EP-BERTGCN: A Simple but Effective Power Equipment Fault Recognition Method

Mingcong Lu*
lmc8133@bupt.edu.cn
Beijing University of Posts and Telecommunications
China

Yusong Zhang*
zhangys2021@bupt.edu.cn
Beijing University of Posts and Telecommunications
China

Qu-An Zheng
quan.zheng@tj.sgcc.com.cn
State Grid Tianjin Electric Power Company
China

Zhenyuan Ma
mazhenyuan@epri.sgcc.com.cn
China Electric Power Research Institute
China

Liqing Liu
liqing.liu@tj.sgcc.com.cn
State Grid Tianjin Electric Power Company
China

Yongping Xiong
ypxiong@bupt.edu.cn
Beijing University of Posts and Telecommunications
China

Ruifan Li†
rfli@bupt.edu.cn
Beijing University of Posts and Telecommunications
China

ABSTRACT

With the advancement of China's State Grid in recent years, text-based power equipment fault recognition has become an essential tool for power equipment maintenance. The task suffers from the domain gap that exists between the electric power domain and the general natural language processing domain. To improve the recognition performance, we proposed a method that combines pre-trained Bidirectional Encoder Representations from Transformers (BERT) and Graph Convolutional Network (GCN), i.e., Electric Power -BERTGCN. Our EP-BERTGCN first builds the graph among documents and words within documents based on pre-trained BERT. Then, the two softmax outputs with pre-trained BERT and GCNs are combined for final classification results. Extensive experiments show that our method outperforms previous baselines.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

*Both authors contributed equally to this research.
†Corresponding Author.

ITCC 2022, June 23–25, 2022, Guangzhou, China

ACM ISBN 978-1-4503-9682-0/22/06…$15.00
https://doi.org/10.1145/3548636.3548646

KEYWORDS

equipment fault recognition, natural language processing, text classification, pre-trained language model, graph convolutional network

ACM Reference Format:
Mingcong Lu, Yusong Zhang, Qu-An Zheng, Zhenyuan Ma, Liqing Liu, Yongping Xiong, and Ruifan Li. 2022. EP-BERTGCN: A Simple but Effective Power Equipment Fault Recognition Method. In *2022 4th International Conference on Information Technology and Computer Communications (ITCC) (ITCC 2022), June 23–25, 2022, Guangzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3548636.3548646

1 INTRODUCTION

With the development of China's State Grid in recent years, the power grid system manages more and more power equipment. The frequency of power equipment faults has also significantly increased, which is challenging for the traditional manual inspection process. Therefore, various automatic fault recognition methods have been proposed to improve the efficiency of power equipment maintenance [14]. Different types of modalities are used to recognize faults of power equipment, such as images, text, video, and so on. Notably, Text-based Power equipment Fault Recognition, i.e., TPFR is the intersection of a text classification task and power equipment fault recognition task, which aims to recognize faults using textual fault information generated by the system of the power grid company. As shown in Figure 1, TPFR is a text classification task that introduces domain knowledge from the electric power.

The text classification task is one of the most popular tasks in natural language processing (NLP). Many effective methods for text classification have been proposed. The classical deep network structures can be used for text classification tasks, such as CNN, LSTM, GCN and Transformer [8]. Traditional NLP methods tend to train end-to-end networks or introduce word embeddings trained on large corpora as initialized embeddings. This results in the weak

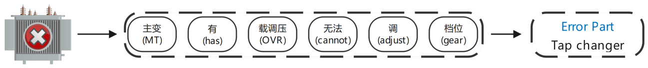

Figure 1: An illustration on TPFR task. MT and OVR denote Main Transformer and On-load Voltage Regulation.

transferability between tasks. The best methods in recent years are based on pre-trained models [1, 2]. With the proposal of these large pre-trained models, most NLP tasks achieved better performance. Therefore, pre-trained models have become the mainstream of NLP.

TPFR task is different from the generic NLP tasks. As shown in Figure 1, there are a lot domain-specific words and semantic relationships. However, there is no pre-trained model based on a large-scale electric power corpus. Applying domain adaption on a general BERT model is an effective solution to these specialized domain problems [17].

In this paper, we propose a method to solve the TPFR task by combining pre-trained BERT model with graph convolution network. Our method is called Electric Power BERT Graph Convolution Neural Network (EP-BERTGCN), EP-BERTGCN incorporates the BERT model into our previous C-TextGCN method [21] to improve the performance. We also apply domain adaption to overcome the gap between different domains. Extensive experiments show that EP-BERTGCN outperforms our previous work and other text classification methods.

Our contributions can be summarized in the following three points:

- We introduced the BERT module to our C-TextGCN and proposed the novel EP-BERTGCN method.
- We conducted extensive experiments on the dataset we collected and compared with multiple baselines. The results show that our method achieves the best performance.
- Based on EP-BERTGCN, we focus on the domain adaption of the BERT model, and improves the performance of our method further.

2 RELATED WORKS

Our work mainly involves several research fields, including text-based power equipment fault recognition, graph neural network, pre-trained BERT, and the cross-domain adaption of pre-trained BERT. We briefly provide the overview of related works as follows.

Power equipment fault recognition is a real problem that is of great interest to grid companies, and lots of related works have been proposed to solve it. Most of the previous work was based on image modal information from surveillance devices such as infrared cameras, which can provide heat features for the captured image [4]. There is also a lot of text modal fault-related information in the power system, which can also be used for fault recognition, and some text-based models have recently been proposed to solve the TPFR task [19].

Graph Convolutional Neural Network (GCN) [7] is a robust neural network model. GCN mainly uses the adjacency relationship between the nodes in a graph to aggregate the information in the graph. The text classification task can also be seen as a kind of graph. Yao et al. proposed a novel TextGCN method of text classification based on graph convolution by constructing graph based on the document [20]. TextGCN defines edges based on the 'belongs to' relationship between words and documents. They define the weight of edges by TF-IDF [16] and PMI. Through graph convolution, TextGCN can capture the relationship between documents and words, which improves the robustness and efficiency of text classification.

Pre-trained models have been proposed changing the paradigm of NLP tasks. The paradigm of pretrain-finetune [15] is to pre-train the model on some subtasks on a huge corpus. The classical subtasks includes Masked Language Model (MLM) and Next Sentence Prediction (NSP). And then the model is trained on real tasks with pre-trained weights. BERT has been proven to be a general natural language processing module that can be easily fine-tuned to other downstream tasks with little training and great performance. Since BERT was proposed, many improved versions of pre-trained models have been published, such as Roberta [12], ALBERT [10], etc.

Since TPFR is a cross-domain task, it is intuitive to integrate power electric knowledge for better performance. Since training from scratch is unacceptable due to the difficulty of collecting data in some specialized domains, there has been a lot of work on the domain adaptation of pre-trained BERT. Ma et al. [13] proposed a domain adaptation method based on domain classification, and Diao et al. [3] proposed a method with low resource consumption with simple n-gram. These works focus on attaching domain knowledge to a general pre-trained BERT to reduce the gap between domains.

3 METHOD

In this section, we will illustrate our EP-BERTGCN framework in detail. Firstly, we introduce the construction of the document word graph in Section 3.1. Then, we discuss the trade-off strategy between BERT and GCN of the model in Section 3.2.

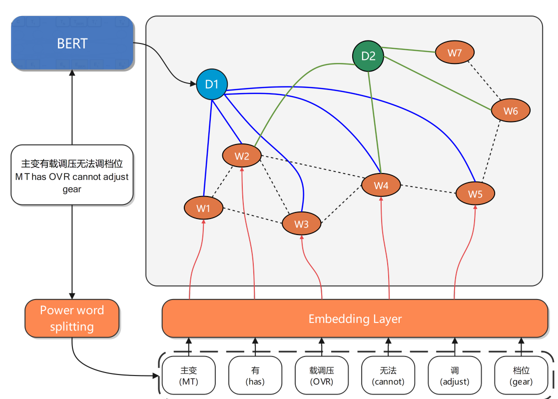

Figure 2: Constructing graph from corpus. The blue and green nodes with D1 and D2 denote two different document nodes, the other orange nodes are word nodes. The blue solid line and black dotted line are document-word edges and word-word edges, respectively.

3.1 GCN

To make use of both labeled and unlabeled data in the training process, we build our text graph from the whole dataset, which contains the training part and the testing part. Specifically, we build the heterogeneous graph containing word nodes and document nodes following TextGCN [20]. In the graph, there are two types of edges, one is the word-document edges based on term frequency-inverse document frequency (TF-IDF) and the other is word-word edges based on point-wise mutual information (PMI). Mathematically, we use the complete dataset build the graph $G = (V, E)$, where $V(|V| = n)$ and E are sets of nodes and edges. And the construction rules of adjacency matrix A are described as follows,

$$A_{ij} = \begin{cases} \mathrm{PMI}(i,j) & i,j \text{ are words, } \mathrm{PMI}(i,j) > 0 \\ \text{TF-IDF}_{ij} & i \text{ is document, } j \text{ is word} \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The calculation rules of PMI are defined as follows,

$$\begin{aligned} \mathrm{PMI}(i,j) &= \log \frac{p(i,j)}{p(i)p(j)} \\ p(i,j) &= \frac{\#W(i,j)}{\#W} \\ p(i) &= \frac{\#W(i)}{\#W} \end{aligned} \tag{2}$$

where $\#W(i)$ is the number of sliding windows containing the word i, $\#W(i,j)$ is the number of sliding windows containing the word i and the word j, and $\#W$ denotes the total number of sliding windows. The window size L is a hyper-parameter, we will discuss it in the experiment section. If the PMI value between two words is positive, then we think the corresponding words are highly related, so we add an edge between them in the graph. Otherwise, we think the two words are not related and there is no edge.

We build the word-document graph shown in Figure 3. Initializing the feature matrix $X \in \mathbf{R}^{n \times m}$, the matrix contains n nodes with their features, the dimensionality of feature vectors is m, where n is the sum of document nodes $n_{doc}$and word nodes n_{word}. In TextGCN, the initial feature matrix is set as an identity matrix $X = I$. In our method, we initialize the feature matrix by document embeddings and word embeddings obtained from a BERT-style model. Document node embeddings and word node embeddings are denoted by $X_{doc} \in \mathbf{R}^{n_{doc} \times m}$ and $X_{word} \in \mathbf{R}^{n_{word} \times m}$. So the feature matrix is denoted as follows,

$$X = \begin{pmatrix} X_{\text{doc}} \\ X_{\text{word}} \end{pmatrix}_{(n_{\text{doc}} + n_{\text{word}}) \times m} \tag{3}$$

We feed the initial feature matrix into the GCN model, and each layer in GCN aggregate the node information through edges and iteratively propagates messages in datasets, which is shown in Figure 3. The output feature matrix of each layer $L^{(i)}$ is described as

$$L^{(i+1)} = \sigma\left(\tilde{A} L^{(i-1)} W_i\right) \tag{4}$$

where σ is an activation function, and $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is Laplacian matrix, $L^{(0)} = X$ is the initial feature matrix. We use cross-entropy

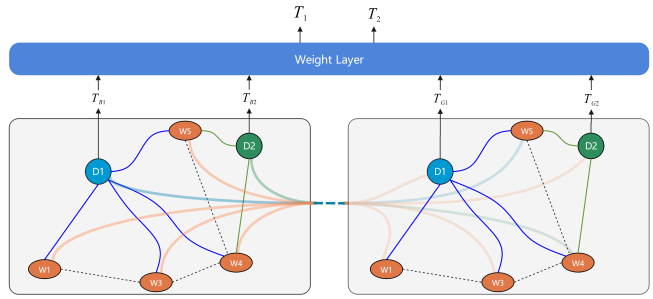

Figure 3: EP-BERTGCN. The output from left part denoted as T_{B1} and T_{B2} is the origin document features from BERT, namely Z_{BERT}. Correspondingly, the output on the right from the GCN which after aggregating the node information is Z_{GCN}

as loss function which is frequently used in classification problems.

$$\mathcal{L} = -\sum_{p} Y_p \ln Z_p \tag{5}$$

$Z = \mathrm{softmax}(GCN(X, A))$ is the model output and Y is the ground truth label, they are both $M-$dimensional vectors where M is equal to the number of document classes. And p denotes the index of training samples.

3.2 Model Trade-off

Considering that the document embeddings output from BERT can be used for sentence level downstream tasks, such as text classification. And combined with the auxiliary classifier that directly oriented to BERT embeddings leads to better performances and faster convergence. Therefore, we use both the outputs from BERT and GCN as the input of the document classifier.

Firstly, we construct an auxiliary classifier using document embeddings from BERT (CLS token) denoted as X_{doc} with softmax activate function:

$$Z_{BERT} = \mathrm{softmax}(W X_{doc}) \tag{6}$$

The final classifier is a linear combination of the auxiliary classifier and the BertGCN output:

$$Z_{final} = \lambda Z_{GCN} + (1 - \lambda) Z_{BERT} \tag{7}$$

where Z_{GCN} is described completely in 3.1. The hyper-parameter λ keeps the trade-off between the two parts. $\lambda = 0$ means we only use the BERT output, and $\lambda = 1$ means we only use the GCN module.

The classifier combined the two outputs achieved better performance. We analyse the reasons as follows. The introduced auxiliary classifier directly accepts the document embeddings from BERT as input. And, critically, the document embeddings from BERT is also the input of BertGCN, which ensures that the input of BERTGCN is optimized in the training process. This can to some extent overcome the drawbacks of GCN model, such as over-smoothing, and leads to better performance.

4 EXPERIMENT

In this section, we first describe the adopted dataset and implementation details. Then, we briefly review the compared baseline

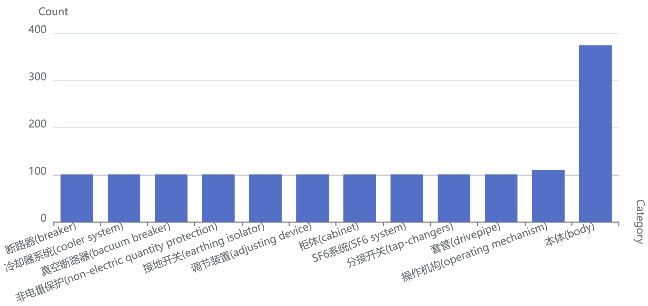

Figure 4: The quantity of each category in the oversampled CPTF dataset

models. Finally, we report the experimental results and the details of ablation studies.

4.1 Dataset

We use the dataset collected in our previous work [21], which is called Chinese Power Text-Fault (CPTF) dataset. The CPTF dataset contains 751 instances of 12 categories, each instance contains a document which is Chinese electric power fault description and the corresponding fault class label. However, the distribution of categories is unbalanced. The dominant category takes almost 50%, which is not conducive for the task of text classification. To overcome the problem, we apply oversampling on the categories with fewer items to avoid overfitting. Finally, the dataset contains 1484 instances for evaluation. The count of each category is shown in figure 4.

4.2 Baselines

We adopt various recently-proposed text classification methods as our baselines. The seven methods are briefly reviewed as follows.

- **TextRNN** [11] integrates RNN into the multi-learning framework, which learns to map arbitrary text into semantic vector representations with both task-specific and shared layers.
- **TextRNN_Att** [22] utilizes neural attention mechanism with Bidirectional Long Short-Term Memory Networks (BiLSTM) to capture the most important semantic information in a sentence. It does not use any features derived from lexical resources or NLP systems.
- **TextRCNN** [9] apply a recurrent structure to capture contextual information as far as possible when learning word representations, which may introduce considerably less noise compared to traditional window-based neural networks.
- **DPCNN** [5] enhance text region embedding with unsupervised embeddings for improving accuracy.
- **FastText** [6] is a simple baseline method for text classification, its word features can be averaged together to form good sentence representations. It is much faster than other methods inspired by deep learning.
- **Transformer** [18] is first proposed to solve sequence to sequence problem such as machine translation. In recent works, transformer has been found effective in sentence level tasks like classification.

Table 1: Experimental Results on the CPTF Dataset.

Model	Acc	Macro-F1	Weighted Macro-F1
TextRNN	0.4950	0.4217	0.4642
TextRNN_Att	0.5538	0.5240	0.5412
TextRCNN	0.6126	0.5800	0.5932
DPCNN	0.5753	0.5515	0.5659
FastText	0.6098	0.5870	0.5978
Transformer	0.4570	0.4245	0.4594
C-TextGCN	0.6607	0.6324	0.6499
EP-BERTGCN	**0.6700**	**0.6645**	**0.6633**
EP-Adaptation	**0.6772**	**0.6668**	**0.6713**

- **C-TextGCN** [21] choose Chinese character instead of word as the basic embedding unit, and find it better than word embedding to feed the text features into GCN and do the text classification.

4.3 Implementation Details

All the experiments were conducted on the NVIDIA GeForce RTX 1080ti GPU and PyTorch 1.5. Following the original GCN [7], we set the layer of GCN equal to 2. We set the feature embedding dimension to be 300. The other hyper-parameter settings are discussed in Section 4.5.

4.4 Domain Adaptation of BERT

We apply domain adaptation for the BERT model. Specifically, before the end-to-end training, we first train BERT alone on classification tasks on the corpus, update the weight, and then use the updated weight as the initial weight of BERT for the end-to-end EP-BERTGCN training. As shown in Table 1, the application of domain adaptation improves the metrics of the model.

4.5 Ablation Study

To evaluate the effect of hyper-parameter λ in balancing the BERT and the GCN modules, and the window size L when building the word document graph. We choose different values while fixing another hyper-parameter, specifically, we fix the window size when using different λ and fix the λ with different window size L. We find the best parameter combinations as $\lambda = 0.4$ and window size $L = 5$, as shown in Fig. 5 and Fig. 6.

For the weight parameter λ, we can see that Weighted Macro-F1 Score first increases as λ becomes larger, but stops and becomes smaller when λ is larger than 0.4. This means that too small λ could not make full use of the abundant relationship between words and documents contained in GCN, while too large λ may suffer from the over smoothing problem in GCN and cannot optimize the input of GCN through BERT module better.

For the window size L, the classification accuracy is similar to the behavior of λ. That is to say, too small window size L can not generate sufficient word co-occurrence information, and too large window size L may add edges between words which are not highly related in graph.

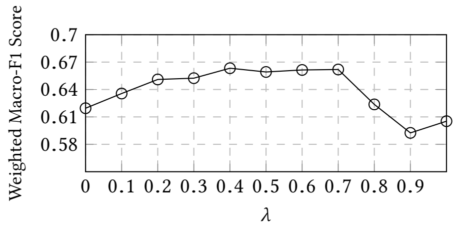

Figure 5: Weighted Macro-F1 Scores with Parameter λ.

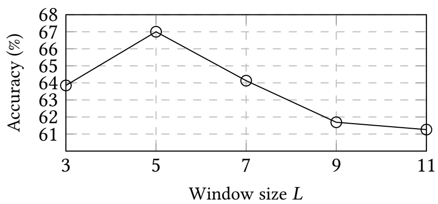

Figure 6: Classification Accuracy with the Window Size L.

5 CONCLUSION

We combine the extensive prior knowledge in the pre-trained language model and the robustness of GCN to construct an effective text classification model for fault recognition in the power field. Facing the domain gap between the power field and the general NLP, we concentrate on the domain adaptation of BERT, and further improve the performance of the model. The comparison with other baseline models demonstrate the effectiveness of our method.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Program of the Headquarters of State Grid Corporation of China, Research on Knowledge Discovery, Reasoning and Decision-making for Electric Power Operation and Maintenance Based on Graph Machine Learning and Its Applications, under Grant 5700-202012488A-0-0-00. The authors would also like to thank anonymous reviewers for their valuable comments that allowed them to improve the final version of this paper.

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

[3] Shizhe Diao, Ruijia Xu, Hongjin Su, Yilei Jiang, Yan Song, and Tong Zhang. 2021. Taming Pre-trained Language Models with N-gram Representations for Low-Resource Domain Adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3336–3349.

[4] Bushra Jalil, Giuseppe Riccardo Leone, Massimo Martinelli, Davide Moroni, Maria Antonietta Pascali, and Andrea Berton. 2019. Fault Detection in Power Equipment via an Unmanned Aerial System Using Multi Modal Data. *Sensors* 19, 13 (2019). https://doi.org/10.3390/s19133014

[5] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 562–570.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).

[7] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[8] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information* 10, 4 (2019). https://doi.org/10.3390/info10040150

[9] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

[11] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[13] Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Hong Kong, China, 76–83. https://doi.org/10.18653/v1/D19-6109

[14] Daniel Martinez, Humberto Henao, and Gerard-Andre Capolino. 2019. Overview of Condition Monitoring Systems for Power Distribution Grids. In *2019 IEEE 12th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED)*. 160–166. https://doi.org/10.1109/DEMPED.2019.8864872

[15] XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science in China E: Technological Sciences* 63, 10 (Oct. 2020), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3 arXiv:2003.08271 [cs.CL]

[16] Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. (01 2003).

[17] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. *arXiv preprint arXiv:1908.11860* (2019).

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Ziyue Xi, Xiaona Chen, Tanvir Almad, and Yinglong Ma. 2019. A Novel Ensemble Approach to Multi-label Classification for Electric Power Fault Diagnosis. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. 267–271. https://doi.org/10.1109/ICCSNT47585.2019.8962410

[20] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph Convolutional Networks for Text Classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 905, 8 pages. https://doi.org/10.1609/aaai.v33i01.33017370

[21] Yusong Zhang, Mingcong Lu, Liqing Liu, Zhijian Li, Fei Jiao, Yongping Xiong, Qinghua Tang, and Ruifan Li. 2021. Chinese Electric Power Equipment Fault Recognition Based on Graph Convolutional Networks. In *2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*. IEEE, 125–129.

[22] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 207–212.