一种基于 BERT 预训练模型的电商工单分类方法

毛新年，徐元甲
上海杉达学院，上海 200136

DOI: 10.61369/TACS.2026080006

摘　　要： 电商行业规模不断扩大，平台客服工单数量激增，传统人工分类存在成本高、处理滞后、分类标准不统一等短板；同时浅层文本分类模型无法适配工单口语化、多歧义、样本类别失衡的文本特点。为此，本文提出融合类别不平衡优化与人机协同阈值调控的 BERT 电商工单自动分类方法。方法以 BERT 预训练模型提取工单深层语义，弥补传统模型语义表征能力不足的问题；通过重采样与损失加权方案改善类别分布失衡引发的分类偏差；构建人机协同动态阈值机制，划分高置信度自动处理样本与低置信度人工复核样本，平衡分类精度与业务处理效率。基于 5000 条真实电商工单数据集开展实验，基础 BERT 模型准确率达 92.3%、宏平均 F1 值 91.4%，较 SVM、TextCNN 基线模型分别提升 11.1、4.8 个百分点；叠加类别平衡采样与数据增强后，准确率、宏平均 F1 值再度提升 4.7、5.5 个百分点。落地应用结果表明，该方案可大幅缩减人力投入，日均 5000 条工单场景下，分类人力由 6 人降至 1 人，每年节约人工成本 12 至 15 万元。该方法既能提升工单自动化处理效率，又兼具理论研究价值与实际落地价值。

关 键 词： BERT；电商工单；文本分类；类别不平衡；人机协同

An E-commerce Work Order Classification Method Based on BERT Pre-trained Model

Mao Xinnian, Xu Yuanjia
Shanghai Sanda University, Shanghai 200136

Abstract: With the continuous expansion of the e-commerce industry, the volume of customer service work orders on platforms has surged. Traditional manual classification suffers from high costs, delayed processing, and inconsistent classification standards. Meanwhile, shallow text classification models fail to adapt to the characteristics of e-commerce work orders, such as colloquial language, multiple ambiguities, and imbalanced sample categories. To address these issues, this paper proposes an automatic e-commerce work order classification method based on the BERT model, integrating class imbalance optimization and human-machine collaborative threshold regulation. The method utilizes the BERT pre-trained model to extract deep semantics from work orders, compensating for the insufficient semantic representation capability of traditional models. It employs resampling and loss weighting schemes to mitigate classification bias caused by imbalanced class distributions. Furthermore, a human-machine collaborative dynamic threshold mechanism is constructed to separate high-confidence samples for automatic processing from low-confidence samples for manual review, thereby balancing classification accuracy and business processing efficiency. Experiments conducted on a real-world dataset of 5,000 e-commerce work orders demonstrate that the baseline BERT model achieves an accuracy of 92.3% and a Macro-F1 score of 91.4%, outperforming the SVM and TextCNN baseline models by 11.1 and 4.8 percentage points, respectively. After incorporating class-balanced sampling and data augmentation, the accuracy and Macro-F1 score are further improved by 4.7 and 5.5 percentage points. Practical application results indicate that this scheme can significantly reduce labor input. In a scenario processing 5,000 work orders daily, the required personnel for classification is reduced from six to one, saving 120,000 to 150,000 RMB in labor costs annually. This method not only enhances the efficiency of automated work order processing but also holds both theoretical research value and practical application value.

Keywords: BERT; e-commerce work orders; text classification; class imbalance; human-machine collaboration

Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.

引言

（一）电商行业工单处理业务背景

数字经济带动线上零售快速发展，电商用户、订单规模持续扩张，2025 年末我国网购用户 9.37 亿，电商交易总额 46.73 万亿元 [1]。平台衍生出咨询、退换货、物流投诉等各类客服工单，日工单量可达数万条 [2]。工单处理效率直接影响用户体验，同时关系平台运营成本，高效工单分类处理已成为电商运营亟需解决的关键问题。

传统工单依靠人工逐条阅读、按规则分类派单，存在三大痛点：一是大批量工单需大量人力，成本高；二是人工判定主观性强，分类标准不统一；三是高峰期工单堆积，处理滞后，用户体验差。采用文本分类技术实现智能分拣，可替代重复人工、统一标准、缩短处理时长，是电商客服数字化转型刚需。但电商工单口语化、网络用语多、文本长短不一、样本类别失衡，通用模型易出现小众类别识别差、模糊工单错分问题，需针对性优化改进。

（二）国内外文本分类研究和电商工单分类存在主要问题

国内外学者就文本自动分类问题围绕文本特征提取、分类模型优化开展大量研究。早期文本分类依托传统机器学习算法，以词袋模型、TF-IDF 作为文本特征，结合 SVM、朴素贝叶斯、随机森林完成分类任务 ,Joachims 将 SVM 应用于长文本分类，验证其在高维稀疏文本特征下的稳定性，但 TF-IDF 仅能统计词频，无法捕捉词语上下文语义关联，对口语化短文本适配性较差 [3][4]。随着深度学习发展，卷积神经网络 TextCNN、循环神经网络 RNN/LSTM 被用于文本建模，通过局部卷积、时序序列提取浅层语义，相较传统机器学习模型分类性能提升，但静态词向量无法根据上下文动态调整语义，存在一词多义识别缺陷 [5][6]。2018 年 提出的 BERT 预训练语言模型采取基于双向 Transformer 结构与大规模无标注语料预训练，可生成上下文感知动态词向量，彻底解决静态词向量语义模糊问题，成为各类文本分类任务基准模型。Devlin 等验证 BERT 在文本分类、情感分析等下游任务上超越 CNN、RNN 系列模型 [7]。但针对电商工单等垂直领域的定制化优化研究仍存在空白 , 同时对电商工单类别不平衡的现实问题较少考虑 [8]，少量研究引入 BERT 做售后工单分类，仅简单微调模型参数，未结合业务设计人机协同流程，模型输出结果无法直接落地客服系统，实用性受限。综上当前电商工单自动分类存在如下问题：首先是语义特征提取能力不足，SVM、TextCNN 依赖静态词向量，无法区分工单中多义词、口语化模糊表述，相似语义不同类别工单易出现错分；其次是未解决类别不平衡问题：电商工单中物流咨询、普通退换货等大类样本占比超 80%，产品质量投诉、虚假宣传等小众类别样本稀缺，模型训练时偏向多数类，小众工单召回率极低；模型输出无业务适配机制：现有模型仅输出固定分类标签，未提供置信度参考，模糊、歧义工单直接自动分类易产生大量错误工单，后续人工修正成本更高；最后是缺乏人机协同处理流程：纯自动分类容错率低，全人工分拣效率低下，现有研究未搭建 “机器自动分拣 + 人工复核” 协同架构，难以平衡准确率与处理效率。

（三）本文主要贡献与论文结构安排

针对现有缺陷，本文构建基于 BERT 的电商工单智能分类方案，主要创新有三点：依托 BERT 提取工单深层语义，适配口语化、长短不均的工单文本；结合重采样与损失加权缓解样本失衡，提升小众工单识别精度；设置置信度动态阈值实现人机协同，高置信度工单自动处理、低置信度工单人工复核，兼顾精度与效率。依托真实工单数据集开展多组对比实验，验证各模块有效性，提供可落地工程方案。

本文章节安排如下：引言介绍研究背景、现状、问题与创新；第二部分详述 BERT 分类方法、不平衡优化与人机协同阈值设计；第三部分完成多维度对比实验并分析结果；第四部分总结研究，说明不足并展望后续工作；文末附参考文献。

一、基于 BERT 预训练模型的电商工单分类方法

（一）BERT 预训练模型基础原理

BERT 是基于 Transformer 双向编码器的预训练模型，由多层双向 Transformer 编码器堆叠而成，依托 MLM 掩码语言模型、NSP 下一句预测两大任务在海量通用文本无监督预训练，习得语义、语法与上下文关联信息。其自注意力机制能建模全局语义、捕捉长距离依赖，弥补 TextCNN 仅提取局部特征、LSTM 单向建模的不足。本文采用中文 BERT 权重处理电商工单，输入融合词、分句、位置三类嵌入，文本首尾添加 [CLS]、[SEP] 标识，[CLS] 向量汇总全文语义送入全连接层完成分类。模型先习得通用中文知识，再用标注工单微调迁移至电商领域，减少标注数据需求，优化小样本分类性能，下图是 BERT 模型进行文本分类原理图。

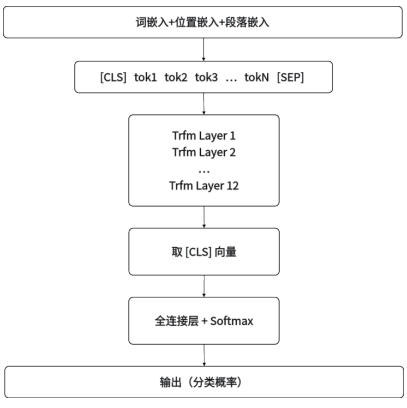


图 1 BERT 模型输入与分类微调整体架构

（二）电商工单分类训练与实时分类流程

本文提出的基于 BERT 预训练模型的电商工单分类系统分

 Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.

为离线模型训练和在线分类两大环节，各自流程如图 2 和图3所示。

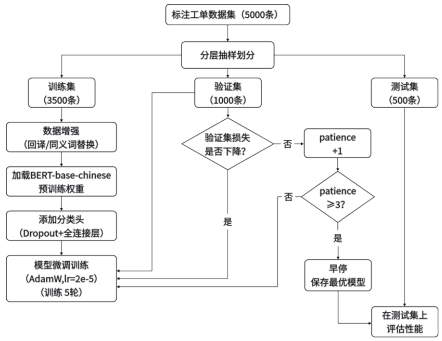


图 2 电商工单分类训练流程

在训练阶段主要步骤为下列6个步骤：

1. 数据预处理：清洗、去重、规整原始工单，划分三类数据集，并通过重采样缓解样本失衡；

2. 文本编码：统一文本长度截断填充，生成 BERT 所需 token、掩码、分段序列；

3. 模型构建：加载中文 BERT 权重，提取 [CLS] 特征连接两层全连接层，经 Softmax 输出分类概率；

4. 损失优化：采用加权交叉熵，降低多数类样本对损失的影响；

5. 迭代训练：配置 AdamW 优化器与超参数，依据验证集指标保存最优模型；

6. 模型固化：导出最优模型用于线上推理。

实时工单分类环节（如图3所示），主要步骤包括

实时分类流程（见图 3）

1. 工单接入：客服系统实时推送工单文本；

2. 在线预处理：过滤特殊字符，统一文本长度与编码；

3. 模型预测：输入模型得到各类别置信概率，取最大值作为初始分类标签；

4. 阈值分流：依靠动态置信阈值划分高低置信工单；

5. 业务分发：高置信工单自动分流处理，低置信工单转交人工复核，修正标签后回流数据集迭代模型。

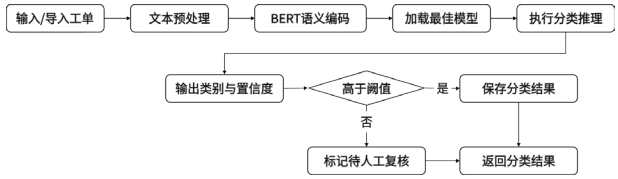


图 3 电商工单实时分类流程图

（三）工单类别不平衡问题双重优化策略

工单数据普遍存在类别失衡，模型训练易偏向多数类，小众售后、质量类工单召回效果差。本文结合数据集重采样与损失加权双重策略改善该问题。

1. 混合重采样

单一过 / 欠采样各有缺陷，故采用混合采样：对样本量大的大类随机欠采样留存关键样本；小众类通过同义词替换、句式改写做带噪数据增强过采样，均衡样本且不丢失有效信息。

2. 加权交叉熵损失

标准交叉熵各类权重均等，模型易忽视小众类别。本文按样本量反比赋予类别权重，小众类损失权重更高，错分时损失惩罚更大，引导模型充分学习小众工单特征。

（四）人机协同动态分类阈值设置

完全自动分类易混淆歧义工单，全人工分拣效率低下。本文基于预测置信度构建人机协同动态阈值机制：最大置信度高于阈值则自动分流处理；低于阈值则交由人工复核。阈值可随业务场景灵活调节，高峰下调提升自动化量，投诉高发期上调减少错分。人工修正后的工单回流训练集，形成预测、复核、数据回流、模型迭代的闭环，持续优化分类精度。

二、实验结果与分析

（一）工单数据集来源与数据清洗

本文实验数据取自国内中小型综合电商平台2025年1-6月真实工单，原始样本共6212条。所有数据已完成隐私脱敏，且获平台官方学术授权，可无偿用于本次研究。数据集涵盖物流、产品质量、退换货、退款、账号安全、优惠券活动、其他咨询七类工单，样本分布差异显著。为消除噪声干扰，本文对原始数据开展标准化清洗：剔除无效短文本、特殊符号与乱码，统一文本简体格式与数字、时间规范，并删除重复工单，最终剔除1212条低质量样本，得到5000条有效实验样本。

为规避随机划分造成的数据分布偏差，保证数据集贴合真实业务场景，本文采用分层抽样方式，按7:2:1比例划分数据集，最终得到3500条训练样本、1000条验证样本、500条测试样本。该划分方式可完整保留各类别样本占比特征，保障模型训练、调优及测试结果的真实性与可靠性。

（二）实验评价指标与对比模型

1. 评价指标

采用文本分类通用量化指标：精确率 P、召回率 R、综合 F1 值，同时统计人机协同模式下人工复核工单占比，兼顾模型精度与业务落地效率。

2. 对比基线模型

选取两类经典文本分类模型作为对照，验证本文 BERT 方案优势：

1）SVM：采用 TF-IDF 提取文本特征，线性核支持向量机分类；

2）TextCNN：基于随机初始化静态词向量，多层卷积提取局部语义特征分类；

同时设置两组消融实验，验证不平衡优化、人机协同模块的有效性：

消融实验 1：仅基础 BERT，无采样与损失加权处理；

消融实验 2：BERT + 不平衡优化 + 损失加权处理，但无人工协同阈值机制。

（三）基础模型分类效果对比

三个方案在测试集上的分类性能通过准确率、精确率、召回

Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.

率和 F1 值等指标进行评估，各自分类性能对比如图 4 所示。

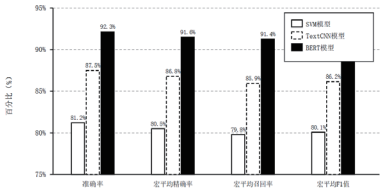


图 4 三种模型分类性能对比

从上图可以看出，基础 BERT 模型比 SVM 模型准确率提升 11.1 个百分点，比 TextCNN 模型准确率提升 4.8 个百分点，充分验证了 BERT 模型在电商工单分类任务上的优越性。

BERT 模型在各类别分类性能表现如图 5 所示：

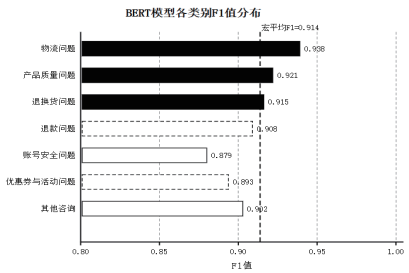


图 5 BERT 预训练模型各类别分类性能

（四）类别不平衡优化策略消融实验

为验证类别平衡采样策略的有效性，本文设置单一变量对照实验。实验组结合类别平衡采样与数据增强方法，通过重采样、回译、同义词替换均衡样本分布，并采用加权交叉熵损失函数，依据样本数量倒数占比分配类别权重，优化优惠券活动等小众样本识别效果。对照组直接使用原始分布数据集训练，未采用采样优化与数据增强操作。两组实验统一模型结构、预训练权重、超参数及验证、测试集，仅变更训练集采样方式。实验以准确率、宏平均精确率、宏平均召回率与宏平均 F1 值为评价指标，两组模型性能对比结果如图 6 所示。

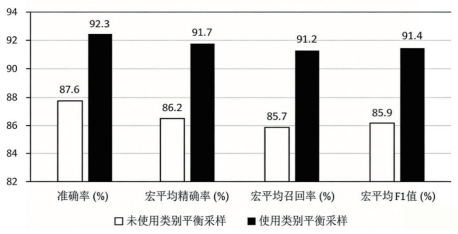


图 6 类别平衡采样策略对比实验结果

实验结果表明，未引入类别平衡采样时，模型分类准确率仅 87.6%，宏平均 F1 值为 85.9%。模型训练过度偏向主流样本类别，对少数类样本学习不足，整体泛化能力受限。引入类别平衡采样并结合数据增强后，模型性能显著提升，准确率、宏平均 F1 值分别提升 4.7 和 5.5 个百分点。其中账号安全、优惠券活动等小众类别优化效果最为突出，账号类 F1 值由 81.3% 提升至 87.9%。由此证实，该优化策略可有效改善数据不均衡导致的模型偏置问题，均衡学习各类工单语义特征，显著提升小众工单的识别精度。

（五）人机协同阈值策略实验效果

本文设置置信阈值 θ =0.85 开展线上推演，对比纯自动分类与人机协同两种模式的业务效果。全机器自动分类时，整体 F1 值为 91.7%，错分工单占比 8.3%，所有错误样本均需人工修正。采用人机协同分流策略后，置信度 ≥ 0.85 的工单由系统自动处理，占比 76.4%；仅 23.6% 的低置信模糊工单转入人工复核，经人工校正后整体 F1 值提升至 97.2%。实验表明，人机协同机制可兼顾效率与精度，既大幅减少人工工作量、降低人力成本，又能通过人工纠错消除错分样本，显著提升整体分类精度。同时，修正工单持续回流迭代，可保障模型长期性能稳步优化。

三、总结与未来工作

针对电商海量工单智能分类落地难题，本文融合 BERT 预训练、类别不平衡优化与人机协同阈值调控，构建了完整的工单分类方案。研究利用 BERT 挖掘文本深层语义，弥补传统模型语义提取不足的缺陷；通过混合重采样与加权交叉熵损失，解决小众工单识别效果差的问题；借助动态置信阈值实现人机协同分流，平衡分类效率与准确率。

真实数据集实验表明：BERT 模型性能显著优于 SVM、TextCNN，更适配口语化、歧义性工单文本；类别平衡优化策略有效提升了小众工单召回率；人机协同机制可实现超 76% 工单自动分拣，人工复核后整体分类精度达 97.2%，工程落地性强。该方案无需复杂特征工程，微调即可快速部署，适配各类型电商平台。

研究仍存在一定局限，后续将从三方面优化：融合图文多模态信息提升模糊工单识别能力；引入智能算法实现阈值自适应调整；构建多标签分类模型，适配多诉求的复杂售后工单场景。

参考文献

[1] 中国互联网络信息中心 . 第 57 次中国互联网络发展状况统计报告 [R]. 北京：中国互联网络信息中心 ,2026.

[2] 阿里巴巴集团 . 阿里智能客服系统白皮书 [R]. 杭州：阿里巴巴 ,2023.

[3] 文飞 . 传统与大模型并举：中文文本分类技术对比研究 [J]. 智能计算机与应用 ,2024, 14 (06): 88-94.

[4] 刘晓明 , 李丞正旭 , 吴少聪等 . 文本分类算法及其应用场景研究综述 [J]. 计算机学报 ,2024, 47 (6): 1321-1352.

[5] 王雪 , 李萌 . 基于深度学习的文本分类模型研究综述 [J]. 计算机工程与科学 ,2022, 44 (09): 1672-1681.

[6]Kim Y. Convolutional Neural Networks for Sentence Classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.

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

[8] 陈云 , 王浩 , 李佳 . 面向不平衡数据的 BERT 文本分类优化研究 [J]. 中文信息学报 ,2023, 37 (2): 56-64.

 Copyright © This Work is Licensed under A Commons Attribution-Non Commercial 4.0 International License.