面向社交网络跨语言虚假信息检测的LLM增强自监督域自适应方法

沈 航,王 旭,王天荆,戴远飞,白光伟
(南京工业大学计算机与信息工程学院(人工智能学院),江苏南京 211816)

摘 要: 在跨平台、跨语言的社交网络环境中,虚假信息的传播具有高隐蔽性和跨文化性,给舆情治理与社会信任体系带来了严峻挑战.由于不同语言和文化背景下文本的表达方式存在显著差异,传统基于深度学习的检测方法在跨域泛化与语义建模方面普遍存在性能退化问题,表现为跨域特征对齐不足、语义表示缺失以及对隐喻、情感和文化语境的理解能力受限.针对这些问题,本文提出一种大语言模型(Large Language Model,LLM)增强的自监督域自适应(Domain Adaptation,DA)检测框架,通过融合LLM的深层语义建模能力与对比学习(Contrastive Learning,CL)的判别特征学习机制,实现高鲁棒性与高泛化性的跨语言虚假信息检测.该方案构建一个从语义增强到特征对齐再到反馈优化的闭环体系.首先,通过基于Prompt的跨语言文本增强机制,引导LLM在生成数据时保持语义完整性与文化适配性,从而在保留原始语义核心的同时,生成符合目标语言风格的高质量文本样本,有效缓解跨语言场景中的语义鸿沟.随后,设计双维度对比策略,在词元层面对齐局部词汇特征,在语句层面对齐全局语义逻辑,从不同层面统一源域与目标域的数据表示,以提升特征分布一致性与跨语言检测的稳定性.最后,构建LLM辅助的跨语言联合训练机制,利用对比损失作为动态反馈信号,引导LLM在迭代微调过程中不断优化生成策略,促使增强样本的分布逐步靠近CL检测器的判别边界,从而实现跨语言数据增强与特征学习的协同演化.在中文社交平台数据集Weibo与英文突发事件数据集PHEME上的实验结果表明,所提方法在精确率和F1指标上显著优于商业LLM直接检测(如ChatGPT-4o)、主流深度学习模型(包括LSTM、TextCNN、RCNN、HAN)及LLM增强检测方法(如LACL).在跨语言检测中,所提方法的平均检测精度相比基准方法提升幅度超过10%.特征可视化分析进一步表明,所提方法能压缩类内特征差异、扩大类间判别间隔,从而获得更清晰的特征边界与更高的判别置信度.

关键词: 社交网络虚假信息;大语言模型;对比学习;跨语言文本增强;域自适应
基金项目: 国家自然科学基金(No.61502230,No.61501224,No.62202221);江苏省自然科学基金(No.BK20201357)
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-15
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250470

LLM-Enhanced Self-Supervised Domain Adaptation for Cross-Lingual Misinformation Detection in Social Networks

SHEN Hang, WANG Xu, WANG Tian-jing, DAI Yuan-fei, BAI Guang-wei
(*College of Computer and Information Engineering* (*College of Artificial Intelligence*), *Nanjing Tech University*, *Nanjing*, *Jiangsu* 211816, *China*)

Abstract: In cross-platform and cross-lingual social network environments, the spread of misinformation is characterized by high concealment and cross-cultural complexity, posing serious challenges to public opinion governance and social trust systems. Due to significant differences in linguistic and cultural expression, traditional deep learning-based detection methods often suffer from performance degradation in cross-domain generalization and semantic modeling, exhibiting insufficient cross-domain feature alignment, incomplete semantic representation, and limited understanding of metaphors, emotions, and cultural contexts. To address these limitations, this paper proposes a large language model (LLM)-enhanced self-supervised domain adaptation (DA) detection framework. By integrating the deep semantic modeling capacity of LLMs with the discriminative feature learning capability of contrastive learning (CL), the framework achieves robust and generalizable cross-lingual misinformation detection. This solution establishes a closed-loop system encompassing semantic augmentation, feature alignment, and feedback optimization. First, a prompt-based cross-lingual text augmentation mechanism

收稿日期:2025-06-01;录用日期:2025-10-09;责任编辑:覃怀银

is designed to guide the LLM in maintaining semantic integrity and cultural adaptability during data generation. This enables the production of high-quality samples that preserve the semantic core of the original text while conforming to the linguistic style of the target language, effectively mitigating semantic gaps in cross-lingual contexts. Next, a dual-dimensional contrastive strategy aligns local lexical features at the token level and global semantic logic at the sentence level, unifying source and target domain representations at multiple levels to enhance feature distribution consistency and cross-lingual detection stability. Finally, an LLM-assisted cross-lingual training mechanism is introduced, where contrastive loss serves as a dynamic feedback signal to guide the iterative fine-tuning of the LLM. This process progressively refines the augmentation strategy, ensuring that the generated data distribution converges toward the CL detector's decision boundary and enabling the co-evolution of cross-lingual data augmentation and feature learning. Experimental results on heterogeneous social media datasets, Weibo (a Chinese social platform) and PHEME (an English dataset of event-related rumor propagation), demonstrate that the proposed method significantly outperforms commercial LLM direct detection (e.g., ChatGPT-4o), mainstream deep learning models (e.g., LSTM, TextCNN, RCNN, HAN), and existing LLM-enhanced methods (e.g., LACL) in terms of accuracy and F1 score. In cross-lingual detection, the average detection accuracy of the proposed approach exceeds baseline methods by more than 10%. Further feature visualization analysis confirms that our method compresses intra-class variance and enlarges inter-class separability, resulting in clearer decision boundaries and higher classification confidence.

Key words: social network misinformation detection; large language model (LLM); contrastive learning (CL); cross-language text augmentation; domain adaptation (DA)

Foundation Item(s): National Natural Science Foundation of China (No.61502230, No.61501224, No.62202221); Natural Science Foundation of Jiangsu Province (No.BK20201357)

1 引言

在互联网时代,社交媒体以其即时性与开放性提升了信息传播效率,但也带来了"信息双刃剑效应",即虚假信息的滋生与扩散愈发严重[1]. 尤其是在网络亚文化环境中,隐喻表达、符号变体等非典型语言形式频繁出现,导致传统算法在语义理解与符号解析上面临困境. 尽管现有社交平台已构建以人工审核与算法过滤相结合的多层治理体系,但面对媒体传播方式的快速演变,该体系在时效性与准确性上均存在局限,亟需设计高适应性的虚假信息检测机制.

在多语言社交网络环境中,语种结构差异、文化认知偏差与时空语境隔阂共同加剧了虚假信息检测难度. 例如,中文社交平台用户更倾向于通过网络流行语、缩略词和表情构建语义场;英文社交平台用户虽然语言逻辑相对规范,但也存在大量非正式表达,如隐喻、俚语、缩写与表情符号. 根据域自适应(Domain Adaptation,DA)[2]理论,当以中文、英文社交平台为源域(source domain)、目标域(target domain)时,两者在语法结构、语用习惯和信息组织等层面的系统性差异会导致特征分布偏移,并造成语义编码错位[3]. 在这种情况下,基于单语种训练的检测模型在跨语言迁移中泛化性能必然受限. 传统神经网络在训练中通常将语言结构固化于参数矩阵,形成语言绑定效应,降低了模型跨语言迁移的适应能力.

自监督学习框架下的对比学习(Contrastive Learning,CL)已成为跨语言建模的重要范式. 通过构建正负样本对和表示学习,CL在计算机视觉中的图像分类[4]、目标检测[5],以及自然语言处理(Natural Language Processing,NLP)中的文本分类[6]、机器翻译[7]等任务上均有优异表现. CL结合图神经网络(Graph Neural Network,GNN)可将社交网络中用户关系、语义线索与传播路径建模为图结构,并利用文本全局-局部的上下文语义关系,为虚假信息检测提供鲁棒的特征表达能力[8~10].

大语言模型(Large Language Model,LLM)[11]依托超大规模语料预训练与强大的语义建模能力,在语言语义理解方面展现出巨大潜力. LLM能够捕捉长距离依赖关系和隐含逻辑,统一表示不同语言中的深层语义结构,尤其在处理隐喻、符号变体和文化差异等复杂表达时展现出显著优势[12,13]. 因此,融合LLM的语义表征能力与CL的判别学习机制,有潜力解决跨语言检测中的语义失配与迁移退化问题.

尽管LLM和CL已在DA任务中展现出潜力,但在跨语言场景中的应用仍面临一些挑战:

(1)跨语言风格迁移与语义完整性难以兼顾. 若源域与目标域在语言风格和语义表达上存在显著差异,文本增强策略往往陷入两难困境:过度强调风格迁移会带来语义扭曲,严格保持语义一致又难以有效适配风格. 对于该问题,DADP[14]将依存句法分析(Dependency Parsing,DP)作为辅助任务,与命名实体识别(Named Entity Recognition,NER)联合训练,通过最大均值差异(Maximum Mean Discrepancy,MMD)实现DP特征在源域与目标域的无监督对齐. Zhang等人[15]则将NER分解为实体检测与类型预测两个子任务,并通过

跨域迁移提升模型适应性. Wang等人[16]在特征空间中融合相邻领域语义,并结合生成式增强策略探索未知域特征,从而增强目标域的一致性保持能力.

(2)浅层与深层特征难以动态协调. 浅层特征(如词汇、句法结构)对域差异较为敏感,易导致模型误判;而深层特征(如语义逻辑)虽然鲁棒性强,但关键细节易被忽视. 主要原因在于跨语言任务同时依赖局部词汇/句法差异和全局语义一致性,而单一粒度特征建模方法难以兼顾二者. 研究者尝试借助图结构建模与层次化编码来缓解该问题. GraphCTA[17]引入图结构变换函数,并结合自训练与邻域对比提升节点表征和DA性能. MHCCL[18]则通过层次聚类构建多粒度正负对比样本,旨在提升时间序列语义表征质量. PDA[19]引入提示学习(Prompt Learning)和双分支结构,以便对齐目标域的分布. 虽然该方法改善了语义对齐,但在细粒度判别方面带来了信息损失风险.

针对上述限制和挑战,本文提出LLM增强的自监督DA方法,旨在实现多语言、多平台社交网络中的高鲁棒性与高泛化性虚假信息检测. 首先,设计基于Prompt的跨语言文本增强机制,以生成兼具语义完整性与跨语言风格适配的样本. 其次,引入LLM增强的双维度对比策略,在词元层面与语句层面对齐局部词汇特征与全局语义逻辑,提升模型的跨语言检测鲁棒性. 最后,构建对比损失驱动的跨语言联合训练框架,其中,CL训练反馈用于引导LLM的微调,以持续优化其文本增强策略. 在Weibo[20]与PHEME[21]两个语境差异显著社交媒体数据集上的实验结果表明,所提方法在精确率、F1分数等指标优于通用商业LLM、主流深度学习方法及LLM辅助的学习方法.

2　方案设计

图1展示了所提LLM辅助的双维度跨语言检测框架,由跨语言Prompt编排、CL检测网络和LLM动态微调三个主要功能模块构成,形成一个从语义增强到特征对齐、再到反馈优化的可持续性闭环优化框架,主要包含:

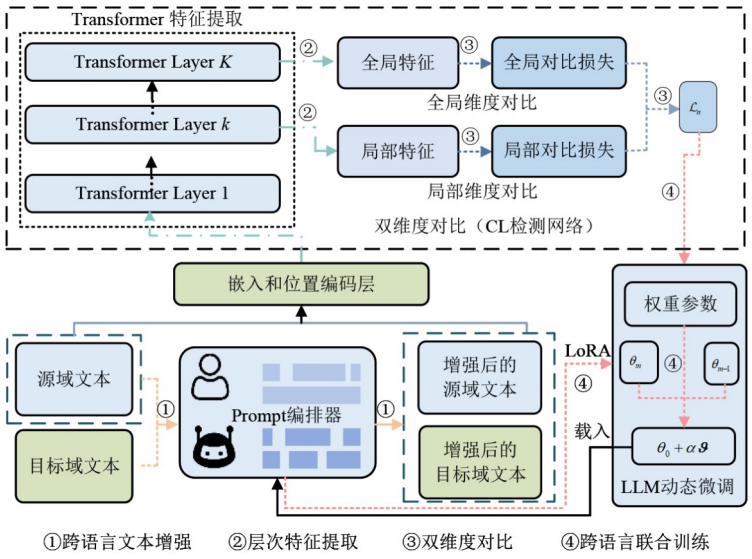


图1　LLM辅助的双维度跨域检测框架

(1)基于Prompt的跨语言文本增强. 通过设计跨语言Prompt约束LLM的生成行为,实现文本风格迁移并保持语义一致.

(2)LLM辅助的双维度对比. 在训练阶段,增强数据与原始数据共同输入CL检测网络. 网络采用词元级(局部)和句子级(全局)的双维度对比机制,将特征映射到共享对比空间,通过计算对比损失来对齐源域与目标域的语义表示,显式提升跨语言特征一致性.

(3)LLM动态微调. LLM通过策略性微调快速适应跨语言风格迁移任务,为CL检测网络持续提供高质量、分布更贴近目标域的训练样本.

(4)LLM辅助跨语言联合训练. CL网络输出的对比损失用于反向调整LLM的生成策略,使其在生成过程中逐步对齐CL网络的判别边界.

该框架中,LLM并非被动的数据生成器,而是演化为与CL网络协同优化、共同学习的组件.

为了便于跟踪,文中主要符号归纳至表1.

2.1　基于Prompt的跨语言文本增强

Prompt被用于引导和约束LLM完成跨语言文本增强,需要规范格式、提升多样性、保持语义一致性以及适配语言与文化语境. Prompt编排用来引导LLM识别输入文本的语种特征,并激活相应的处理路径. 对于中文输入,执行中译英,并在语境中融入西方社交媒体的

表1 主要符号和变量含义

符号	含义
$h_S^{(i,j,k)}$	第k层Transformer中第j个词的隐藏状态
$\boldsymbol{h}_S^{(i,k)}$	源域样本i在第k层的局部特征表示
$\mathcal{L}_n^{(k,b)}$	第n个epoch中第b个批次第k层的损失
$\mathcal{L}_n^{(k)}$	第n个epoch中第k层的局部损失
$\mathcal{L}_n$	第n轮总对比损失
M	最大微调轮次
N	总训练轮数(epochs)
$x_S^{(i)}/x_T^{(i)}$	源域/目标域第i条原始样本
$x_{S\to T}^{(i)}$	源域样本增强为目标域风格
$x_{T\to T'}^{(i)}$	目标域样本i的伪增强样本
$\mathcal{S}/\mathcal{T}$	源域/目标域数据集
T	文本增强间隔(每T轮增强一次)
$\boldsymbol{z}_{S,i}^{(k)}$	源域样本i的局部投影向量
θ_0	LLM的初始参数
$\theta_m/\boldsymbol{\vartheta}_m$	第m次微调后的LLM参数/偏移向量
λ_m	第m次微调的权重
φ_n	第n轮LLM增强效果评估指标

表达方式;对于英文输入,执行英译中,并映射至本土化的互联网表达方式,贴合集体主义等文化框架. 这种语言转换不仅是表层的词汇替换,更是深层文化框架的迁移. 通过语种识别和文化语境适配,Prompt编排使LLM生成既忠实于原始语义又符合目标语境的高质量增强样本,提升跨语言虚假信息检测的鲁棒性.

本节编排如图2所示的Prompt,相应的中英双向增强样例对见图3. 针对原始和增强数据,Prompt设计需要达成如下目标:

(1)语义一致性与风格增强. 增强文本在保留原始信息核心事件与立场的基础上,采用更具表现力的语言进行重写. 例如,“好主妇、好母亲”被转换为“homemaker & mother”,规避了直译为“housewife”可能带来的负面语义,更契合西方性别语境. 同时,通过添加表情符号和话题标签(如 #GenderEquality)增强社交媒体风格.

(2)跨语言与文化适配. 借助LLM的语义重构能力转换语言与迁移文化,如“抢地盘”被转化为“fight for resources”,避免物理空间意象的局限;“直男癌”被翻译为“toxic masculinity”,引入了女权核心术语;“共创社会繁荣”表达为“ELEVATE each other”,契合西方强调个体成长的文化逻辑,并使用#PatriarchyAlert等标签贴合西方议题表达.

```
You are an expert in cross-language social media rumor data augmentation. Please process the input text according to the following rules:
1. **Holistic Enhancement**:
   -Determine whether the input text is a rumor (0 for non-rumor, 1 for rumor), and augment the data based on your judgment. During augmentation, please focus on the overall semantic meaning and strictly maintain semantic consistency.
2. **Cross-lingual Conversion**:
   -Identify the input language type and perform language conversion: If the input is Chinese, please output an English version; if the input is English, please output a Chinese version. Perform cultural adaptation: When converting from Chinese to English, incorporate Western cultural perspectives and English social media expression habits; when converting from English to Chinese, incorporate Chinese cultural perspectives and Chinese internet expression habits.
3. **Increase Diversity**:
  - Focus on making the data more diverse in terms of language expression, without altering the underlying meaning. The enhanced data should be more varied but semantically identical to the original.
4. **Format Consistency**:
   -Ensure the output format is consistent with the input, including text structure and expression style.
Please return only the converted text without explaining your judgment process.
```

图2 Prompt编排

(3)表达多样性增强. 在不改变基本语义的前提下,增强样本引入网络俚语(如“Sis”“ain't”)、缩写(如“smh”“bc”)等地道用法,增强文本的语言多样性与真实感;同时引入表情符号和话题标签使内容更具社交平台风格.

(4)格式与结构一致性. 增强样本严格保留原始结构字段(如uid、string_value、replies),确保可直接用于下游模型训练;各回复条目与原始样本对应,且保留原有情感色彩,如批评、讽刺等语气通过强化语气词与表情进行保真迁移.

增强后的文本不仅完成了中英互译,还在保持语义一致性的前提下,增加了表达多样性,并使内容更加符合目标语言的社交网络文化环境.

2.2 LLM辅助的双维度对比

在跨语言场景中,源域与目标域的特征分布差异主要体现在语言体系差异(如语法结构的系统性区别与词汇使用偏好)、文化认知鸿沟(如隐喻表达、社会规约的解读差异)以及时空语境隔阂(如历时语义演变与地域性变体). 这些差异导致跨语言特征空间出现显著偏移,表现为同类样本的分布边界错位和异类样本的嵌入重叠,从而加剧跨语言检测的困难.

为缓解这种偏移,特征空间重构过程引入了双维度对比机制. 在词元层面,通过细粒度语义对比消除词汇的跨语言歧义;在语句层面,通过整体语义对比捕捉语言结构的深层关联. 这种分层对比策略能够同步缩小词汇含义的局部差异与语句语义的全局偏差. 当分

图3　跨语言文本增强样例

层对比达到理想效果时,源域特征会被逐步校准至目标域特征空间,从而达成两个目标:(1)同类样本(无论来自源域还是目标域)在嵌入空间中形成紧凑聚类;(2)异类样本之间具有清晰可辨的决策边界.

图4展示了兼顾词元和句子级的双维度对比网络的构造. 源域表示为 $\mathcal{S}=\{w_S^{(1)},w_S^{(2)},\cdots,w_S^{(I)}\}$,其中 $x_S^{(i)}=\{w_S^{(i,1)},w_S^{(i,2)},\cdots,w_S^{(i,C)}\}$ 代表源域 $\mathcal{S}$ 中第 $i\in\{1,2,\cdots,I\}$ 条源域样本的词序列,C 为相应的序列长度. 目标域 $\mathcal{T}=\{x_T^{(1)},x_T^{(2)},\cdots,x_T^{(I)}\}$ 同样包含 I 条社交网络文本实例. 基于2.1节设计的Prompt,源域样本 $x_S^{(i)}$ 被增强为具有目标域风格的样本 $x_{S\to T}^{(i)}$,而目标域样本 $x_T^{(i)}$ 则被增强为风格保持的伪样本 $x_{T\to T}^{(i)}$. 此后,原始样本 $x_S^{(i)}$、增强样本 $x_{S\to T}^{(i)}$ 和 $x_{T\to T}^{(i)}$ 一并输入到CL网络进行统一建模. 在输入前,离散语言序列需映射至连续语义空间. $x_S^{(i)}$ 中的第 j 个词单元 $w_S^{(i,j)}$ 通过如下嵌入函数

$$h_S^{(i,j,0)}=\mathcal{E}(w_S^{(i,j)})\in\mathbb{R}^d \tag{1}$$

得到带位置信息的 d 维词向量. 完成嵌入后,样本的初始特征矩阵表示为

$$\boldsymbol{h}_S^{(i,0)}=\left[h_S^{(i,1,0)},h_S^{(i,2,0)},\cdots,h_S^{(i,C,0)}\right] \tag{2}$$

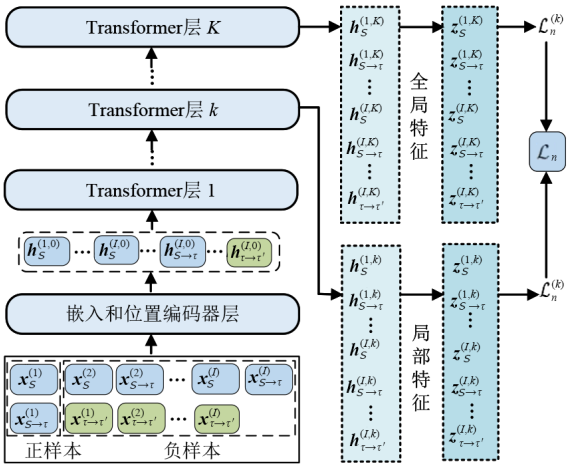


图4　双维度对比架构

同样的嵌入过程也应用于 $x_{S\to T}^{(i)}$ 和 $x_{T\to T}^{(i)}$.

在得到初始特征表示 $\boldsymbol{h}_S^{(i,0)}$ 后,CL网络通过 K 层堆叠的Transformer编码器[22]建模语义. 每层编码器采用多头自注意力机制,逐层捕获词级交互关系. 设 $\{h_S^{(i,c,k-1)}\}_{c=1}^{C}$ 为 k-1层的词级特征集合. 在第 k<K 层,所有的位置特征 $\{h_S^{(i,c,k-1)}\}_{c=1}^{C}$ 作为输入,经过Transformer

第k层编码器建模后，位置j的特征被更新为

$$h_S^{(i,j,k)}=\mathcal{F}^{(k)}(\{h_S^{(i,c,k-1)}\}_{c=1}^{C})_j \tag{3}$$

为实现不同粒度的语义信息建模，双维度对比框架分别从浅层与深层编码器中提取局部与全局特征. 对于第k层输出，词级特征序列表示为

$$\boldsymbol{h}_S^{(i,k)}=\left[h_S^{(i,1,k)},h_S^{(i,2,k)},\cdots,h_S^{(i,C,k)}\right]\in\mathbb{R}^d \tag{4}$$

该特征聚焦于句子的局部语义结构（如短语、依存关系等），通过浅层编码器捕捉短距离语义约束. 从最终层（第K层）提取的上下文感知特征则用于建模全局语义关联，即$\boldsymbol{h}_S^{(i,K)}\in\mathbb{R}^d$. 该特征通过第$K$层编码器整合全句上下文信息，捕捉长距离语义关联（如因果推理等）. 增强后的样本$x_{S\to T}^{(i)}$和$x_{T\to T'}^{(i)}$也通过相同流程提取特征，以确保双向跨语言风格迁移的信息得到一致性处理.

双维度对比方法旨在协同优化局部语义与全局语义结构表示，包含双维度特征投影、分层对比和损失加权融合三部分：

双维度特征投影：令$\boldsymbol{W}_1$和$\boldsymbol{W}_2$代表MLP的第一层和二层权重矩阵. 局部特征$\boldsymbol{h}_S^{(i,k)}$和全局特征$\boldsymbol{h}_S^{(i,K)}$被$\boldsymbol{W}_1$经MLP投影到一个统一的对比空间后，得到$\boldsymbol{z}_S^{(i,k)}$和$\boldsymbol{z}_S^{(i,K)}$，表示为

$$\begin{cases}\boldsymbol{z}_S^{(i,k)}=\boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1\boldsymbol{h}_S^{(i,k)}\right)\\ \boldsymbol{z}_S^{(i,K)}=\boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1\boldsymbol{h}_S^{(i,K)}\right)\end{cases} \tag{5}$$

其中，$\sigma(\cdot)$为非线性激活函数. 在投影空间中，用来衡量$\boldsymbol{z}_S^{(i,k)}$和$\boldsymbol{z}_S^{(i',k)}$相似度的函数定义为

$$\mathrm{sim}(\boldsymbol{z}_S^{(i,k)},\boldsymbol{z}_S^{(i',k)})\triangleq\frac{\boldsymbol{z}_S^{(i,k)}\cdot\boldsymbol{z}_S^{(i',k)}}{\left\|\boldsymbol{z}_S^{(i,k)}\right\|\times\left\|\boldsymbol{z}_S^{(i',k)}\right\|} \tag{6}$$

同理可计算全局特征相似度$\mathrm{sim}(\boldsymbol{z}_S^{(i,K)},\boldsymbol{z}_S^{(i',K)})$.

分层对比：对于局部特征或全局特征，正样本与锚点的相似性定义为

$$\phi^{(i,k)}\triangleq\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_S^{(i,k)},\boldsymbol{z}_{S\to T}^{(i,k)}\right)}{\tau}\right) \tag{7}$$

其中，τ是温度超参数. 源域负样本和锚点的相似函数（包含增强数据）为

$$\phi_S^{(i,i',k)}\triangleq\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_S^{(i,k)},\boldsymbol{z}_S^{(i',k)}\right)}{\tau}\right) \tag{8}$$

目标域负样本和锚点的相似函数为

$$\phi_T^{(i,i',k)}\triangleq\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_S^{(i,k)},\boldsymbol{z}_{T\to T'}^{(i',k)}\right)}{\tau}\right) \tag{9}$$

相似性归一化因子表示为

$$\Theta=\frac{\phi^{(i,k)}}{\phi^{(i,k)}+\sum\limits_{\substack{i'\in\mathcal{S}^{(b)}\\ i'\neq i}}\phi_S^{(i,i',k)}+\sum\limits_{\substack{i'\in\mathcal{T}^{(b)}\\ i'\neq i}}\phi_T^{(i,i',k)}} \tag{10}$$

每个epoch参与CL检测网络训练的样本数量为$3I$，包括I个原始样本及两种类型的（$2I$个）增强样本. 设每批次（batch）数据量为a，则第n个epoch包含$B=\lceil 3I/a\rceil$个批次的数据. 批次b的样本集合表示为$\mathcal{S}^{(b)}\cup\mathcal{T}^{(b)}$，其中$\mathcal{S}^{(b)}$为源域及其增强文本数据集合，$\mathcal{T}^{(b)}$为目标域增强样本集合. 第$n$个epoch中第$b$个批次下第$k$层的局部损失被计算为

$$\mathcal{L}_n^{(k,b)}=-\frac{\log\Theta}{|\mathcal{S}^{(b)}\cup\mathcal{T}^{(b)}|} \tag{11}$$

通过最小化式（11），模型在投影空间中实现类内紧凑与类间分离，从而增强特征的判别性. 第n个epoch的局部损失被表示为

$$\mathcal{L}_n^{(k)}=\frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_n^{(k,b)} \tag{12}$$

同理可得全局损失$\mathcal{L}_n^{(K)}$.

损失加权融合：为平衡局部和全局语义的贡献，在第n个epoch结束后的总损失计算为

$$\mathcal{L}_n=\delta\mathcal{L}_n^{(K)}+(1-\delta)\mathcal{L}_n^{(k)} \tag{13}$$

其中，$\delta\in[0,1]$为学习权重因子.

2.3 LLM动态微调

主流的LLM微调方法包括全参数微调（Full Fine-Tuning）[23]、提示学习（Prompt Learning）[24]和参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）[25]. 低秩适应（Low-Rank Adaptation，LoRA）[26]属于PEFT中的代表性方法. 通过在原始权重矩阵旁引入低秩分解矩阵，该方法在冻结大部分参数的同时实现高效更新，从而显著降低显存占用与计算开销. 由于公共数据集规模有限，直接全量微调难以保证增强样本的多样性和语义一致性. 本研究采用LoRA微调策略，使LLM能在每轮迭代后快速适应跨语言风格迁移任务，从而为CL检测网络提供语义一致性和多样性的高质量训练样本.

设原始LLM的参数为θ_0，第m次微调后的参数记为θ_m. 在第m次微调前，θ_{m-1}被用于跨语言文本增强，以生成更贴近目标域分布的样本，缩小源域与目标域在特征空间中的差距. 整个训练过程包含N个epoch，在每个第n个epoch结束时周期性地触发一次LoRA微调. 令$\mathcal{D}_{n,m}$为第n个epoch结束时，第m次微调所使用的训练数据集. 相应的LLM参数表示为

$$\theta_m=\mathrm{LoRA}\left(\mathcal{D}_{n,m},\theta_0\right) \tag{14}$$

上式中，LoRA$(\cdot)$代表低秩适应过程，即在冻结原始权重θ_0的同时，通过引入低秩分解矩阵高效更新参数. 这使得LLM能够在有限的计算资源下快速适应跨语言文本增强任务，同时保持原有模型的语义建模能力.

为便于描述LLM的动态微调过程，相对于初始参数θ_0的偏移向量表示为$\boldsymbol{\vartheta}_{m-1}=\theta_{m-1}-\theta_0$和$\boldsymbol{\vartheta}_m=\theta_m-\theta_0$，

分别表示在$\dim(\theta_0)$维空间($\mathbb{R}^{\dim(\theta_0)}$)中的参数的更新方向与幅度. 令$\gamma_{m-1}=\mathrm{sgn}(\boldsymbol{\vartheta}_{m-1})\in\{-1,0,1\}^{\dim(\theta_0)}$为参数的更新方向(正向、零更新或反向),$\gamma_{m-1}=\mathrm{sgn}(\boldsymbol{\vartheta}_{m-1})\in\mathbb{R}^{\dim(\theta_0)}$表示每个参数的更新幅度. 符号函数满足$\mathrm{sgn}(\boldsymbol{\vartheta}_{m-1})*|\boldsymbol{\vartheta}_{m-1}|=\boldsymbol{\vartheta}_{m-1}$,即通过方向向量与幅度向量的组合,完整地重建参数的偏移向量. 为刻画微调的更新模式,偏移向量$\boldsymbol{\vartheta}_{m-1}$被分解为符号向量$\gamma_{m-1}$和幅度向量$\eta_{m-1}$,满足$\boldsymbol{\vartheta}_{m-1}=\gamma_{m-1}\odot\eta_{m-1}$.

算法1描述了基于LoRA的LLM权重合并策略. 该策略融合了多轮微调结果,以提升跨语言泛化能力. 从任务角度,整个工作流被分为如下三步:

(1)冗余修剪. 修剪第$m-1$次微调得到的参数偏移向量$\boldsymbol{\vartheta}_{m-1}$的冗余,得稀疏向量$\hat{\boldsymbol{\vartheta}}_{m-1}$. 在$\boldsymbol{\vartheta}_{m-1}$中,幅度(magnitude)排序前$q$%的重要参数被保留,其余置零. 随后,$\hat{\boldsymbol{\vartheta}}_{m-1}$被分解为符号与幅度向量,满足$\hat{\boldsymbol{\vartheta}}_{m-1}=\hat{\gamma}_{m-1}\odot\hat{\eta}_{m-1}$,其中$\hat{\gamma}_{m-1}=\mathrm{sgn}(\hat{\boldsymbol{\vartheta}}_{m-1})$、$\hat{\eta}_{m-1}=|\hat{\boldsymbol{\vartheta}}_{m-1}|$. 同样的方式被用于处理第$m$次微调后的参数$\theta_m$,得到$\hat{\boldsymbol{\vartheta}}_m$、$\hat{\gamma}_m$和$\hat{\eta}_m$.

(2)符号选举. 在合并第$m-1$次和第m次微调的参数θ_{m-1}与θ_m之前,需要解决同一维度参数上的符号冲突,从而在不同微调结果之间保持语义一致性. 对于参数偏移向量$\hat{\boldsymbol{\vartheta}}_{m-1}$、$\hat{\boldsymbol{\vartheta}}_m$及对应的符号向量$\hat{\gamma}_{m-1}$、$\hat{\gamma}_m$,在第$e$维的元素分别记为$\hat{\vartheta}^{(e)}_{m-1}$、$\hat{\vartheta}^{(e)}_m$、$\hat{\gamma}^{(e)}_{m-1}$、$\hat{\gamma}^{(e)}_m$. 令合并后的目标偏移向量为$\boldsymbol{\vartheta}$,相应的符号向量为$\gamma$,它的第$e$维的符号被计算为(见第8行)

$$\gamma^{(e)}=\mathrm{sgn}(\hat{\vartheta}^{(e)}_{m-1}+\hat{\vartheta}^{(e)}_m)\tag{15}$$

当$\hat{\vartheta}^{(e)}_{m-1}$与$\hat{\vartheta}^{(e)}_m$符号一致时,$\gamma^{(e)}$继承该共同符号;当符号冲突时,$\gamma^{(e)}$根据两者幅度之和决定最终符号,幅度大的方向将主导合并后的更新方向;当两者的幅度近似相等但符号相反时,$\gamma^{(e)}=0$表示在该维度上未更新参数. 这种基于幅度感知的符号选举策略能保持跨语言语义一致性并且避免符号冲突导致的梯度抵消,提高权重合并的稳定性.

(3)参数合并. 完成修剪和选举后,针对$\boldsymbol{\vartheta}$中第e个参数,基于符号一致性原则,执行加权合并. 设第$r\in\{m-1,m\}$轮微调后的参数偏移为$\hat{\vartheta}^{(e)}_r$,我们仅保留符号与$\gamma^{(e)}$一致的参数,以避免梯度方向的冲突. 令$\mathcal{R}^{(e)}=\{\hat{\gamma}^{(e)}_r=\gamma^{(e)}\}_{r\in\{m-1,m\}}$为满足条件的参数集合. 基于此,第$e$个参数被加权合并为

$$\vartheta^{(e)}=\frac{1}{|\mathcal{R}^{(e)}|}\sum_{r\in R^{(e)}}\lambda_r\hat{\vartheta}^{(e)}_r\tag{16}$$

其中,λ_r为加权系数. 聚合后的参数偏移向量为

$$\boldsymbol{\vartheta}=\left[\hat{\vartheta}^{(1)},\hat{\vartheta}^{(2)},\cdots,\hat{\vartheta}^{(\dim(\theta_0))}\right]^{\top}\tag{17}$$

加权合并所有维度的参数后,最终返回的LLM参数为$\theta=\theta_0+\alpha\boldsymbol{\vartheta}$(见第21行),其中,$\alpha$为合并超参数,用来控制整体更新幅度.

算法1　Fuse($\theta_m,\theta_{m-1},\theta_0,\lambda_m,\lambda_{m-1},\alpha,q$)

输入: θ_m, θ_{m-1}, θ_0, λ_m, λ_{m-1}, α, q

输出: $\theta_0+\alpha\cdot\boldsymbol{\vartheta}$

1. For $r\in\{m-1,m\}$ do
2. 　$\boldsymbol{\vartheta}_r\leftarrow\theta_r-\theta_0$;
3. 　$\hat{\boldsymbol{\vartheta}}_r\leftarrow\mathrm{TopMag}(\boldsymbol{\vartheta}_r,q)$;
4. 　$\hat{\gamma}_r\leftarrow\mathrm{sgn}(\hat{\boldsymbol{\vartheta}}_r)$;
5. 　$\hat{\eta}_r\leftarrow|\hat{\boldsymbol{\vartheta}}_r|$;
6. End For
7. For $e=1$ to $\dim(\theta_0)$ do
8. 　$\gamma^{(e)}\leftarrow\mathrm{sgn}(\hat{\vartheta}^{(e)}_{m-1}+\hat{\vartheta}^{(e)}_m)$;
9. 　$\mathcal{R}^{(e)}\leftarrow\{\hat{\gamma}^{(e)}_r=\gamma^{(e)}\}_{r\in\{m-1,m\}}$;
10. 　If $\gamma^{(e)}=0$
11. 　　$\vartheta^{(e)}\leftarrow 0$;
12. 　　Continue;
13. 　End If
14. 　If $\mathcal{R}^{(e)}=\varnothing$
15. 　　$\vartheta^{(e)}\leftarrow 0$;
16. 　Else
17. 　　$\vartheta^{(e)}\leftarrow\sum_{r\in\mathcal{R}^{(e)}}\lambda_r\hat{\vartheta}^{(e)}_r\Big/\sum_{j\in R^{(e)}}\lambda_j$;
18. 　End If
19. End For
20. $\boldsymbol{\vartheta}\leftarrow[\vartheta^{(1)},\vartheta^{(2)},\cdots,\vartheta^{(\dim(\theta_0))}]^{\top}$;
21. Return $\theta\leftarrow\theta_0+\alpha\cdot\boldsymbol{\vartheta}$

LLM-CL对齐训练实现细节被归纳为递归算法2,其中的输入S_1和S_2是中间变量,用于归一化损失平滑和动态权重. 在联合训练中,每T个epoch执行一次文本增强操作,即在总计N个epoch的训练周期内将实现$w=\lceil N/T\rceil$次文本增强. 该策略利用LLM的跨语言语义生成能力,提升在目标域上的样本质量,并通过约束机制减少噪声,从而保证生成的样本对齐CL模型的检测目标. 在跨语言对齐中,$\mathcal{L}_n$不仅体现源域与目标域在共享特征空间中的分布差异,还间接反映LLM生成的增强样本质量. $\mathcal{L}_n$较小意味着LLM生成的样本有更丰富的语义信息和更强的区分度,有助于提升CL检测网络在源域与目标域上的判别能力,反之表明增强数据未能充分贴近目标域分布.

基于历史损失,这里采用动量更新机制[27]以避免单次异常训练周期过度影响LLM的调整方向. 第n次训练的加权损失为

$$\varphi_n=\mu\mathcal{L}_n+(1-\mu)\frac{1}{n-1}\sum_{n'=1}^{n-1}\mathcal{L}_{n'}\tag{18}$$

其中,μ为平衡因子,用于控制当前损失与历史均值的

算法2 Alignment(θ, n, m, S_1, S_2, λ)

输入: θ, n, m, S_1, S_2, λ

输出: θ

1. If $m > M$
2. Return θ;
3. End If
4. $(\mathcal{L}_n, \mathcal{D}_n) \leftarrow$ TrainCL(θ);
5. $S_1 \leftarrow S_1 + \mathcal{L}_n$;
6. If (n mod $T = 0$) and ($m \leqslant M$)
7. If $n = 1$
8. $\varphi_n \leftarrow \mathcal{L}_n$;
9. Else
10. $\varphi_n \leftarrow \mu \cdot \mathcal{L}_n + (1-\mu) \cdot \dfrac{S_1 - \mathcal{L}_n}{n-1}$;
11. End If
12. $S_2 \leftarrow S_2 + \varphi_n$;
13. $\lambda_m \leftarrow \dfrac{\varphi_n}{S_2}$;
14. $\mathcal{D}_{n,m} \leftarrow$ Augument(θ, $\mathcal{D}_n$); //增强文本
15. $\theta_m \leftarrow$ LoRA($\mathcal{D}_{n,m}$, θ_0);
16. $\theta \leftarrow$ Fuse(θ, θ_m, θ_0, λ, λ_m, α, q);
17. $\lambda \leftarrow \lambda_m$;
18. $m \leftarrow m + 1$;
19. End If
20. Return Alignment(θ, $n+1$, m, S_1, S_2, λ)

权重. 第m次微调的权重被归一化为

$$\lambda_m = \frac{\varphi_m}{\sum_{m'=1}^{m} \varphi_{m'}} \tag{19}$$

确保LLM在跨语言文本增强策略中自适应调整,优先保留历史上对齐效果更佳的微调方向.

在每次完成文本增强后,算法2被调用并执行LoRA微调. 在第n轮CL训练结束时,利用LLM生成贴近目标域风格的增强样本,形成新训练数据集$\mathcal{D}_{n,m}$(见第14行). 随后,调用算法1合并执参数(见第16行),以更新跨语言适配模型(见第17行). 最大微调轮次M可按需设定,以控制联合训练代价. 当$m > M$时,微调终止.

2.4 LLM辅助跨语言联合训练

本小节设计LLM辅助跨语言联合训练机制. 在每轮迭代中,LLM利用源域样本生成风格接近目标域的增强数据,以缩小特征空间中的跨语言语义差异;CL检测网络则基于原始数据与增强样本共同优化特征提取与分类能力,并通过双维度对比在词元级局部语义与句子级全局语义上对齐跨语言特征. 训练中产生的对比损失反馈被用来指导LLM的微调方向,促使其在后续联合训练中生成更贴合目标域特征分布的文本样本.

需要说明的是,在联合训练中,LLM并非直接参与特征提取或分类过程,而是被策略性地用于增强CL网络的训练. 在测试阶段,样本被输入CL网络并通过分类器预测标签. 跨语言联合训练工作流归纳为算法3,主要包含四个步骤:

(1)跨语言文本增强. 在每次文本增强,源域$\mathcal{S}$与目标域$\mathcal{T}$的原始数据被统一输入LLM,以生成贴近目标域风格的多样化训练样本. 在2.1节所设计Prompt的约束下,源域中的实例$x_{\mathcal{S}}^{(i)}$被增强为符合目标域风格的样本$x_{\mathcal{S}\to\mathcal{T}}^{(i)}$;同理,目标域中的实例$x_{\mathcal{T}}^{(i)}$被增强为$x_{\mathcal{T}\to\mathcal{T}'}^{(i)}$(见第1~4行). 最终构建的混合样本集合$\{x_{\mathcal{S}}^{(i)}, x_{\mathcal{S}\to\mathcal{T}}^{(i)}, x_{\mathcal{T}\to\mathcal{T}'}^{(i)}\}$为后续CL提供了接近目标域分布的训练样本.

(2)层次特征抽取. 在获得增强后的输入序列后,模型对源域样本$x_{\mathcal{S}}^{(i)}$进行双维度表征建模. $x_{\mathcal{S}}^{(i)}$的词单元$w_{\mathcal{S}}^{(i,j)}$被嵌入函数Embed(·)映射为包含位置编码的d维向量$\boldsymbol{h}_{\mathcal{S}}^{(i,j,0)}$(见第7行),并基于式(2)得到初始特征表示. 随后,第K层Transformer编码器执行层次语义建模:在第k层,利用式(3)更新词级特征(见第11行),并由式(4)输出浅层局部特征$\boldsymbol{h}_{\mathcal{S}}^{(i,k)}$;在第$K$层的编码器会整合全句上下文,生成全局特征$\boldsymbol{h}_{\mathcal{S}}^{(i,K)}$(见第14行). 这一过程同样用于处理$x_{\mathcal{S}\to\mathcal{T}}^{(i)}$和$x_{\mathcal{T}\to\mathcal{T}'}^{(i)}$,实现跨语言一致性语义表示.

(3)双维度对比. 在投影空间中,浅层局部和全局语义特征$\boldsymbol{h}_{\mathcal{S}}^{(i,k)}$和$\boldsymbol{h}_{\mathcal{S}}^{(i,K)}$被式(5)映射为投影向量$\boldsymbol{z}_{\mathcal{S}}^{(i,k)}$和$\boldsymbol{z}_{\mathcal{S}}^{(i,K)}$(见第15行). 对于批次$b$的样本投影,采用式(7)计算锚点与正样本的相似度,并用式(8)、(9)计算锚点与同语言、跨语言负样本的相似度. 它们被代入式(11)后得到分层对比损失$\mathcal{L}_n^{(k,b)}$,随后,通过式(12)在批次维度聚合得到$\mathcal{L}_n^{(k)}$(见第16行). 最终,浅层与全局损失被融合为当前epoch的对比损失$\mathcal{L}_n$(见第19行).

(4)LLM-CL对齐. 在整个训练过程中,采用周期性增强策略实现LLM-CL的协同优化:每经过T个epoch,执行一次文本增强操作. 在获得当前epoch的$\mathcal{L}_n$后,将其代入式(18)计算LLM增强效果的评估指标φ_n. 算法2通过参考φ_n执行基于LoRA的微调,随后调用算法1合并参数,形成新的跨语言适配LLM(见第21行).

联合训练过程中,LLM生成具有目标域风格的增强样本,以缓解CL网络跨语言训练种源域与目标域之间的特征分布偏移;训练过程中的产生对比损失用于间接评估LLM生成样本的语义丰富性与可区分性. 通过迭代,LLM的跨语言文本增强策略持续对齐CL的辨别能力,形成一个耦合闭环.

3 实验设计与结果分析

Weibo与PHEME两个具有显著语境差异的数据集被选为实验对象. 前者是中文社交媒体数据集,主要记

算法3　联合DA训练

输入: $S=\{x_S^{(i)}\}_{i=1}^{I}, T=\{x_T^{(i)}\}_{i=1}^{I}, a, \delta, \theta_0, \boldsymbol{W}_1, \boldsymbol{W}_2$

输出: θ

For i=1 to I do
　$x_{S\to T}^{(i)} \leftarrow \mathrm{LLM}(x_S^{(i)})$;
　$x_{T\to T'}^{(i)} \leftarrow \mathrm{LLM}(x_T^{(i)})$;
End For
For n=1 to N do
　For j=1 to C do:
　　$h_S^{(i,j,0)} \leftarrow \mathrm{Embed}(x_S^{(i,j)})$;
　End For
　For k'=1 to K do:
　　For j=1 to C do:
　　$h_S^{(i,j,k)} \leftarrow \mathcal{F}^{(k)}(\{h_S^{(i,c,k-1)}\}_{c=1}^{C})_j$;
　　End For
　　If $k'=k$ or $k'=K$
　　$\boldsymbol{h}_S^{(i,k')} \leftarrow [h_S^{(i,1,k')}, \cdots, h_S^{(i,C,k')}]$;
　　$z_S^{(i,k')} \leftarrow \boldsymbol{W}_2 \cdot \sigma(\boldsymbol{W}_1 \cdot \boldsymbol{h}_S^{(i,k')})$;
　　$\mathcal{L}_n^{(k')} \leftarrow \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_n^{(k',b)}$;
　　End If
　End For
　$\mathcal{L}_n \leftarrow \delta\mathcal{L}_n^{(K)} + (1-\delta)\mathcal{L}_n^{(k)}$;
　If $n \bmod T = 0$
　　$\theta \leftarrow \mathrm{Alignment}(\theta, n, m, S_1, S_2, \lambda)$;
　End If
End For

录日常社会话题讨论;后者是英文数据集,专门收集突发事件相关的传播链条. 两者不仅在语言与文化层面存在差异,在内容主题上也有不同侧重. 前者聚焦常规舆情,后者针对危机事件. 这样的跨语言、跨场景差异,为验证所提方法的泛化能力与适应性提供了理想测试环境. 在数据规模方面,Weibo包含4 664条样本(谣言2 356条、非谣言2 308条);PHEME含有5 805条样本(谣言2 374条、非谣言3 431条). 两者均按照70%、20%、10%的比例划分为训练集、测试集和验证集.

为了全面、量化地评估性能,选取了如下三种类型的基准方法:

(1)主流商业LLM(ChatGPT-4o)直接检测;

(2)LLM辅助的深度学习方法(LACL[28]):未适配跨语言检测任务,仅支持单向文本增强;

(3)主流深度学习检测模型(包括LSTM[29]、Text-CNN[30]、RCNN[31]和HAN[32]):LSTM依赖序列建模但缺乏跨语言语义对齐;Text-CNN通过卷积提取局部特征,泛化能力有限;RCNN结合局部与全局建模,性能优于前者.

广泛采用的标准精确率(Precision)和F1分数[对应式(20)和式(21)]被用于性能分析.

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{20}$$

$$F1=\frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{21}$$

表2归纳了所提方法的消融分类及相应的实验设置. LLM规模(7B与14B)、文本增强方向(单向/双向)和微调阶段(一至三轮)三个维度被用来构建对比实验. 其中,Proposed-1作为消融实验的对照,被用来量化双向文本增强策略的性能增益;Proposed-2至-4为无微调条件下的增强轮次对照,分别执行一至三轮双向文本增强;Proposed-5和-6在三轮增强基础上进行1-2轮微调;所有实验均基于7B模型. Proposed-7则采用14B规模的Qwen模型执行三轮双向增强. 为确保实验的公平性,所有的超参数均保持一致,详细参数设置见表3.

表2　所提方法的消融分类

所提方法	基座LLM参数量/B	增强轮次(w)	微调轮次(m)	增强方向
Proposed-1	7	3	0	单向
Proposed-2	7	1	0	双向
Proposed-3	7	2	0	双向
Proposed-4	7	3	0	双向
Proposed-5	7	3	1	双向
Proposed-6	7	3	2	双向
Proposed-7	14	3	0	双向

表3　默认参数设置

参数	数值
每批次数据量/a	256
epoch轮次/N	25
文本增强次数/T	10
权重参数/δ	0.7

设计了三组实验:(1)方法性能边界验证实验,确定算法理论上下限;(2)消融实验用于验证双向增强机制的有效性,并量化分析增强轮次对模型性能的影响;(3)微调轮次与检测效果相关性实验,探究不同的LLM微调次数对性能的影响.

3.1　DA性能的基准测试

在所提联合训练框架下,LLM的语义建模与生成能力用于增强CL模型的跨语言检测性能. 换言之,实际检测任务依赖轻量化的CL分类器,而非高计算代价的LLM. 为验证通用LLM直接检测的局限性,我们直接采用GPT-4o对社交媒体内容进行二分类,并在Weibo与PHEME中各随机抽取200条样本进行测试. 如表4所示,它的平均精确率仅为52.2%和56.3%,表明主流通用LLM缺乏针对特定任务的判别能力,难以在跨语言与文化差异场景下稳定识别虚假信息.

表4 ChatGPT-4o的直接检测结果

数据集	True	False	精确率/%
Weibo	58.3	46	52.2
PHEME	63.6	49.2	56.4

进一步对比传统CNN与最新LACL的跨语言检测性能(见表5). 以Weibo为源域时,CNN的精确率和F1均值分别为38.53%和38.51%;若以PHEME为源域,平均精确率为43.32%,而False类别的F1值仅为5.20%,F1均值为31.50%. LACL的跨语言检测性能虽存在下降,但总体保持稳定. 以Weibo为源域时,其精确率为51.90%,较同域测试精度(89.70%)下降约37%,平均F1分数为44.49%,其中True类别的F1分数下降了近66%. 以PHEME为源域时,精确率为53.01%,较同域测试精度(90.61%)下降37.6%,其中Non-Rumor类别的F1值仅为14.45%,下降幅度达76.54%.

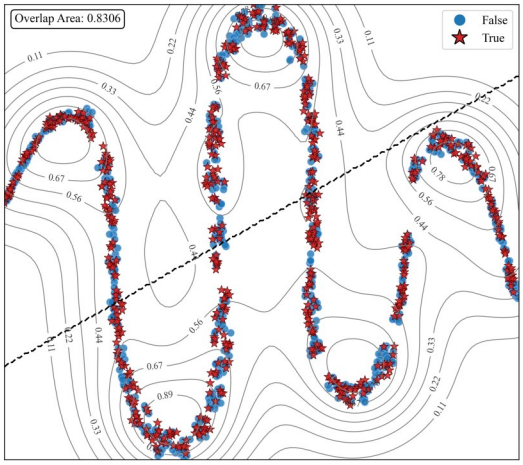
表5 跨语言检测性能对照

跨语言任务	方法	精确率/%	F1/%		平均F1/%
			False	True	
Weibo→PHEME	LACL	51.90	64.76	24.23	44.49
Weibo→PHEME	CNN	38.53	37.95	39.07	38.51
PHEME→Weibo	LACL	53.01	14.45	67.60	41.02
PHEME→Weibo	CNN	43.32	5.20	57.80	31.50

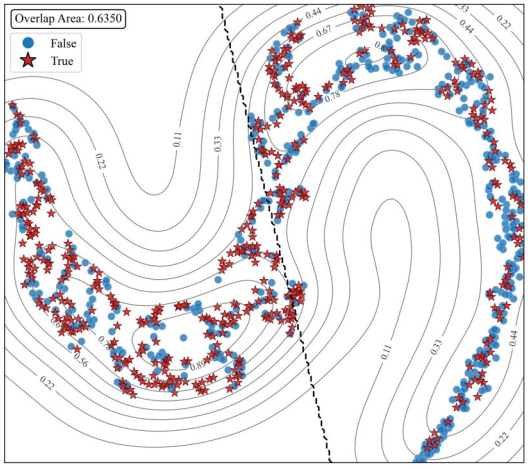
图5可视化了LACL的跨语言特征分布. 特征通过t-SNE投影至二维空间,随后基于核密度估计(Kernel Density Estimation,KDE)计算True/False类样本的概率密度分布. 为了量化两类样本在特征空间中的可分离性,在每个网格点计算两类概率密度的最小值并进行积分,获取重叠面积作为指标. 此外,我们在二维特征空间上训练一个线性支持向量机(Support Vector Machine, SVM)分类器,用来决定绘制分类边界(虚线),以反映两类样本的可分离性(弥补KDE重叠面积在局部判别上的不足). 通过观测可见,无论以Weibo还是PHEME为源域,True与False类的样本在特征空间中均存在大面积重叠,导致跨语言检测性能受限. 后续可视化特征图均采用上述方法绘制,细节不再赘述.

以上结果与表5中LACL在跨语言检测上精确率和F1分数显著下降的特点高度一致,表明该方法在跨语言与跨文化差异下泛化能力不足. 虽然在精确率和F1值上比CNN平均提升约10%,但高度混叠的特征表明LACL缺乏稳健性和鲁棒性.

对于Weibo→PHEME的检测任务(见表6),Proposed-2的精确率与F1分数均未超越基准模型. 在False和True类别上,它的精确率比HAN和Text-CNN分别低4.23%和2.83%. Proposed-7在所有类别上的指标上均优于其他模型. 对于PHEME→Weibo(见表7),虽



(a) 源域:Weibo



(b) 源域:PHEME

图5 LACL的DA特征分布

表6 不同增强轮次、增强方向下Weibo→PHEME检测性能

方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
LSTM	81.81	82.89	82.35	82.83	81.81	82.32
Text-CNN	80.21	85.31	82.76	83.47	81.40	82.44
RCNN	81.54	80.40	80.97	81.20	80.75	80.97
HAN	83.75	82.76	83.25	83.47	83.04	83.26
Proposed-2	79.52	82.48	81.00	81.71	80.00	80.85
Proposed-7	89.03	88.81	88.92	89.11	88.73	88.92

然Proposed-2在False类别上的精确率(83.51%)略逊于Text-CNN(84.49%),但在True类别及综合指标上超越了LSTM、RCNN和HAN. Proposed-7延续其优势,在False/True类别精确率和平均F1值上均达到峰值.

图6中,在Weibo→PHEME的跨语言任务下,Proposed-2的False(79.52%)、True(82.48%)和平均精确率

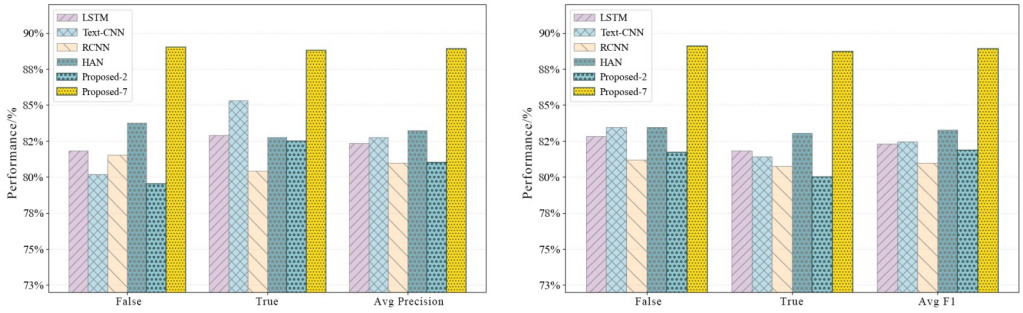
表7　不同增强轮次、增强方向下PHEME→Weibo检测性能

方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
LSTM	77.11	75.31	76.21	75.76	76.58	76.17
Text-CNN	84.49	73.11	78.80	75.23	79.68	77.46
RCNN	78.50	75.71	77.11	76.43	77.61	77.02
HAN	81.62	77.09	79.35	78.34	79.95	79.15
Proposed-2	83.51	84.53	84.02	84.13	83.89	84.01
Proposed-7	91.96	88.84	90.40	90.15	90.52	90.33

(81.00%)均处于低位,F1值也未见提升;Proposed-7的柱形高于其他方法. 类似地,图7中,Proposed-2仅在False类别略低于Text-CNN,其余指标柱形均高于基准方法. Proposed-7再次呈现出对基线方法的显著优势.

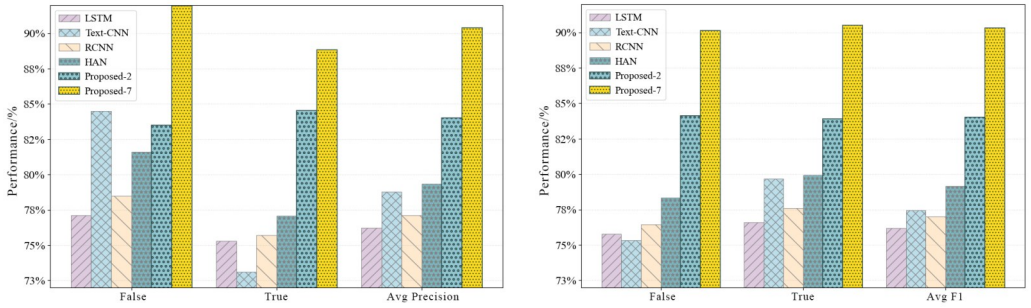
在PHEME→Weibo的任务上,Proposed-2在F1和精度指标上优于基准方法,但对于Weibo到PHEME,Proposed-2性能稍弱于基准方法,间接证实了联合训练对增强轮次存在一定的敏感性. Proposed-7凭借三轮增强提升了数据多样性和检测鲁棒性,这是性能领先的重要原因之一.

MF-XLM-R+ST+GL[33]是目前跨语种虚假信息检测的代表性方法之一. 根据文献中的结果,它在Weibo到PHEME和PHEME到Weibo的跨语言任务下的平均准确率为77.9%和77.2%,低于所提方法(特别是Proposed-7)的性能上界.



(a) 精确性　(b) F1分数

图6　Weibo→PHEME检测性能对比



(a) 精确性　(b) F1分数

图7　PHEME→Weibo检测性能对比

3.2　消融实验:双向增强有效性分析

本节通过消融实验验证双向文本增强策略的有效性,并考察不同增强轮次对检测性能的影响. 如表2所示,Proposed-2至-7均采用双向增强,即从源域迁移到目标域时,同步实施面向目标域风格的文本增强;而Proposed-1仅使用三轮单向增强(源域到目标域),作为对比基准.

从表8中的结果可见,在Weibo→PHEME的检测任务中,Proposed-1的平均精确率和F1值相较Proposed-2分别下降4.45%和4.52%;与经过三轮文本增强的Proposed-7相比,降幅接近13%. 对于表9的PHEME→Weibo任务,Proposed-1同样处于劣势,其平均精确率和F1分数相较Proposed-2与-7分别下降约5%和11%.

在双向增强的内部对比中(Proposed-2至-4),表8显示Weibo→PHEME任务下,Proposed-4仅在True类别的精确率较Proposed-3微降0.32%,其余指标均持续提升. 根据表9中PHEME→Weibo的结果,Proposed-4在False类别上的精确率相较Proposed-3下降了0.49%,而Proposed-3在True类别上的精确率较Proposed-2降低0.34%. 总体而言,这些方法的平均精确率和F1分数与

表8 不同增强轮次、增强方向下Weibo→PHEME检测性能

方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
Proposed-1	77.57	74.34	75.95	75.08	76.58	75.83
Proposed-2	79.52	82.48	81.00	81.71	80.00	80.85
Proposed-3	83.60	84.11	83.86	84.22	83.46	83.84
Proposed-4	87.76	83.79	85.77	85.59	85.83	85.71
Proposed-7	89.03	88.81	88.92	89.11	88.73	88.92

表9 不同增强轮次、增强方向下PHEME→Weibo检测性能

方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
Proposed-1	81.03	78.11	79.57	78.98	80.00	79.49
Proposed-2	83.50	84.53	84.02	84.13	83.89	84.01
Proposed-3	87.41	84.19	85.80	85.40	86.04	85.72
Proposed-4	86.92	85.83	86.38	86.27	86.47	86.37
Proposed-7	91.96	88.84	90.40	90.15	90.52	90.33

增强轮次呈正相关,仅在类别级别上存在轻微波动.

3.3 LLM微调轮次对跨语言检测的影响

如表10所示,在Weibo→PHEME检测任务中,Proposed-5相较于Proposed-4的平均精确率提升约2.3%,其中True类别的检测精确率增幅达3.9%. 然而,在Proposed-6下,True类别和平均精确率分别下降1.56%和0.11%,但在其余指标上仍保持优势. 表11中PHEME→Weibo的结果呈现相似规律:Proposed-5在平均精确率和F1值上相较前代模型实现3.32%的提升,而Proposed-6则出现0.69%和1.35%的轻微下降. Proposed-7在大部分指标上展现了优越性能,仅在True类别的精确率上比Proposed-5低0.68%.

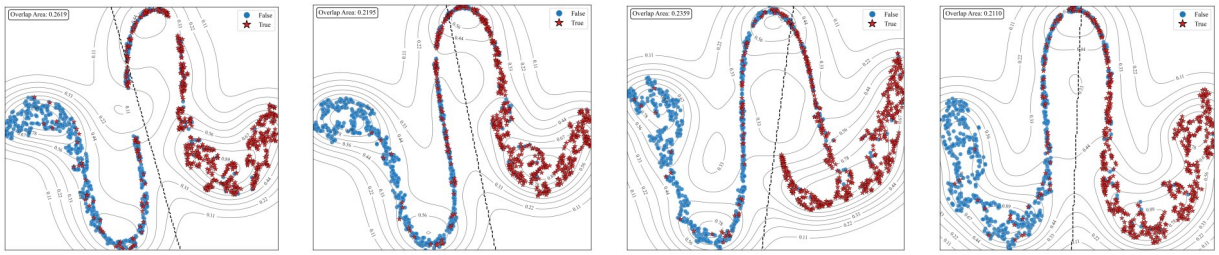
表10 不同微调轮次下Weibo→PHEME检测性能

方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
Proposed-4	87.76	83.79	85.77	85.59	85.83	85.71
Proposed-5	88.46	87.69	88.07	88.24	87.91	88.07
Proposed-6	89.89	86.04	87.96	88.82	87.99	88.41
Proposed-7	89.03	88.81	88.92	89.11	88.73	88.92

表11 不同微调轮次下PHEME→Weibo检测性能

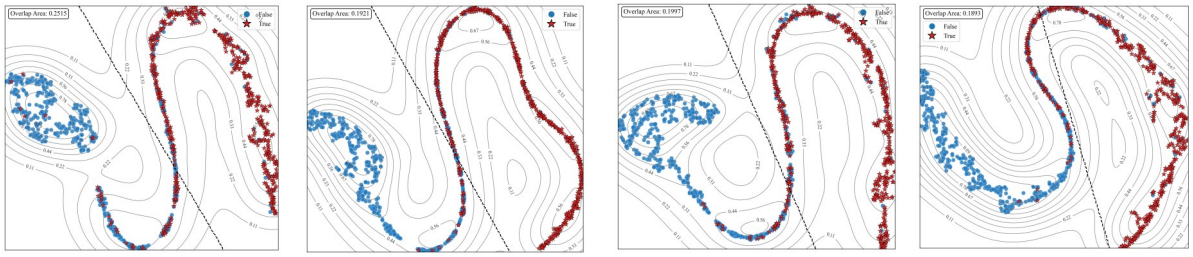
方法	精确率/%		平均精确率/%	F1/%		平均F1/%
	False	True		False	True	
Proposed-4	86.92	85.83	86.38	86.27	86.47	86.37
Proposed-5	89.87	89.52	89.70	89.67	89.72	89.69
Proposed-6	90.60	87.42	89.01	88.71	89.16	88.94
Proposed-7	91.96	88.84	90.40	90.15	90.52	90.33

我们仍然用特征可视化分析(图8和9)解释这些现象. 根据图8(*b*)、(*c*),Proposed-6在True类别上的特征点在0.89等密度线外呈双侧分散分布,解释了其精确率下降的潜在原因. 相对而言,关于False类别的密度峰值从Proposed-5的89%和56%提升至67%和89%,对应检测精度的提升. 进一步对比图9可见,Proposed-5密度双峰分别为76%和78%,而Proposed-6则下降为56%和44%,表明特征空间聚合度的减弱,这与True类别精度的下降相关.



(*a*) Proposed-4下的结果 (*b*) Proposed-5下的结果 (*c*) Proposed-6下的结果 (*d*) Proposed-7下的结果

图8 不同微调轮次下Weibo→PHEME检测的特征可视化



(*a*) Proposed-4下的结果 (*b*) Proposed-5下的结果 (*c*) Proposed-6下的结果 (*d*) Proposed-7下的结果

图9 不同微调轮次下PHEME→Weibo检测的特征可视化

Proposed-6在部分指标上退化的原因在于第二轮微调受前一轮增强数据的影响. 由于LLM参数合并的权重设置参考了对比损失,此时微调后的LLM可能过度拟合前一轮增强数据中的局部噪声模式(如特定句式结构的误增强),降低了生成的文本与目标域真实分布的语义一致性,导致CL检测模型在个别指标上性能下降.

4　结束语

本文提出了一种面向社交网络虚假信息检测的LLM增强的自监督DA方法. 该方法通过跨语言风格迁移的LLM文本增强机制,克服传统分布对齐方法对浅层特征的依赖,减轻文化差异带来的语义偏移. 双维度对比框架建模词汇与语义两层特征,提升了模型对复杂表达与跨语言偏移的鲁棒性. 联合训练框架实现跨了语言文本生成与特征学习的闭环优化. 在跨语言数据集Weibo与PHEME上的实验结果表明,所提方法在整体性能上优于通用LLM直接检测、基准深度学习模型及最新LLM增强方法. 未来研究将探索多模态融合与持续学习机制,以增强检测模型在复杂社交网络中的适应性与实用性.

参考文献

[1] 高玉君, 梁刚, 蒋方婷, 等. 社会网络谣言检测综述[J]. 电子学报, 2020, 48(7): 1421-1435.
GAO Y J, LIANG G, JIANG F T, et al. A summary of social network rumor detection[J]. Acta Electronica Sinica, 2020, 48(7): 1421-1435. (in Chinese)

[2] WILSON G, COOK D J. A survey of unsupervised deep domain adaptation[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(5): 1-46.

[3] SHANG L Y, ZHANG Y, CHEN B Z, et al. MMAdapt: A knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection[C]//Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 4653-4663.

[4] LIM J Y, LIM K M, LEE C P, et al. SCL: Self-supervised contrastive learning for few-shot image classification[J]. Neural Networks, 2023, 165: 19-30.

[5] ZHENG P, QIN J, WANG S, et al. Memory-aided contrastive consensus learning for co-salient object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 3687-3695.

[6] LIU Y H, HUANG L, GIUNCHIGLIA F, et al. Improved graph contrastive learning for short text classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 18716-18724.

[7] YIN Y J, ZENG J L, SU J S, et al. Multi-modal graph contrastive encoding for neural machine translation[J]. Artificial Intelligence, 2023, 323: 103986.

[8] 欧阳祺, 陈鸿昶, 刘树新, 等. 基于Bert-GNNs异质图注意力网络的早期谣言检测[J]. 电子学报, 2024, 52(1): 311-323.
OUYANG Q, CHEN H C, LIU S X, et al. Early rumor detection based on bert-GNNs heterogeneous graph attention network[J]. Acta Electronica Sinica, 2024, 52(1): 311-323. (in Chinese)

[9] JIANG Y Q, HUANG C, HUANG L H. Adaptive graph contrastive learning for recommendation[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2023: 4252-4261.

[10] LIU N, WANG X, HAN H, et al. Hierarchical contrastive learning enhanced heterogeneous graph neural network[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(10): 10884-10896.

[11] 兰玉乾, 饶元, 李冠呈, 等. 基于内在质量约束的文本生成和评价综述[J]. 电子学报, 2024, 52(2): 633-659.
LAN Y Q, RAO Y, LI G C, et al. A survey of text generation and evaluation based on intrinsic quality constraints[J]. Acta Electronica Sinica, 2024, 52(2): 633-659. (in Chinese)

[12] HU B Z, SHENG Q, CAO J, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(20): 22105-22113.

[13] HU W B, XU Y F, LI Y, et al. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(3): 2256-2264.

[14] DOU C X, SUN X H, WANG Y S, et al. Domain-adapted dependency parsing for cross-domain named entity recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(11): 12737-12744.

[15] ZHANG X H, YU B W, CONG X, et al. Cross-domain NER under a divide-and-transfer paradigm[J]. ACM Transactions on Information Systems, 2024, 42(5): 1-32.

[16] WANG Y, XIE H, HE J Y, et al. Cross-domain semantic transfer for domain generalization[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2025, 21(5): 1-24.

[17] ZHANG Z, LIU M H, WANG A H, et al. Collaborate to adapt: Source-free graph domain adaptation via bi-directional adaptation[C]//Proceedings of the ACM Web Conference 2024. New York: ACM, 2024: 664-675.

[18] MENG Q W, QIAN H W, LIU Y, et al. MHCCL: Masked hierarchical cluster-wise contrastive learning for multivariate time series[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(8): 9153-9161.

[19] BAI S H, ZHANG M, ZHOU W Q, et al. Prompt-based distribution alignment for unsupervised domain adaptation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(2): 729-737.

[20] MA J, GAO W, MIRTRA P, et al. CHA M. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). California: IJCAI, 2016: 3818-3824.

[21] ZUBIAGA A, LIAKATA M, PROCTER R. Learning reporting dynamics during breaking news for rumour detection in social media[EB/OL]. (2016-10-24)[2025-02-24]. https://arxiv.org/abs/1610.07363.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. San Diego: NeurIPS, 2017: 6000-6010

[23] LV K, YANG Y Q, LIU T X, et al. Full parameter fine-tuning for large language models with limited resources[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 8187-8198.

[24] GU Y X, HAN X, LIU Z Y, et al. PPT: Pre-trained prompt tuning for few-shot learning[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 8410-8423.

[25] DING N, QIN Y J, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.

[26] HU E J, WALLIS P, ALLEN-ZHUl Z, et al. LoRA: Low-rank adaptation of large language models[C]//Proceedings of International Conference on Learning Representations. Appleton: ICLR, 2022, 1-20.

[27] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 9726-9735.

[28] SHEN H, LI X, WANG X, et al. LLM-augmented contrastive learning for misinformation detection in social networks[J]. IEEE Transactions on Computational Social Systems, 2025. DOI:10.1109/TCSS.2025.3599080.

[29] BECK M, PÖPEL K, SPANRING M, et al. xLSTM: extended long short-term memory[C]//Advances in Neural Information Processing Systems 37, San Diego: NeurIPS, 2024: 107547-107603.

[30] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1746-1751.

[31] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[32] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2016: 1480-1489.

[33] TIAN L, ZHANG X Z, LAU J H. Rumour Detection via Zero-shot Cross-lingual Transfer Learning[M]//Machine Learning and Knowledge Discovery in Databases. Research Track. Cham: Springer International Publishing, 2021: 603-618.

作者简介

沈　航　男,1984年3月出生于江苏省南京市.现为南京工业大学计算机与信息工程学院(人工智能学院)副教授、硕士生导师.主要研究方向为领域大模型、网络安全和空天地一体化网络.

E-mail: hshen@njtech.edu.cn

王　旭　男,1999年10月出生于江苏省镇江市.2025年6月硕士毕业于南京工业大学,现为镇江电信工程师.主要研究方向为大模型、无监督和半监督深度学习及网络安全应用.

E-mail: Lhasahi@163.com

王天荆　女,1977年7月出生于江苏省扬州市.现为南京工业大学计算机与信息工程学院(人工智能学院)副教授、硕士生导师.主要研究方向为网络安全、车联网及自动驾驶.

E-mail: wangtianjing@njtech.edu.cn

戴远飞　男,1992年10月出生于山东省威海市.南京工业大学计算机与信息工程学院(人工智能学院)副教授、硕士生导师.主要研究方向为大模型、知识图谱技术及应用.

E-mail: daiyuanfei@njtech.edu.cn

白光伟　男,1961年11月出生于辽宁省沈阳市.现为南京工业大学计算机与信息工程学院(人工智能学院)教授、博士生导师.主要研究方向为未来网络、人工智能和区块链.

E-mail: bai@njtech.edu.cn