10.1109/ICHI.2018.00058.

QINGLIN MENG (Member, IEEE) was born in Tianjin, China. He received the B.S. and M.S. degrees in electrical engineering from Tianjin University, China, in 2003 and 2010, respectively. As a Senior Engineer, he is currently the Deputy Director of the Operational Audit Division, Comprehensive Service Center, State Grid Tianjin Electric Power Company. He is named Digital Audit Expert of SGCC, Reserve Excellent Expert of SGCC, Science Communication Expert of CSEE, and Postgraduate Supervisor of Enterprise at Tiangong University. His main research interests include new energy power generation technology, integrated energy system, distribution system automation, power cable technology, energy blockchain, and digital audit technology.

YAN SONG was born in Tianjin, China. He received the B.S. degree in electrical engineering from Tianjin University in 2002, and the M.S. degree in automation engineering from the Tianjin University of Technology in 2012. As a Senior Economist, he is currently the Deputy Director of the Comprehensive Service Center, State Grid Tianjin Electric Power Company. He is named Power Supply Inspection Expert at State Electricity Regulatory Commission, an Advanced Individual in Marketing at SGCC, and a Young Post Expert of North China Power Group. His main research interests include marketing electricity fees technology and digital audit technology.



JIAN MU was born in Dezhou, Shandong, China. He received the B.S. degree in electrical engineering from the Lanzhou University of Technology in 2010, and the M.S. degree in electrical engineering from Tianjin University in 2018. He is currently an Engineer at the Audit Department, State Grid Tianjin Electric Power Company. He is named Digital Audit Expert at SGCC. His main research interests include thermal power generation technology and digital audit technology.



YUANXU LV was born in Jinan, Shandong, China. He received the B.A. degree in accounting from the Shandong University of Finance and Economics in 2012, and the M.B.A. degree from the University of International Business and Economics in 2020. He is currently an Accountant at the Audit Supervision Department at SGCC. His main research interest includes digital audit technology.

JIACHEN YANG (Senior Member, IEEE) was born in Anshan, Liaoning, China. He received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, China, in 2005 and 2009, respectively. He is currently an Endowed Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with the Department of Computer Science, Loughborough University, U.K. and Embry-red Aeronautical University, USA. His main research interests include information processing, the Internet of Things, cloud/edge computing, big data, pattern recognition, and image quality assessment. He served as a Guest Editor for many reputed journals, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *IS*, and *Digital Communications and Networks*. He is also an Associate Editor of many international journals, such as *Alexandria Engineering Journal*, *Journal of Ambient Intelligence and Humanized Computing*, IEEE ACCESS, and *Sensors*.

LIANG XU was born in Tianjin, China. He received the B.S. degree in electrical engineering from the Tianjin University of Science and Technology in 2005, and the M.S. degree in electrical engineering from Tianjin University in 2019. As a Senior Engineer, he is currently the Executive Director at Tianjin Tianyuan Power Engineering Company Ltd., Baodi Power Supply Branch, State Grid Tianjin Electric Power Company. He is named Executive Deputy Director of Baodi Power Engineering Quality Supervision Station. His main research interests include power system automation and digital technology.

JIN ZHAO was born in Yuncheng, Shanxi, China. He received the B.S. degree in electronic information engineering from the Changsha University of Science and Technology in 2013, and the M.S. degree from Xi'an Jiaotong University in 2016. He is currently an Engineer at the Electric Power Research Institute, State Grid Shanxi Electric Power Company. His main research interests include new energy power generation technology, power system automation, and digital technology.

JUNWEI MA was born in Jining, Shandong, China. He received the Ph.D. degree in control theory and control engineering from the Dalian University of Technology, China, in 2011. As a Senior Engineer, he is currently the Deputy Director of the Technology Development Department, Information and Communication Branch, State Grid Shanxi Electric Power Company. He is named the Head of the Energy Blockchain Laboratory, State Grid Shanxi Electric Power Company, and the Doctor Workstation of State Grid Shanxi Electric Power Company. His main research interests include energy blockchain and artificial intelligence technology.

WEI YAO was born in Taiyuan, Shanxi, China. He received the B.S. degree in electronic and information engineering from Shanxi University, China, in 2009, and the M.S. degree in software engineering from the Dalian University of Technology, China, in 2013. He is a Senior Engineer at Taiyuan Power Supply Company, State Grid Shanxi Electric Power Company. He is named the Excellent Expert of State Grid Shanxi Electric Power Company, Science Communication Expert at CSEE, Sanjin Talent in Shanxi. His main research interests include big data and information and communication technology.

RUI WANG was born in Tianjin, China. He received the B.S. degree in electrical engineering from Tianjin University, China, in 2004. As a Senior Engineer, he is currently the Chief Economist at Chengxi Power Supply Branch, State Grid Tianjin Electric Power Company. He is named the Supervisor at Tianjin Electric Power Engineering Association. His main research interests include power system automation and digital technology.

MAOXIANG XIAO was born in Qingdao, Shandong, China. He received the B.S. and M.S. degrees in structural engineering from Northeast Electric Power University, in 2008 and 2011, respectively. As a Senior Engineer, he is currently the General Manager at Ningdongshengyuan Electric Power Engineering Company Ltd., Ninghe Power Supply Branch, State Grid Tianjin Electric Power Company. He is named Reserve Excellent Expert of SGCC. His main research interests include new energy power generation technology, power system automation, and digital technology.

QINGYU MENG was born in Zhangjiakou, Hebei, China. She received the B.S. degree in electrical engineering from the Hebei University of Architecture in 2004, and the M.S. degree in electrical engineering from Tianjin University in 2018. As a Senior Engineer, she is currently the Secretary of the Discipline Inspection Committee and the Chairperson of Trade Union of the Zhangjiakou Wanquan District Power Supply Branch, State Grid Jibei Electric Power Company Ltd. She is named the Director of the Electrical Engineering and Automation Discipline Construction Committee and a Postgraduate Supervisor of enterprise at the Hebei University of Architecture. Her main research interests include power system automation and digital audit technology.