System Report for CCL23-Eval Task 6: A Method For Telecom Network Fraud Case Classification Based on Two-stage Training Framework and Within-task Pretraining

Guanyu Zheng[1], **Tingting He**[2], **Zhenyu Wang**[1*], **Haochang Wang**[2]
[1] South China University of Technology, Guangzhou, China
[2] Northeast Petroleum University, Daqing, China
mounthuangdj@gmail.com, hetingtingjiayou@163.com
wangzy@scut.edu.cn, kinghaosing@gmail.com

Abstract

Domain-specific text classification often needs more external knowledge, and fraud cases have fewer descriptions. Existing methods usually utilize single-stage deep models to extract semantic features, which is less reusable. To tackle this issue, we propose a two-stage training framework based on within-task pretraining and multi-dimensional semantic enhancement for CCL23-Eval Task 6 (Telecom Network Fraud Case Classification, FCC). Our training framework is divided into two stages. First, we pre-train using the training corpus to obtain specific BERT. The semantic mining ability of the model is enhanced from the feature space perspective by introducing adversarial training and multiple random sampling. The pseudo-labeled data is generated through the test data above a certain threshold. Second, pseudo-labeled samples are added to the training set for semantic enhancement based on the sample space dimension. We utilize the same backbone for prediction to obtain the results. Experimental results show that our proposed method outperforms the single-stage benchmarks and achieves competitive performance with 0.859259 F1. It also performs better in the few-shot patent classification task with 65.160% F1, which indicates robustness.

1 Introduction

The official implementation of *Law of the People's Republic of China on Anti-Telecom and Network Fraud* demonstrates China's determination to combat telecom and network fraud in our society. As an essential part of the fight against telecom and network fraud crimes, accurate classification of fraud cases facilitates the public security services to grasp the distribution characteristics of current fraud cases. It assists them in making targeted measures, such as prevention, supervision, suppression, and detection. Currently, there are two problems with the task: (1) The categories of fraud cases are more finely grained. (2) There need to be sufficient text features and data resources.

To cope with these problems, we introduce the two-stage training to enhance the differentiation among samples and to motivate the model to distinguish fine-grained case samples and thus achieve effective classification. In the first stage, the model extracts semantic features in fraud cases to obtain above-threshold prediction samples and use them as prior knowledge. The model then fuses the obtained samples into the second stage to enhance the representation capability of the model. In the second stage, the model uses BERT (Devlin et al., 2019) based fine-tuning to improve the semantic feature mining capability for a few descriptive texts of the cases and finally predicts the labels. In the training process, we introduce a multi-dimensional semantic enhancement method to improve the contextual modeling capability of the model. It contains two dimensions: feature space dimension and sample space dimension. The former uses a combination of adversarial training and multi-sampling. At the same time, the latter introduces semi-supervised learning that uses a model fusion-based approach to generate pseudo-labeled data. We implement the model fusion-based approach by summing the predicted probabilities. In addition, to enhance the model's fitness, we introduce the within-task pretraining approach to maximize

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 206-212, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the use of data from the dataset and further improve the classification effect without introducing external knowledge.

In this paper, we proposed a two-stage training framework, and here are our main contributions:

(1) We propose a two-stage training-based text classification framework. The framework uses in-domain samples to guide data classification and effectively improves the classification results. Meanwhile, it is simple to implement and reusable for similar tasks.

(2) The introduction of the within-task pretraining approach and the multi-dimensional semantic enhancement method effectively compensates for the loss of crucial information caused by the lack of external knowledge and less contextual details on the task data. The experimental result shows that the proposed model outperforms other benchmark models.

(3) Our proposed framework ranks eighth in CCL23-Eval Task 6 (Telecom Network Fraud Case Classification, FCC) with 0.859259 F1 and performs well in the few-shot patent classification task.

2 Proposed Framework

2.1 Within-task Pretraining

As justice-relevant corpus is generally unavailable, we feed BERT the training data for domain pretraining. This approach is known as within-task pretraining. Research shows that the method efficiently improves the performance of the model on specific tasks despite the less training corpus and achieves task-adaptive effects (Gururangan et al., 2020). This makes the within-task pretraining method much less expensive to run than the domain-adaptive pretraining approach that continues pretraining on a large corpus of unlabeled domain-specific text.

2.2 Multi-dimensional Semantic Enhancement

Feature Space dimension: For the case with fewer case details, we perform random perturbation in the embedding layer to generate adversarial samples. This increases the feature space's diversity and enhances the model's robustness to adversarial samples, thus improving the model's performance. Adversarial training is standardized in the following format (Madry et al., 2017):

$$\min_{\theta} E_{(x,y)\sim\mathcal{D}}\left[\max_{\Delta x\in\Omega} L(x+\Delta x, y;\theta)\right] \tag{1}$$

where $\mathcal{D}$ is training dataset, x represent input, y represent label, θ represent model parameters, $L(x, y;\theta)$ is the loss of a single sample, Δx is adversarial perturbation, and Ω is the perturbation space. We use Fast Gradient Method (FGM) (Goodfellow et al., 2014) to optimize equation (1):

$$\Delta x = \epsilon \nabla x L(x, y;\theta) = \epsilon \frac{\nabla x L(x, y;\theta)}{\|\nabla x L(x, y;\theta)\|} \tag{2}$$

substitute the normalized Δx back into equation 1 to complete the optimization:

$$\min_{\theta} E_{(x,y)\sim\mathcal{D}}\left[L(x+\Delta x, y;\theta)\right] \tag{3}$$

We also introduce Adversarial Weight Perturbation (AWP) (Dong et al., 2020) to optimize the training process, and our experimental results show that AWP underperforms FGM on two tasks.

In addition, we randomly sample the hidden layer representation output from the model backbone several times and then calculate the arithmetic mean of all the results. This approach speeds up training convergence and improves generalization. We exploit the randomness of Dropout (Inoue, 2019) by feeding the hidden layer representation into it multiple times to obtain a consistent number of slightly different but within-controllable vector representations, achieving the effect of numerous random samplings.

Sample Space dimension: To expand the case context, we augment the sample space from a data augmentation perspective to improve model performance. As external data in the same domain as the fraud cases are difficult to obtain, the unlabeled test data are a suitable target for semantic augmentation based

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 206-212, Harbin, China, August 3 - 5, 2023.

on the sample space dimension. For more straightforward implementation and to include as much feature information as possible, we first fuse several BERTs with different structures from within-task pretraining to predict the test dataset, generating classification labels and corresponding probabilities. A threshold is then set, and the predicted labels are filtered to select samples above the threshold as the new training data. This data is known as pseudo-labeled data (Lee, 2013). According to the cluster assumption, these sample points with higher probability are usually more likely to be in the same class. Hence, the confidence level of their corresponding pseudo-labels is higher.

2.3 Two-stage Training Framework

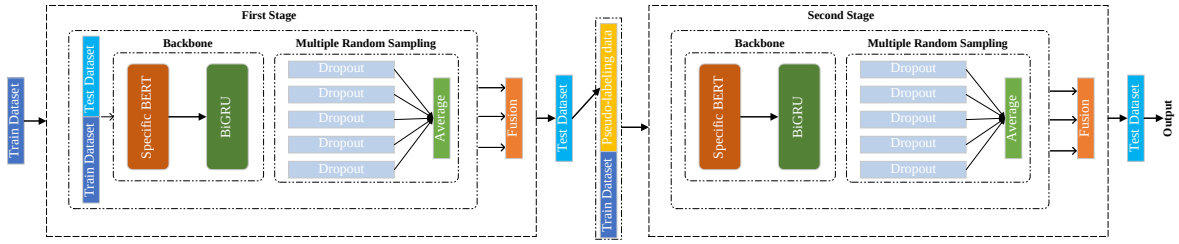


Figure 1: The two-stage training framework

As shown in Figure 1, the two-stage training framework integrates within-task pretraining and multi-dimensional semantic enhancement to improve the classification of fraud cases.

First stage: We use the training corpus for within-task pretraining to obtain a specific BERT, stack the BERT and BiGRU as our model backbone, and introduce semantic enhancement based on the feature space dimension to improve the model's contextual modeling capability and enhance robustness. Finally, pseudo-labeled data is generated by model fusion to predict the test data. We implement the model fusion-based approach by summing the predicted probabilities. Specifically, the method first sums the probability results of the same class predicted by different models in each sample and then takes the maximum value. The class corresponding to that value is the final predicted class.

Second stage: The pseudo-labeled data generated in the first stage is added to the training set to achieve semantic enhancement based on the sample space dimension. We use the same backbone and model fusion for prediction to obtain the results.

Summary: The two stages of the training process have four points in common: the model backbone, the specific BERT, the semantic enhancement based on the feature space dimension, and the model fusion approach. The difference is in the training data used, the former being the original dataset and the latter a new dataset containing pseudo-labeled data fused by multi-dimensional semantic enhancement and within-task pretraining. This simple but effective implementation facilitates low-density separation between fine-grained fraud classes and enhances classification results.

3 Experiments and Analysis

3.1 Datasets

We conduct experiments on two different-scale domain-specific text classification datasets: (1) the **Telecom Network Fraud Case** (or TNFC for short) dataset consists of 82,210 fraud cases and is divided into 12 genres, and (2) the **Few-shot Patents** (or FSP for short) dataset consists of 985 patent data (including title, assignee, and abstract) and encompasses 36 genres.

3.2 Experimental Settings

For the backbone, the structure of bidirectional transformers is implemented with a specific BERT, and a single-layer variant recurrent network is implemented with BiGRU. The specific BERT is implemented using RoBERTa-wwm-ext-large (Cui et al., 2021), Chinese-LERT-large (Cui et al., 2022), and MacBERT-large (Cui et al., 2020). All three are developed by Harbin Institute of Technology. We use F1 score based on macro-averaged as our evaluation metric. We set the probability threshold for

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 206–212, Harbin, China, August 3 – 5, 2023.

#	Type	Model	F1/%
1	single-stage	input^{template}+backbone^{RoBERTa}	83.642
2	single-stage	input^{template}+backbone^{LERT}	84.195
3	single-stage	input^{template}+backbone^{MacBERT}	84.255
4	single-stage	input^{abstract}+backbone^{RoBERTa}	84.308
5	single-stage	input^{abstract}+backbone^{LERT}	84.454
6	single-stage	input^{abstract}+backbone^{MacBERT}	84.590
7	single-stage	input^{abstract}+backbone$^{RoBERTa^{pre}}$ + adv^{FGM}	84.739
8	single-stage	input^{abstract}+backbone$^{LERT^{pre}}$ + adv^{FGM}	84.821
9	single-stage	input^{abstract}+backbone$^{MacBERT^{pre}}$ + adv^{FGM}	84.910
10	single-stage	input^{abstract}+backbone$^{LERT^{pre}}$ + adv^{AWP}	84.187
11	single-stage	fusion^{1+2+3}	84.578
12	single-stage	fusion^{4+5+6}	85.179
13	single-stage	fusion^{7+8+9}	85.284
14	two-stage	two-stage$^{pseudo^{13}+7}$	85.322
15	two-stage	two-stage$^{pseudo^{13}+8}$	85.526
16	two-stage	two-stage$^{pseudo^{13}+9}$	85.752
17	two-stage	fusion$^{14+15+16}$ (**final published result**)	**85.926** (0.859259)

Table 1: Experimental results on the Telecom and Network Fraud Cases dataset

filtering pseudo-labeled data to 0.99. We set the $epsilon$ involved in the adversarial training to 0.01. We set $epochs$, $batch\_size$, and lr in within-task pretraining to 5, 100, and 2e-5, respectively. We set $encoder\_lr$, $epochs$, and $batchsize$ to 2e-5, 10, and 8 respectively. We optimize the models using AdamW (Loshchilov and Hutter, 2017) and implement the proposed models using Pytorch. All experiments are run on a single NVIDIA RTX A5000 24GB GPU.

3.3 Experimental Results and Analysis

The experimental results on the TNFC dataset are shown in Table 1. Exp.1 to Exp.13 are single-stage frameworks, and Exp.14 to Exp.17 are two-stage frameworks. Note that input* represents the input format, where input^{template} has the format **"case number: id. case description: description"**, and input^{abstract} has **only description**. $backbone$ consists of BERT, a single-layer BiGRU, and Multi-sample Dropout. $RoBERTa^{pre}$, $LERT^{pre}$, and $MacBERT^{pre}$ denote specific BERT model by within-task pertraining. $fusion$ and adv* represent the model fusion method and adversarial training approach, respectively. Meanwhile, two-stage^{a+b} represents the two-stage framework, where a indicates the approach used in the first stage and b indicates the method used in the second stage. We have experimental results that lead to the following conclusions:

(1) Comparing Exp.1 & Exp.4, Exp.2 & Exp.5, and Exp.3 & Exp.6, the practical part of the training data is the case description, and the overly detailed prompt template affects the model recognition and training efficiency.

(2) Comparing Exp.4 & Exp.7, Exp.5 & Exp.8, and Exp.6 & Exp.9, RoBERTa, MacBERT, and LERT introduce the multidimensional semantic enhancement to improve the model's ability to mine the migrated knowledge features in the domain after within-task pretraining, which lead to a significant increase in F1 score and verify the effectiveness of Within-task Pretraining and multidimensional semantic enhancement.

(3) Comparing Exp.5 & Exp.10 and Exp.8 & Exp.10, AWP adversely affects the model performance. A possible reason is that the more complex AWP is prone to overfitting. It performs less well than the simpler structured FGM under few-shot scenarios or few contents in Chinese-specific domains.

(4) After introducing the two-stage framework, the F1 score of Exp.14, Exp.15, and Exp.16 improve substantially, which verifies the effectiveness of the proposed framework. It indicates that the two-stage framework can enhance the model performance from both data and feature perspectives. The

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 206-212, Harbin, China, August 3 - 5, 2023.

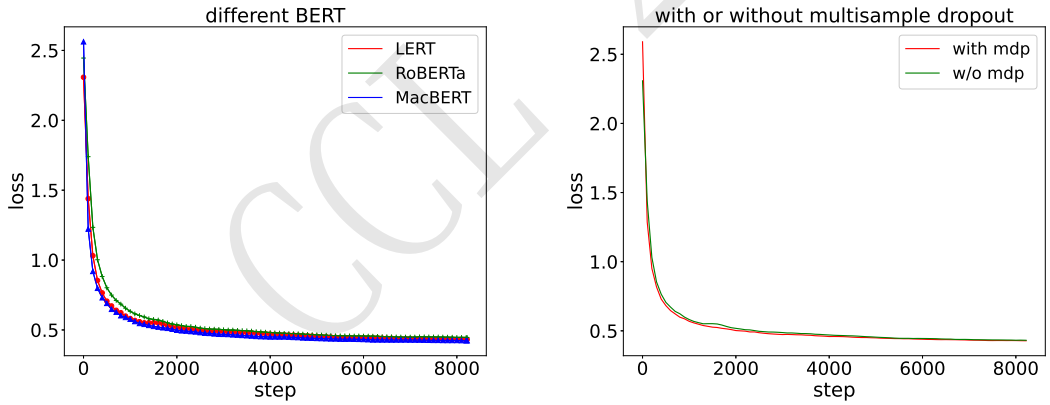
#	Model	F1/%
1	Our Model$^{two-stage}$	**85.752**
2	- two-stage	85.284
3	Our Model$^{single-stage}$	**84.910**
4	- msdr	84.753
5	- fgm	84.806
6	- within-task	84.611

Table 2: Ablation Studies

semantic information of the training corpus is maximized without introducing external knowledge. The F1 score of Exp.11, Exp.12, Exp.13, and Exp.17 illustrate the effectiveness of model fusion, and its simple implementation can improve model performance more substantially.

3.4 Efficiency Analysis

For the same experimental conditions, we output the loss variation curves in the same training epoch as shown in Figure 2. Fig. 2(a) represents the comparison of loss variation between MacBERT, LERT, and RoBERTa. Fig. 2(b) represents the comparison of loss variation between MSDR and without MSDR. It leads to the following conclusions: (1) The training convergence speed of the three types of BERT is MacBERT, LERT, and RoBERTa in order from high to low, among which MacBERT has the fastest convergence speed, and the loss is always at a relatively low level, indicating that MacBERT is more effective in this task. (2) The convergence speed of the model training is accelerated after adding MSDR. The lower mean value of loss in the convergence process indicates that MSDR can speed up the convergence of the model and avoid the overfitting problem.



(a) loss variation of different BERT (b) Loss variation with and without MSDR

Figure 2: Loss variation curves

3.5 Ablation Studies

To further validate the effectiveness of each part of the proposed framework, we conduct ablation experiments on the TNFC dataset, and the results are shown in Table 2. The validity of the two-stage framework is verified by comparing Exp.1 & Exp.2. Comparing with Exp.3, the F1 score of Exp.4, Exp.5, and Exp.6 all show different degrees of decrease, indicating the validity of all three. Among them, Exp.6 has the most significant reduction, meaning that BERT generated based on large-scale corpus training has a more substantial impact on the effect of downstream tasks after within-task pretraining.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 206-212, Harbin, China, August 3 - 5, 2023.

#	Model	F1/%
1	input+simple_backbone$^{RoBERTa^{pre}}$ + adv^{FGM}	61.021
2	input+simple_backbone$^{LERT^{pre}}$ + adv^{FGM}	60.737
3	input+simple_backbone$^{MacBERT^{pre}}$ + adv^{FGM}	60.775
4	input+simple_backbone$^{bert_for_patent^{pre}}$ + adv^{FGM}	60.146
5	fusion$^{1+2+3+4}$	64.539
6	two-stage$^{pseudo^1+1}$	61.815
7	two-stage$^{pseudo^2+1}$	61.100
8	two-stage$^{pseudo^3+1}$	61.078
9	two-stage$^{pseudo^5+1}$	**65.160**

Table 3: Experimental results on the Few-shot Patents dataset

3.6 Robustness Analysis Based on Few-shot Text Classification

To verify the robustness of our proposed framework, we conduct experiments on the FSP dataset, and the results are shown in Table 3. Due to the limitation of the prediction model size, we only use a single model for prediction in the second stage. Since the context of the training corpus is scarce in small sample scenarios, a more straightforward model structure is used for encoding. It should be specified that simple_backbone consists of BERT only. We have experimental results that lead to the following conclusions:

(1) The F1 score of Exp.9 is higher than the F1 score of Exp.5, indicating the effectiveness of our proposed two-stage framework, which is higher than the F1 score of Exp.6, Exp.7, and Exp.8. The reason is that the higher the F1 score of the first-stage model, the higher the confidence level of the prediction and the higher the quality of the generated pseudo-labeled data, affecting the prediction of the second-stage labels.

(2) Our proposed framework achieves better results in both tasks. Despite the absence of fine-tuned parameters, it still ranks at the top of the evaluation, which validates the robustness of our proposed framework.

4 Conclusion

In this paper, we propose a two-stage framework based on within-task pretraining and multi-dimensional semantic enhancement for CCL23-Eval Task 6 (FCC). It enhances the model's ability to represent the case text in feature and sample space dimensions. Experiments show that our proposed architecture achieves good results in this review, outperforming baseline models. In future work, we plan to use the framework extensively for domain text classification tasks in other scenarios. Inspired by model pruning and knowledge distillation, we try to refine the sub-stage to improve our architecture.

References

Devlin J, Chang M W, Lee K, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019: 4171-4186.

Gururangan S, Marasović A, Swayamdipta S, et al. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020: 8342-8360.

Madry A, Makelov A, Schmidt L, et al. 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

Goodfellow I J, Shlens J, Szegedy C. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

Dong Y, Deng Z, Pang T, et al. 2020. Adversarial distributional training for robust deep learning. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020: 8270-8283.

Inoue H. 2019. Multi-sample dropout for accelerated training and better generalization. arXiv preprint arXiv:1905.09788, 2019.

Lee D H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning (ICML)*, 3(2): 896.

Loshchilov I, Hutter F. 2017. Fixing weight decay regularization in adam.

Cui Yiming, Che Wanxiang, Liu Ting, et al. 2021. Pre-Training with Whole Word Masking for Chinese BERT. *IEEE Transactions on Audio, Speech and Language Processing*

Cui Yiming, Che Wanxiang, Wang Shijin, et al. 2022. LERT: A Linguistically-motivated Pre-trained Language Model. arXiv preprint arXiv:2211.05344, 2022.

Cui Yiming, Liu Ting, Qin Bing, et al. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020: 657–668.