CCL23-Eval 任务6系统报告：基于深度学习的电信网络诈骗案件分类

李晨阳[1,2]，张龙[1,2]，赵中杰[1,2]，郭辉[1,2]
[1]中原工学院 前沿信息技术研究院，河南 郑州 450007
[2]河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007
2312826399@qq.com

摘要

文本分类任务作为自然语言处理领域的基础任务，在面向电信网络诈骗领域的案件分类中扮演着至关重要的角色，对于智能化案件分析具有重大意义和深远影响。本任务的目的是对给定案件描述文本进行分类，案件文本包含对案件的经过脱敏处理后的整体描述。我们首先采用Ernie预训练模型对案件内容进行微调的方法得到每个案件的类别，再使用伪标签和模型融合方法对目前的F1值进行提升，最终在CCL23-Eval任务6电信网络诈骗案件分类评测中取得第二名的成绩，该任务的评价指标F1值为0.8628，达到了较为先进的检测效果。

关键词： 文本分类；网络诈骗；预训练模型；伪标签；模型融合

System Report for CCL23-Eval Task 6: Classification of Telecom Internet Fraud Cases Based on Deep Learning

Chengyang Li[1,2], Long Zhang[1,2],Zhongjie Zhao[1,2],Hui Guo[1,2]
[1]Frontier Information Technology Research Institute,
Zhongyuan University of Technology, Zhengzhou 450007 China
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou China
2312826399@qq.com

Abstract

As the basic task in the field of Natural language processing, text classification plays a crucial role in the case classification in the field of telecom Internet fraud, and has great significance and far-reaching impact on intelligent case analysis. The purpose of this task is to classify the given case description text, which contains the overall description of the case after being desensitized. We first used Ernie's pre training model to fine tune the case content to get the category of each case, and then used pseudo tags and model fusion methods to improve the current F1 value. Finally, we won the second place in the CCL23-Eval task 6 Telecom Internet fraud case classification evaluation. The evaluation index F1 value of this task is 0.8628, achieving a more advanced detection effect.

Keywords: Text classification , Internet fraud , Pretraining model , Pseudo label , Model fusion

基金项目： 河南省高等学校重点科研项目（22B520054）；嵩山实验室预研项目（YYJC032022021）；中原工学院自然科学基金（K2023MS021）

1 任务介绍

诈骗案件分类问题是打击电信网路诈骗犯罪过程中的关键一环，通过根据不同的诈骗方式、手法等对其分类，一方面能够便于统计现状，有助于公安部门掌握当前电信网络诈骗案件的分布特点，进而能够对不同类别的诈骗案件作出针对性的预防、监管、制止、侦查等措施，另一方面也有助于在向群众进行反诈宣传时抓住重点、突出典型等。然而，人工分类的方法不仅耗时耗力，还容易受到人为主观因素的干扰，分类结果难以达到高度准确。在现代社会，电信网络诈骗案件种类繁多、数量巨大，人工分类的方法已经难以满足需求。面对人工智能技术高速发展的时代，急需一项更加高效、准确的自动分类方法来解决此问题。为此，CCL2023发布了依靠深度学习技术解决这一问题的任务。该任务提供的数据集由公安部反诈大数据平台导出，数据样例如图1所示，由82210条训练集和10276条测试集组成。如图1所示，每条数据由案件编号、案情描述、案件类别3部分组成。案件文本内容为案件简述，即为受害人笔录，全部经过脱敏处理去除个人隐私或敏感信息。案件类别标签采用的是反诈大数据平台导出的12个类别，类别体系来源于反诈大数据平台的分类标准，主要依据受害人的法益及犯罪分子的手法进行分类。该任务需要通过测试集的案件内容来预测所属案件的类别。在评测性能时，主要采用宏F1值作为评价标准，即对每一类计算F1值，最后计算算数平均值。


```
{
  "案件编号": 48788,
  "案情描述": "2022年111月11日，接报警称其在家中，于当天晚上18时左右接到陌生电话对方自称时快递客服人员，称因疫情原因要销毁快递并会给我进行补偿，后下载了一个“会讯通云会议”APP，后称其操作失误银行卡冻结了，要解冻的话要先打钱到对方提供的银行卡上，按对方操作向对方转账了19999元。（受害人已向嫌疑人转账19999元）（嫌疑人电话）（涉案网址）",
  "案件类别": "冒充电商物流客服类"
},
{
  "案件编号": 56818,
  "案情描述": "2022年11月28日16时30分至2022年11月28日17时40分，报警人在里，因房子贷款要开结清证明，在支付宝上搜索贷款公司上海尚城消费金融有限公司，就在百度上搜索这家公司的客服，联系对方电话（），电话那边问我有没有20000元以上的卡我只有这样才能把结清证明发到工商银行上，按照客服的要求汇款给对方指定账户99976元，报警人叫对方退钱，对方让报警人接着操作，发现被骗，损失人民币99976元。（受害人银行卡：；嫌疑人卡号：，，吉林银行）",
  "案件类别": "贷款、代办信用卡类"
},
```


图 1. 数据样例

2 相关工作

文本分类在自然语言处理和文本挖掘中具有重要的作用，通过不断学习文本特征进行预测分类，在各个方面的研究中都具有十分重要的意义和研究价值 (Minaee et al., 2021)。传统的文本分类是基于机器学习方法 (Cheng, 2020)，包括支持向量机、决策树、朴素贝叶斯等，但这些方法都只解决了词汇层面的问题，无法有效学习和反映语句之间的语义相关性和深层语义特征。

近年来，深度学习技术在计算机视觉和自然有语言处理领域都取得了显著的进展。在自然语言处理任务中，基于深度学习的文本分类模型备受关注和研究，如CNN (Wan et al., 15) (Wang et al., 2017)、RNN (Le et al., 2017) (Cui et al., 2019)、GNN (Yao et al., 2018)、Attention (Kim et al., 2018)和预训练模型。它们在文本分类等自然语言处理任务中都表现出了优秀的效果。特别是预训练模型，由于在预训练时就已经接触大量的文本数据，因此能学习到更加丰富的语义信息，使其在文本分类等任务中具有更高的准确性和泛化能力。

3 分类模型介绍

Bert-wwm (Cui et al., 2019)是哈尔滨工业大学和科大讯飞联合发布的一个Bert升级版本，主要更改了原预训练截断的训练样本生成策略。相较于Bert，Bert-wwm的改进是用mask标签替换一个完整的词而不是字词。中文和英文不同，英文最小的token是一个单词，而中文最小的token却是字，词由一个或多个字组成，且每个词之间没有明显的分割，包含更多信息的是词。对比基于字的掩码，基于词的掩码能够让模型学到更多的语义信息。

第二十二届中国计算语言学大会论文集，第167页-第172页，哈尔滨，中国，2023年8月3日至5日。

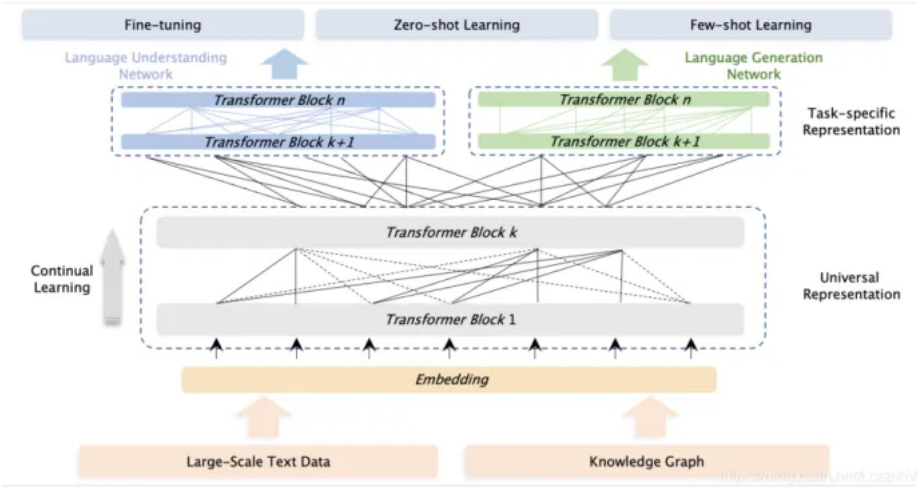

图 2. Ernie3.0结构图

Ernie3.0 (Sun et al., 2019)模型是一种融合了自回归网络和自编码网络的大规模知识增强模型，在纯文本和大规模知识图谱组成的语料库上训练得到。如图2所示，通过将大规模知识图谱的实体关系与大规模文本数据同时输入到预训练模型中进行联合掩码训练，促进了结构化知识和无结构文本之间的信息共享，大幅度提升了模型对知识的记忆和推理能力。通过这种方式，Ernie能够捕捉到更加细微的语义区别，并将其与知识图谱中的实体关系结合，从而实现更精准的分类。

模型融合是一种训练多个模型并进行融合的方法，旨在通过融合模型结果来超越单个模型的表现。常用的模型的融合方法共有三种，第一种是投票法，适用于分类任务，即对多个学习模型的预测结果过进行投票，以少数服从多数的方式来确定最后的结果，还可以根据人工设置或者根据模型评估分数来设置权重。第二种是平均法，适用于回归和分类任务，即对于学习模型的预测概率进行平均。第三种是交叉融合法，主要思路就是把原始的训练集先分成两部分，例如按9：1划分训练集和测试集，在第一轮训练时，使用训练集训练多个模型，然后对测试集进行预测，在第二轮训练时，直接用第一轮训练的模型在测试集上的预测结果作为新特征继续训练。

4 实现方法

如图3所示，我们先对文本内容进行预处理，然后对当前主流深度学习模型进行评估，选出表现较好的模型作为我们的基线模型，最后用伪标签、模型融合等方法提升F1指标。

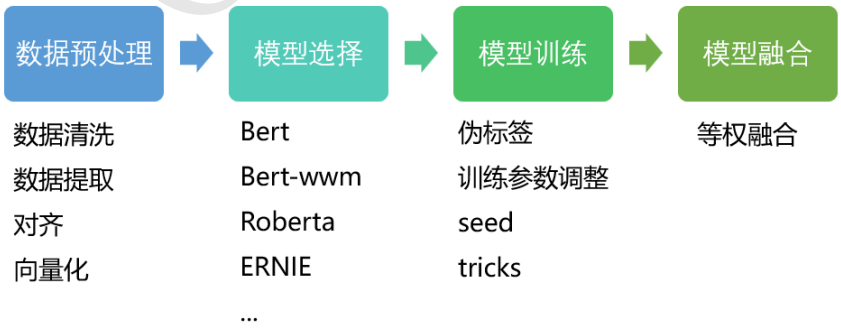

图 3. 模型的数据预处理、预测以及后处理过程

4.1 数据预处理

在获取训练集后，我们首先进行了清洗工作，我们发现训练集中存在29条重复文本和2条

内容相同但标签却不相同的文本，因此我们删除这30条数据。接下来分析文本内容，我们发现经过脱敏处理的文本有多个重复标点或符号的情况，我们在代码中用正则表达式对这些符号进行规范化。接下来我们编写了代码，对数据集中的内容进行提取，包括案件描述和案件类别。这些文本的平均长度为362，其中75%的文本长度在437以下，最大的文本长度为1865。由于Ernie预训练模型能够处理的最大长度为512，为使模型在训练中能学到更多特征，我们选择512为最大长度，来对案件描述的文本进行截断和补齐。

4.2 模型选择

由于本任务仅提供了训练集和无标签的测试集，因此我们按照9：1的比例把训练集重新划分为训练集和验证集。为了确保在后续步骤中可以控制每次分割训练集和验证集的异同，我们在代码中动态实现了数据集的随机划分，并固定随机种子为42。如表1所示，我们选取了当前主流的预训练模型，包括Bert，Bert的变体模型和Ernie，在此基础上对每个模型进行微调。从直观上来讲，网络模型越大，层数越深，学习能力越强大，因此我们的Ernie模型选择了20层网络结构的ernie3.0-xbase进行测试,同时我们后面提出的Ernie均指ernie3.0-xbase。经过验证，我们发现只有Bert-wwn和Ernie两个模型在验证集上超过了任务的baseline，因此我们把这两个模型当作我们任务的基线模型。

模型	bert-base	bert-wwm	bert-wwm-ext	ernie3.0	roberta-wwm-ext
F1	0.8475	0.8506	0.8486	0.8512	0.8483

表 1. 各模型评测F1分数

4.3 数据分析

为了提高F1分数，我们对每个类别的总数和每个类别对应的F1值进行了比较，如图4所示。我们发现数据分布不平均，部分标签对应的数据量较少的问题。例如虚假购物、服务类和网络婚恋、交友类两个类别的F1值相对于较低，且它们对应的数据集数量也相对较少。为了提升这两个类别的F1分数，我们采取增加数据集的方式来使模型学到更多的相关特征。我们首先采用同义词替换、随机词插入和随机词删除等方法对这两个标签的数据集进行了数据增强，在我们的两个基线模型上训练后，验证集分数都有很大的提升，但在测试集上却低于了官方的baseline，出现了过拟合状态，这表明数据增强这种方法可能对该任务不起作用。于是我们采用了同样能增加训练集数量的伪标签方法。

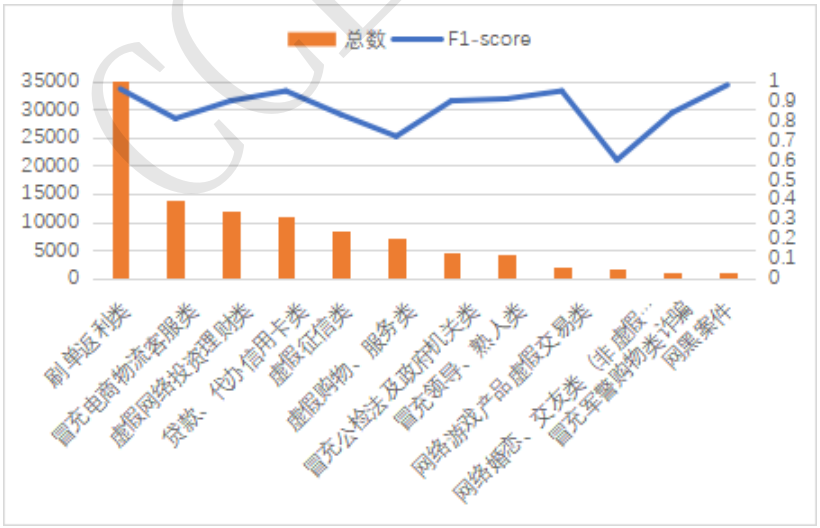

图 4. 各类别总数对应的F1值

4.4 伪标签

伪标签方法主要是将模型对无标签的测试数据的预测结果加入到训练集中 (Rizve et al.,

2021),从而增大数据量以提升模型效果。这种方法适用于模型精度较高的情况。此时我们模型预测的准确率已经接近90%，故可以采用该技巧。我们将Bert和Ernie两个模型进行微调，并取两者预测标签相同的部分加入到训练集重新训练。每次的伪标签数量都接近1万条，这表明我们的数据集在原本数量的基础上又增加了1万条数据。经过4轮伪标签法的训练后，F1值从0.85提升到了0.8616。

4.5 模型融合

经过多轮的伪标签训练后，筛选出的伪标签会越来越接近，最终模型达到了拟合的状态，此时再进行后续的伪标签方法已经不能够再提升测试集的F1值。于是我们采用模型融合的方法来进一步提升F1分数。关于模型融合，周志华教授的机器学习一书中提到：模型融合需要好而不同，即模型差异越大，融合效果越好。我们从两方面来增加差异化，一是使用不同的两个模型，Bert和Ernie。二是重新划分训练集和验证集来改变模型输入。在使用这两个模型对测试集进行预测时，我们没有直接输出预测的类别，而是输出了每个类别的概率，然后使两个模型预测的类别概率进行等权相加，得出模型融合后预测的新概率，最后选取具有最大概率的那一个类别作为预测结果。最终我们的分数由0.8616提升为0.8628。

5 实验结果

如表2所示，我们列出了我们的确定的两个基线模型Bert-wwm和Ernie在单模型情况下、数据增强、伪标签和模型融合方法下的F1值，相对于任务官方的基线模型0.8503，我们在此基础上提升了0.0125的分数。这再次证明了我们方法的有效性。

模型	F1
Bert-wwm	0.8506
Bert-wwm+数据增强	0.8300
Bert-wwm+伪标签	0.8608
Ernie3.0	0.8512
Ernie3.0+数据增强	0.8410
Ernie3.0+伪标签	0.8616
Bert-wwm+伪标签+Ernie3.0+模型融合	0.8628

表 2. 模型在伪标签和模型融合下的F1分数

如表3所示，展示了我们的模型具体参数。

模型	Max-len	Batch-size	seed	epoch	Learning-rate
Bert-wwm	512	18	42/43	5	2e-5
Ernie3.0	512	22	42/43	5	2e-5

表 3. 模型参数

6 总结

通过大量实验发现，对于本任务数据集，大多数模型预测的结果分数相近，并且由于本任务数据集数规模并不算小，因此采用类似随机词插入和随机词删除等通过添加噪声来实现数据增强的方法对本任务并没有提升。反而，使用伪标签方法可以增加数据集的规模，提升测试集的F1分数，并增加模型的泛化性。使用多轮伪标签方法后，后续筛选得出的伪标签几乎不会有变化，导致模型的性能不再有提升。这时可以采用模型融合技术，取差异较大的多个模型，分别学习不同的输入，使得多个模型之间学到的知识尽量不同，这样使得多个模型可以更好的融合，提高性能。

进一步优化方面，针对训练集存在的过拟合问题，可以考虑在划分训练集和验证集时进行数据均衡。使用伪标签方法时，尝试在预测结果时对输出的预测概率进行阈值判断，选取概率

较高的结果作为伪标签加入训练集。使用模型融合时，可以先采用五折交叉验证法来训练多个模型，然后在对多个预测结果取平均。

参考文献

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*,363:366 – 374.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504-3514.

Jiang Cheng. 2020. Research and implementation of Chinese long text classification algorithm based on deep learning. *University of the Chinese Academy of Sciences (Institute of artificial intelligence, Chinese Academy of Sciences)*.

Le, H. T., Cerisara, C., and Denis, A. 2017. Do convolutional networks need to be deep for text classification? *arXiv preprint arXiv:1707.04108*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. 2021. Deep learning based text classification: a comprehensive review. *ACM computing surveys (CSUR)*,54(3), 1-40.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. 2021. In defense of pseudo-labeling:an uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Seonhoon Kim, Jin Hyun Hong, inho Kang, and nojun kwak. 2019. Semantic sense matching with densely connected recurrent and co-attentive information. *Proceedings of the AAAI Conference on Artificial Intelligence*,33(01), 6586-6593.

Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., and Cheng, X. 2016. A deep architecture for semantic matching with multiple positive sense representations. *Proceedings of the AAAI Conference on Artificial Intelligence*,30(1). https://doi.org/10.1609/aaai.v30i1.10342.

Wang, Z., Hamza, W., and Florian, R. 2017. Bilateral multi-perspective matching for natural language sentences. *In procedures of the twenty Sixth International Joint Conference on artistic intelligence*, ijcai-17, pages 4144 – 4150.

Yao, L., Mao, C., and Luo, Y. 2019. Graph revolutionary networks for text classification. *In Proceedings of the AAAI conference on artificial intelligence*,Vol. 33, No. 01, pp. 7370-7377

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, danxiang Zhu, Hao Tian, and Hua wuErnie. 2019. Enhanced representation through knowledge. *arXiv preprint arXiv:1904.09223*.