CCL23-Eval 任务6系统报告:面向电信网络诈骗案件分类的优化策略研究

余俊晖
CVTE
junhuy@163.com

李智
CVTE
lizhi@cvte.com

摘要

电信网络诈骗案件的激增给社会带来了巨大的安全威胁，因此准确、高效地分类和检测电信网络诈骗成为了当务之急。本研究旨在针对电信网络诈骗案件分类问题，探索了一系列优化策略，并在“电信网络诈骗案件分类评测”技术评测比赛中最终成绩排名第一。本研究基于文本分类模型，并采用了BERT的继续预训练、FreeLB的对抗训练和模型融合等trick。通过BERT的继续预训练，使模型具备更好的语义理解能力和特征提取能力。而通过FreeLB的对抗训练，增强了模型的鲁棒性，使其能够更好地应对噪声和干扰。此外，本文采用模型融合的方法将多个模型的预测结果进行融合，进一步提高了分类的准确性。实验结果表明，本文的优化策略在比赛中取得了显著的成绩，证明了其在电信网络诈骗案件分类中的有效性和优越性。本研究的成果对于提高电信网络诈骗案件的分类性能具有重要意义，为相关领域的研究和实践提供了有益的参考。

关键词： 预训练；对抗训练；模型融合

CCL23-Eval Task 6 System Report: Research on Optimization Strategies for Telecom Internet fraud Case Classification

Junhui Yu
CVTE
junhuy@163.com

Zhi Li
CVTE
lizhi@cvte.com

Abstract

The proliferation of telecom internet fraud cases has brought huge security threats to society, so it is urgent to classify and detect telecom Internet fraud accurately and efficiently. The purpose of this study is to explore a series of optimization strategies for the classification of telecom Internet fraud cases, and rank first in the technical evaluation contest of "telecom Internet fraud case classification evaluation". This study is based on a text classification model and employs techniques such as BERT's continuous pre training, FreeLB's adversarial training, and model fusion. Through the continuous pre training of BERT, the model has better semantic understanding and feature extraction capabilities. Through FreeLB's adversarial training, the robustness of the model is enhanced, enabling it to better cope with noise and interference. In addition,

第二十二届中国计算语言学大会论文集，第173页-第178页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

this article adopts the method of model fusion to fuse the prediction results of multiple models, further improving the accuracy of classification. The experimental results show that the optimization strategy in this paper has made significant achievements in the competition, which proves its effectiveness and superiority in the classification of telecom Internet fraud cases. The results of this study are of great significance for improving the classification performance of telecom Internet fraud cases, and provide a useful reference for research and practice in related fields.

Keywords: pre-training , adversarial training , model fusion

1 引言

随着信息技术的迅猛发展，电信网络诈骗案件在全球范围内呈现出愈演愈烈的态势，给个人和组织的财产和安全造成了巨大威胁。电信网络诈骗的手段和技术不断更新换代，使得传统的防御方法逐渐失去了效果。因此，准确、高效地识别和分类电信网络诈骗案件成为了当今社会安全领域的重要任务。

在此背景下，文本分类技术作为一种关键的手段被广泛应用于电信网络诈骗案件的防范和打击。通过对电信网络诈骗案件文本进行分类，可以实现对嫌疑案件的及时发现和预警，从而有效减少诈骗行为造成的损失。然而，由于电信网络诈骗案件文本的复杂性和多样性，传统的文本分类方法在处理此类问题时面临着挑战。

为了解决这些挑战，本研究着重探索了一系列优化策略，旨在提升电信网络诈骗案件分类的性能和鲁棒性。其中，我们采用了BERT(Devlin et al., 2018)的继续预训练技术，通过在大规模电信网络诈骗数据上进行预训练，使模型具备更好的语义理解和特征提取能力。同时，我们引入了FreeLB(Zhu et al., 2019)的对抗训练方法，以增强模型对干扰和噪声的抵抗能力。此外，我们还采用了模型融合的策略，通过集成多个分类模型的预测结果，进一步提高分类性能。

2 模型介绍

本文模型结构如图1所示，基线模型采用BERT(包括其变种)+Linear的架构。并采用预训练、对抗训练和模型融合等三种主要优化策略提升基线模型的性能。

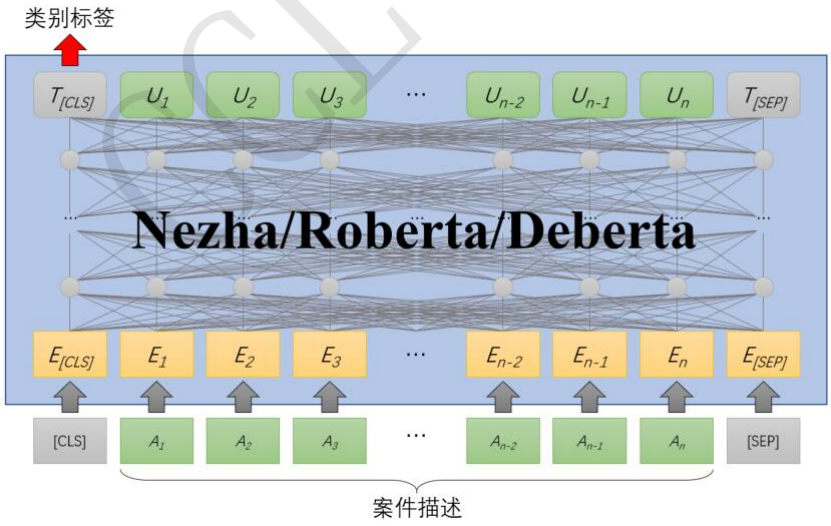

Figure 1: 模型结构

2.1 预训练

有效的预训练可以提升模型在下游任务微调的性能。本文提取数据集中的案情描述文本，在预训练阶段添加MLM预训练任务，通过无监督学习使得预训练语言模型获得案件领

域的知识，从而使模型具备对案件文本更好的语义理解和特征提取能力。MLM预训练使用了与Roberta(Cui et al., 2021)一致的方式，将输入的案情描述文本随机遮蔽，即为存在15%的概率决定对该token进行修改，其中有80%的概率改为"[MASK]",有10%的概率被替换为一个随机的token,有10%的概率保持不变。MLM预训练任务使用交叉熵损失进行训练，其损失表示为公式1：

$$L_{mlm} = -\sum_{i=0}^{V-1} y_i^{mask} log(p_i^{mask}) \tag{1}$$

其中，V为模型词表大小，y_i^{mask}是遮蔽字符的标签，p_i^{mask}表示模型预测的概率。

本文在预训练阶段，分别预训练了三种中文模型，分别为nezha、Roberta和Deberta。在使用Nezha-base-wwm[0]预训练语言模型时，输入序列的最大长度为1024，在使用chinese-roberta-wwm-ext-large[1]与Deberta[2]预训练语言模型时，输入序列的最大长度为512。

2.2 对抗训练

为了增强模型对干扰和噪声的抵抗能力，本文实验了PGD(Madry et al., 2017)、FGM(Miyato et al., 2016)、FreeLB等对抗训练技巧提升模型的鲁棒性，通过实验性能对比，最终主要采用了FreeLB对抗训练。FreeLB的核心思想是通过增加对抗样本的生成空间，引入自由生成的方法来提高模型的鲁棒性。传统的对抗训练方法通常使用固定的扰动方法来生成对抗样本，这可能会限制模型的泛化能力和鲁棒性。相比之下，FreeLB提出了自由生成的概念，它允许生成过程中的扰动更加多样和自由，从而提供更丰富的训练信号。都是在word embedding空间上加入扰动，然后对扰动后的embedding进行look up，得到的词向量再喂给模型。其原理伪代码如表1所示。

Table 1: FreeLB算法伪代码

输入： 原始训练数据集D
输出： 防御后的模型参数

Procedure AdversarialTraining(D):
  初始化BERT模型参数
  while 未达到停止条件**do**:
    for each 样本(x, y) in D **do**:
      计算原始样本的word embedding E_x
      生成对抗样本的word embedding E_{adv} using Freelb算法
      将E_x和E_{adv}分别作为输入喂给BERT模型
      前向传播计算模型的输出O_x 和O_{adv}
      计算原始样本的损失函数L_x
      计算对抗样本的损失函数L_{adv}
      计算总体损失函数$L_{total} = \alpha \cdot L_x + \beta \cdot L_{adv}$，其中$\alpha$ 和β 是权重
      反向传播更新BERT模型参数
  return 更新后的模型参数

2.3 模型融合

模型融合是一种常用的技术，在文本分类比赛中被广泛应用，旨在提高分类模型的性能和泛化能力。模型融合通过结合多个不同的分类模型的预测结果，从而得到更准确、更稳定的最终预测结果。本文的模型融合的方法是对于每个分类模型的输出概率进行简单的相加，得到最终的融合概率分布，进一步求取最大概率的下标获取对应的类别标签。

[0]https://huggingface.co/sijunhe/nezha-base-wwm
[1]https://huggingface.co/hfl/chinese-roberta-wwm-ext-large
[2]注：这里使用了两个权重进行实验，320M的进行了预训练，710M的没有进行预训练，相关权重链接：1、Erlangshen-DeBERTa-v2-320M-Chinese：https://huggingface.co/IDEA-CCNL/Erlangshen-DeBERTa-v2-320M-Chinese；2、Erlangshen-DeBERTa-v2-710M-Chinese：https://huggingface.co/IDEA-CCNL/Erlangshen-DeBERTa-v2-710M-Chinese

3 实验设置

3.1 数据集介绍

本文数据集来自于“CCL 2023 电信网络诈骗案件分类评测”任务[3]，该数据集案件文本内容为案情简述，即为受害人的笔录，由公安部门反诈大数据平台导出。去除了案件文本中的姓名、出生日期、地址、涉案网址、各类社交账号以及银行卡号码等个人隐私或敏感信息。最终将案件类别分为12个类别，具体类别信息及分布情况如表2所示。

类别名称	样本数量
刷单返利类	35459
冒充电商物流客服类	13772
虚假网络投资理财类	11836
贷款、代办信用卡类	11105
虚假征信类	8464
虚假购物、服务类	7058
冒充公检法及政府机关类	4563
冒充领导、熟人类	4407
网络游戏产品虚假交易类	2155
网络婚恋、交友类（非虚假网络投资理财类）	1654
冒充军警购物类诈骗	1092
网黑案件	1197

Table 2: 电信网络诈骗案件分类数据集类别及分布

3.2 实验参数设置

本文实验参数设置如表3所示，并且如图2分析了案情文本的长度分布，因此实验了两种输入长度策略，分别为1024和512。所有实验均使用Pytorch深度学习框架，并在一台A6000服务器上进行。

3.3 评价指标

评测性能时，本文参照比赛任务要求，主要采用宏平均F1值作为评价标准，即对每一类计算F1值，最后取算术平均值，其计算方式如公式2：

$$Macro - F1 = \frac{1}{n}\sum_{i=1}^{n} F1_i \tag{2}$$

其中$F1_i$为第i类的$F1$值，n为类别数，在本任务中n取12。

[3]https://github.com/GJSeason/CCL2023-FCC

模型参数	预训练	微调
Mask probabity	0.15	-
训练轮数	5	3
学习率	5e-5	2e-5
权重衰减系数	0.01	0.01
batch size	128	64
随机种子	42	42
输入序列最大长度	1024/512	1024/512
优化器	AdamW	AdamW
Warm up ratio	0.1	0.1
Lr schedule	0.1	0.1

Table 3: 实验参数设置

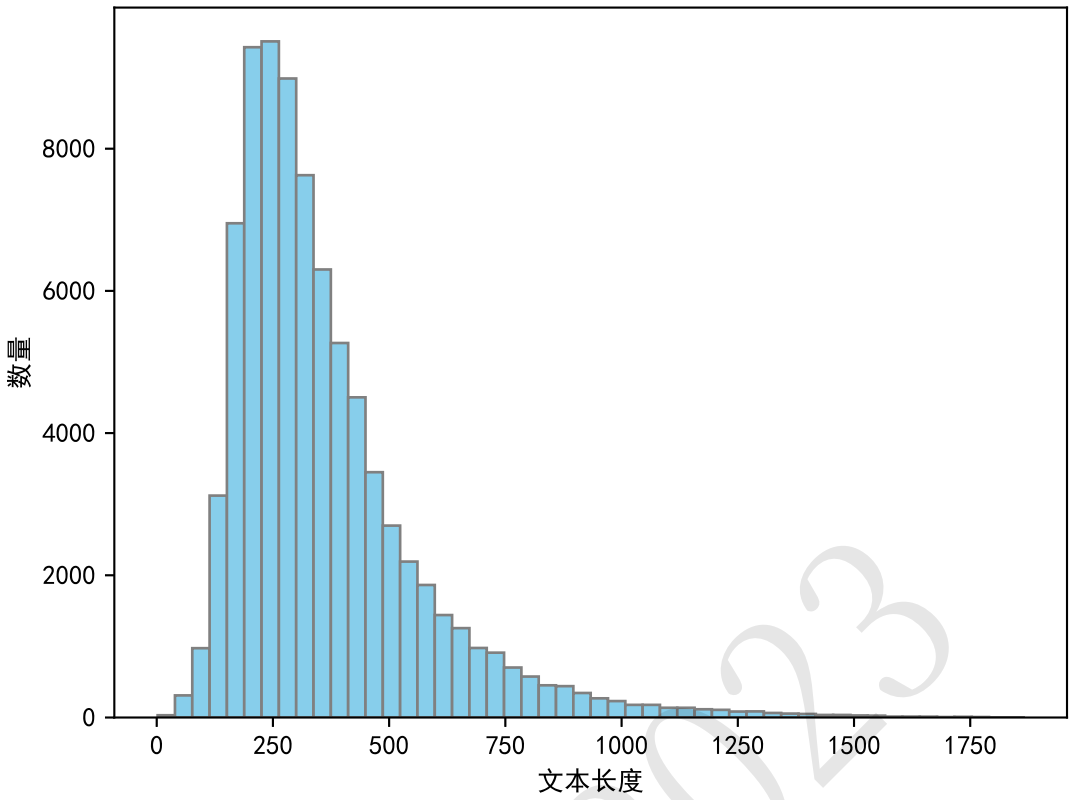

Figure 2: 案情描述长度分布

4 评测结果

表4展示了本文模型的线上评测结果, 本文方法取得了最佳性能。

提交模型	数据划分	输入长度	线上得分
①chinese-roberta-wwm-ext-large(预训练前)	9：1	512	0.846889077
②chinese-roberta-wwm-ext-large(预训练后)	全量数据	512	0.8589623748
③nezha-base-wwm(预训练前)	9：1	1024	0.8582825466
④nezha-base-wwm(预训练前)	全量数据	1024	0.8596631322
⑤nezha-base-wwm(预训练后)	全量数据	1024	0.8619108232
⑦Erlangshen-DeBERTa-v2-320M-Chinese(预训练前)	9：1	512	0.8582825466
⑧Erlangshen-DeBERTa-v2-320M-Chinese(预训练后)	全量数据	512	0.8595019184
⑨Erlangshen-DeBERTa-v2-710M-Chinese(预训练前)	全量数据	512	0.8611721092
④+⑦	-	-	0.8621353191
⑤+⑦	-	-	0.864710651
⑤+⑧+⑨	-	-	0.8660677395

Table 4: 评测最终公布结果

5 结果分析与讨论

模型对比：本文使用了多个不同的预训练模型进行评测，包括chinese-roberta-wwm-ext-large、nezha-base-wwm和Erlangshen-DeBERTa-v2系列模型。从线上得分来看，预训练后的模型普遍表现比预训练前的模型更好。

数据划分：大部分模型使用了9:1的数据划分比例，即将数据集划分为训练集和验证集。只有两个模型（②和③）使用了全量数据进行训练。使用全量数据进行训练通常会有更好的效果，因为模型可以更充分地学习数据中的模式和规律。

输入长度：所有模型的输入长度都为512或1024。较长的输入长度可以提供更多的上下文信息，有助于模型理解文本的语义和逻辑关系。然而，较长的输入长度也会增加模型的计算负担和训练时间。

模型融合：根据给出的实验结果，可以看出模型组合⑤+⑧+⑨获得了最高的线上得分（0.8660677395）。这是因为这个组合中的模型相互补充，模型的融合能够有效的提升模型的泛化能力。

此外，由于比赛提交次数有限，未提交验证FreeLB对抗训练对于结果的影响，根据本人在其他比赛的经验，该策略能有效提升模型的鲁棒性。

6 结论

本研究针对电信网络诈骗案件的分类问题，通过采用一系列优化策略和技巧，包括BERT的继续预训练、FreeLB的对抗训练和模型融合，取得了显著的成果。实验结果在“CCL23-Eval-任务6-电信网络诈骗案件分类评测”技术评测比赛中最终成绩排名第一，证明了所提出的优化策略在提高电信网络诈骗案件分类性能方面的有效性和优越性。

通过BERT的继续预训练，研究者使模型具备更好的语义理解和特征提取能力，有助于准确地分类和检测电信网络诈骗案件。同时，通过FreeLB的对抗训练，模型的鲁棒性得到增强，使其能够更好地处理噪声和干扰，提高了分类的准确性。此外，采用模型融合的方法将多个模型的预测结果进行融合，进一步提升了分类的效果。

参考文献

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

第二十二届中国计算语言学大会论文集，第173页-第178页，哈尔滨，中国，2023年8月3日至5日。