

基于“话题—情感”协同的网络语言暴力动向检测

阴凯东¹, 张茂元^{1,2}, 李晨阳³

(1. 华中师范大学 计算机学院; 2. 国家语言资源监测与研究网络媒体中心;
3. 华中师范大学 人工智能教育学部, 湖北 武汉 430079)

摘要: 网络欺凌现象愈发凸显, 日益成为不容忽视的严重社会问题。现有网络语言暴力检测方法通常仅从话题或情感单一视角建模, 忽略了两者的不一致性, 如情感特征未与其所属话题正确匹配, 从而导致检测结果存在语义偏差与误判现象。为此, 提出一种基于“话题—情感”协同演化的网络语言暴力动向检测与预警方法, 通过融合BERT语义表示、情感、话题与时间特征构建协同演化模型 Co-Evolution Model, 挖掘网络暴力的潜在演化规律, 生成协同分布图以直观展示动态趋势, 并基于动态阈值与一阶差分方法实现暴力风险的自动识别预警。实验结果显示, 所提方法在SCCD和CHNCI两个数据集的准确率和F1得分均优于对照方法, 特别是在标签稀疏的情况下表现出更强的鲁棒性, 为舆情预警与干预提供了可靠技术支持。

关键词: 网络语言暴力; 协同演化; 话题分类; 情感分析; 预警机制

DOI: 10.11907/rjdc.251566

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)008-0001-06

扫描二维码阅读全文:



Online Language Violence Trends Detection Based on Topic-Sentiment Synergy

YIN Kaidong¹, ZHANG Maoyuan^{1,2}, LI Chenyang³

(1. School of Computer Science, Central China Normal University;
2. National Language Resources Monitoring and Research Network Media Center;
3. Faculty of Artificial Intelligence Education, Central China Normal University, Wuhan 430079, China)

Abstract: Cyberbullying has become increasingly prominent and is emerging as a severe social issue that cannot be overlooked. Existing detection methods for online verbal violence often model from a single perspective, either topic or emotion, neglecting the inconsistency between the two—such as emotional features not matching their corresponding topics, leading to semantic bias and misjudgment in detection results. To address this, a novel method for detecting and warning about online verbal violence trends based on “topic-emotion” co-evolution is proposed. By integrating BERT semantic representations, emotion, topic, and temporal features, a co-evolution model is constructed to uncover the underlying evolutionary patterns of online violence. A co-distribution map is generated to visually display dynamic trends, while automatic risk identification and early warning are achieved through dynamic thresholds and first-order difference methods. Experimental results demonstrate that the proposed method outperforms baseline approaches in both accuracy and F1 scores on the SCCD and CHNCI datasets, particularly exhibiting stronger robustness in scenarios with sparse labels, providing reliable technical support for public opinion monitoring and intervention.

Key Words: cyberbullying; co-evolution; topic classification; sentiment analysis; early-warning mechanism

收稿日期: 2025-08-19

基金项目: 国家语委“十四五”重点科研项目 (No. ZDI145-94)

作者简介: 阴凯东 (2002-), 男, CCF 学生会员, 华中师范大学计算机学院硕士研究生, 研究方向为自然语言处理; 张茂元 (1975-), 男, 华中师范大学计算机学院教授、博士生导师, 研究方向为自然语言处理、人工智能; 李晨阳 (1999-), 男, 华中师范大学人工智能教育学部博士研究生, 研究方向为自然语言处理。本文通讯作者: 张茂元。

0 引言

网络语言暴力以文字、语音或图像的形式实施攻击、侮辱、威胁行为,且具有匿名性强、传播快、影响广等特征,对受害者的身心健康乃至公共安全构成风险^[1]。早期研究多采用支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(Naive Bayes, NB)或随机森林(Random Forest, RF)等经典机器学习模型,利用 n-gram、TF-IDF(Term Frequency-Inverse Document Frequency)及情感词典特征实现分类,但存在两方面局限:一是特征工程过于依赖人工经验,难以覆盖隐形侮辱、讽刺等复杂表达;二是模型泛化能力受限,当训练语料领域或话题变化时,性能会显著下降^[2]。鉴于此,政府与学界正积极推进基于人工智能技术的网络语言暴力自动化检测,以实现其早期识别、风险预警与干预,从而降低社会危害。

1 相关研究

目前已有多种网络暴力自动检测方法,如基于深度学习的文本分类模型、基于预训练语言模型以及大语言模型的暴力内容识别方法^[3]。卷积神经网络(Convolutional Neural Network, CNN)和循环神经网络(Recurrent Neural Network, RNN)能够学习文本的局部或序列特征,被广泛应用于网络暴力检测^[4]。随后,基于Transformer的预训练语言模型成为主流,例如BERT引入双向上下文编码,对语料具有更强的语义捕捉能力^[5];HateBERT则通过预训练使攻击性表达更加敏感^[6]。近年来,轻量化 Prompt-learning 与参数高效调优技术应用广泛,其能在仅有几百条标注的少样本情况下进行匹配,甚至超越全量微调性能^[7]。此外,大语言模型跨域推断能力表现突出,例如 Qwen-7B 在多语言仇恨检测上取得了优于传统微调模型的效果^[8]。然而,以上方法存有明显局限性,其往往只考虑文本本身的语义特征,或以话题或情感单一视角建模,忽视了两种特征之间的不一致性,未将二者对齐,如情感特征未与其

所属话题正确匹配,从而产生语义偏差并导致检测误判。如表1所示,单靠情感或话题的分析很容易出现偏差。此外,其还缺乏对网络暴力风险趋势的精准量化与预警,难以实现对网络暴力动向的早期识别与及时响应。

Table 1 Example sentences

表1 例句

句子	情感分析	话题分析	正确标签	检测错误
我觉得他们的文化很有趣,但他们还是应该回到自己的国家去(仅情感)	正面	种族	网暴	偏差
你真是没用(仅话题)	负面	无	网暴	偏差

情感分析中的情感词典、情感分类器常被用作网络暴力检测的辅助特征^[9]。Shen 等^[10]发现在公共健康谣言场景中,主题与情绪存在显著的同频共振关系,验证了“主题—情绪”协同演变对风险治理的价值。为此,本文提出一种“话题—情感”协同的网络语言暴力演化动向检测与预警方法 Co-Evolution Model,主要创新点为:①借助迁移学习建立话题、情感与时间三维信息的跨域映射,实现多维标签的自动对齐与补全,从数据层面提升协同建模的一致性完整性;②将话题、情感及时间特征联合投射到统一潜空间,通过非线性变换生成具备协同演化语义的高阶表示,为后续检测与预测提供具有强表达力的特征向量;③利用协同演化分析增强网暴演化动向的感知力,既能捕捉突发性爆发,又能对中度或短时上涨进行及时、精准的风险预警。实验结果表明,所提方法具有更高的准确性和鲁棒性,为舆情治理提供了理论与工具支撑,具有一定的现实应用价值。

2 网络语言暴力检测方法

如图1所示,本文提出的模型架构由基于迁移学习的标签补全、协同演化建模及预警机制组成。

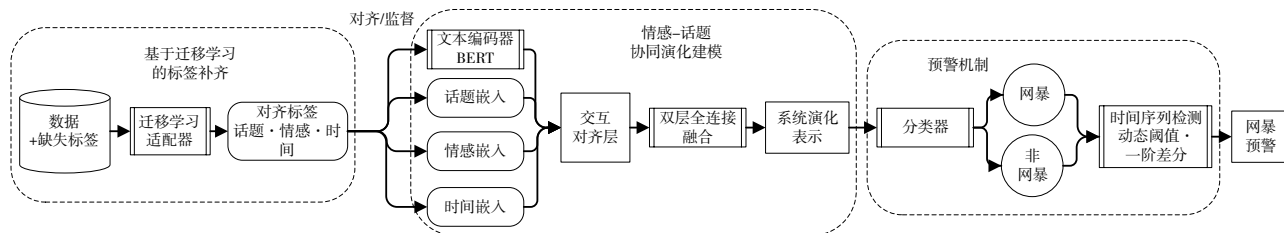


Fig. 1 Architecture of temporal evolution of cyberbullying

图1 网络语言暴力演化检测模型架构

2.1 基于迁移学习的标签补全

在公开的中文社交媒体网暴数据集中,SCCD数据集标注了评论相对时间^[11],CHNCI数据集标注了绝对时间

戳^[12]。由于时间维度对本文研究至关重要,二者是最适合的数据集。然而,SCCD提供了话题标签,但缺少情感标签;CHNCI除缺乏情感维度外,还未提供话题分类。缺失

的多维标签不仅削弱了数据的描述力,而且限制了模型对“话题—情感—时间”三元演化机制的学习深度,因此本文设计了“人工标注+迁移学习+人工核验”的标签补全流程。

组织具有语言学背景的标注员为 SCCD 数据集逐条补充三元情感标签,完善其维度并构建高质量源域数据。CHNCI 数据集同时缺失话题与情感标签,且包含 220 676 条评论,完全依赖人工标注成本过高,因此采用标注完整的 SCCD 微调大语言模型,对 CHNCI 实施标签迁移。为确保标签质量,随机抽取 20% 的自动标注样本进行双人交叉核验,若一致性低于 85%,则提高抽检比例、细化准则,并经人工讨论消歧。另将 CHNCI 的绝对时间戳转换为相对时间,以满足协同演化模型的相对时序分析要求。上述流程系统性补全并统一了两个数据集的标签空间,为后续协同演化研究奠定了数据基础。

标签补全后实施系统化清洗与预处理,对全体样本进行缺失值检测,剔除任一维度存在缺失的记录。社交平台上网暴的发生往往与特定时间段有关,因此将评论相对时间离散为 24 个时间桶,以刻画时间特征并识别时段聚集性。采用 TOPIC2ID 与 SENT2ID 将话题与情感标签映射为数值编码,便于后续建模。为保证与预训练 BERT 的兼容性,文本仅经 BERT 分词器统一编码,不额外进行去噪或正则化处理。以上流程在保证原始语义完整性的同时实现了数据格式的标准化与高效加载。

2.2 “情感—话题”协同演化建模

联合话题、情感与时间特征,构建“话题—情感”协同演化模型。首先,由 BERT 获取文本语义表示,与 3 类离散特征进行向量级融合;其次,通过两层感知机提炼融合表示中的协同演化信号;最后,由分类头完成网暴判别回传梯度,实现文本编码、多维嵌入、特征融合与分类决策的完整流程。

2.2.1 文本编码与多维嵌入

首先,将评论文本编码为 $input_ids$, 配套 $attention_mask$ 屏蔽补齐位,输入 BERT 中得到上下文语义表示 $h_{BERT}^{[7]}$ 。表示为:

$$h_{BERT} = \text{BERT}(input_ids, attention_mask) \quad (1)$$

话题和情感标签由嵌入层编码,得到话题特征表示 e_{topic} 和情感特征表示 e_{sent} 。公式分别为:

$$e_{topic} = \text{Embed}_{topic}(topic_id) \quad (2)$$

$$e_{sent} = \text{Embed}_{sent}(sentiment_id) \quad (3)$$

式中: $topic_id$ 为离散的话题编号,包括地域、性别、少数民族群体、种族和无,用于捕获不同话题之间的潜在语义关联; $sentiment_id$ 采用三元情感体系,分别为负面、中性和正面,为模型注入宏观情感线索。

为使模型感知网暴演化节律,通过时间嵌入层生成小时级时间 e_{time} 。表示为:

$$e_{time} = \text{Embed}_{time}(time_index) \quad (4)$$

式中: $time_index$ 表示相对会话起点的小时编号,范围

为 0~23。

2.2.2 特征融合

协同演化模块将文本语义、话题、情感和时间信息放在同一个连续空间内。表示为:

$$c = [h_{BERT}; e_{topic}; e_{sent}; e_{time}] \quad (5)$$

信息经过两层全连接网络 f_{coevo} 处理,两个线性层均采用 RELU 激活函数,并采用 0.2 的 Dropout 防止过拟合。上述 4 种信息进行非线性交互得到 z_{coevo} 。表示为:

$$z_{coevo} = f_{coevo}(c) \quad (6)$$

2.2.3 分类决策

分类头 f_{abuse} 为两层全连接网络,映射出 2 维 logits。表示为:

$$o = f_{abuse}(z_{coevo}) \quad (7)$$

输出经 Softmax 转化为“网暴/非网暴”概率分布。训练采用交叉熵损失与真实标签对齐,以优化二分类目标。

2.3 预警机制

为实现早期预警并遏制网暴扩散,构建基于时间序列分析的预警机制。在二分类的基础上形成“演化分析—风险识别—可视化呈现”的闭环,为网暴治理提供即用型技术支撑。按照事件粒度统计各小时网暴占比并绘制其时间序列,基于上节分类结果计算小时级时间桶的网暴比例 a_h 。公式为:

$$a_h = \frac{\sum_{i: \Delta t_i = h} y_i}{\sum_{i: \Delta t_i = h} 1}, \quad h \in \{\Delta t_{\min}, \Delta t_{\min} + 1, \dots, \Delta t_{\max}\} \quad (8)$$

式中: Δt_i 表示第 i 条评论相对会话首条评论的整点小时差;时间范围 h 覆盖会话全程的小时级网暴密度。直接跳过无评论的小时桶。在进行可视化时,峰值对应网暴高发时段,可用于预警与重点分析;低谷则表明互动相对平和。

为捕捉网暴急增情况,采用一阶差分度量其变化速率,用于支撑趋势分析、峰值检测与预警工作。计算公式为:

$$d_h = a_h - a_{h-1} \quad (9)$$

式中: d_h 表示从 $h-1$ 小时到第 h 小时网暴密度的增减量。通过一阶差分序列 $\{d_h\}$ 可以快速定位趋势变化点以及加速上升、下降的时段。

为跨会话自适应识别网暴显著暴发情况,基于网暴比例序列 $\{a_h\}$ 的整体水平与波动范围构建动态阈值。首先计算各小时桶网暴比例的算术平均 μ 。公式为:

$$\mu = \text{mean}(a_h) \quad (10)$$

标准差 σ 反映了数据的波动幅度。计算公式为:

$$\sigma = \text{std}(a_h) \quad (11)$$

根据算术平均 μ 和标准差 σ 进行动态阈值构建。公式为:

$$\theta_{\text{out}} = \mu + 1.8\sigma \quad (12)$$

多会话调优结果表明,当 $k \in [1.6, 2.1]$ 时,最终预警结

果的变化均在 2% 以内,对结论影响可忽略,故折中取 $k = 1.8$ 。模型随事件的网暴密度与波动自适应设定阈值,既能抑制偶发小幅波动,又能捕捉真实暴发。预警分为暴发点、持续上升和单点上升 3 类。当 $a_h > \theta_{out}$ 时判定为暴发点,以确保仅对显著高于常态水平的突发网暴进行预警。当至少连续 3 h 的一阶差分 $d_h > 0.03$ 时判定为持续上升,用于捕捉稳步增势。在未检测到持续上升,且某小时段满足 $a_h > 0.06$ 时判定为单点上升,以捕捉无连续上升但出现急剧单小时跃升的场景。这 3 类互补的判定规则既能不漏突发暴发,又能兼顾对中度或短暂上涨的及时响应。

3 实验方法与结果分析

3.1 实验环境与参数设置

软件环境: Python 3.8、PyTorch 1.6 和 NumPy 1.19.1。硬件环境: RTX 2060。参数设置: 学习率为 3×10^{-5} , 训练轮数为 3, 批量大小为 16, 随机种子为 42。训练策略: 每个 epoch 结束即验证并保存; 采用五折交叉验证以检验稳健性; 模型训练过程中实时记录 loss 值、准确率、F1 分数和召回率。

3.2 数据集

采用 SCCD 和 CHNCI 两个基准数据集。其中, SCCD 来源于微博, 共有 38 999 条评论, 以会话为单位, 保留了上下文信息, 每条评论均附有相对原微博的发表时延, 这有利于识别网暴中的连续攻击行为^[11]; CHNCI 来源于抖音、微博、小红书、哔哩哔哩等, 共有 220 676 条评论, 涵盖了 91 个真实网络热点事件, 以事件为单位构建, 每条评论均有发表时间, 适合演化趋势分析与预警研究^[12]。为保证可比性与一致性, 在随机抽样下复现 SCCD、CHNCI 的实验设定, 即 CHNCI 分别取 600、1 000 和 1 600 条样本为训练集, 其余为测试集; SCCD 按照 7:3 的比例划分为训练集和测试集。

3.3 基线模型

在 SCCD 数据集上复现两类基线实验: 第一类是面向单条评论的欺凌实例识别任务 CL-CD, 用于判断每条评论是否包含欺凌内容^[11]。该类实验选取的基线模型为 Baidu-TC、COLDETECTOR^[13]、BERT^[5]和 RoBERTa^[14]; 第二类是面向完整会话的欺凌标签预测任务 SL-CD, 用于判断整个对话会话是否存在欺凌行为^[11]。该类实验选取的基线模型为 CNN^[15]、LSTM^[16]、BERT^[5]、RoBERTa^[14]以及在少样本 ($k=4$) 设置下的大语言模型 GPT-5 与 DeepseekV3.2^[17]。本文研究重点为“情感—话题”协同演化, 而非直接比较评论级与会话级粒度下的性能差异, 因此采用同一组实验结果, 分别与上述两种基线结果进行比较分析。

在 CHNCI 数据集上选取 9 种基线模型, 覆盖 3 种实验设定: 第一种是预训练模型微调, 在 600、1 000 和 1 600 条

标注样本上分别对 BERT^[5]与 HateBERT^[6]进行微调; 第二种是提示微调, 在 30、40 和 50 个少样本示例下分别使用 P-tuning^[18]、KPT^[19]、KPT++^[20]与 Conprompt^[21]方法进行提示微调; 第三种是大型语言模型评估, 分别在零样本、少样本 ($k=4$) 以及思维链设置下评估 Llama3-8B^[22]、Qwen-7B^[23]。

3.4 实验结果与分析

3.4.1 比较实验

各模型在 SCCD 数据集上的实验结果如表 2 所示。可以看出, 本文方法在网暴检测任务中表现优异, 尤其在准确率和 F1 分数上显著优于其他基线模型, 召回率虽略低于 GPT-5 模型, 但亦处于较高水平。

Table 2 Experimental results of different models on the SCCD dataset

表 2 各模型在 SCCD 数据集上的实验结果 (%)				
方法	模型	准确率	召回率	F1 分数
CL-CD	Baidu TC	62.33	21.92	76.64
	COLDETECTOR	64.22	39.07	79.21
	BERT	70.23	62.77	84.33
	RoBERTa	73.66	65.92	85.78
	CNN	86.56	63.82	72.66
SL-CD	LSTM	81.66	62.67	70.74
	BERT	89.55	84.82	84.32
	RoBERTa	91.98	83.66	86.34
	GPT-5	91.87	88.21	90.41
	DeepseekV3.2	91.55	85.32	89.17
本文方法	Co-Evolution Model	93.84	83.46	91.15

各模型在 CHNCI 数据集上的实验结果如表 3 所示。可以看出, 在 3 种数据集划分方式中, 本文方法均处于领先地位, 表明其在有标签数据和少量示范数据情况下均具有显著优势。

本文方法能精准检测网络语言暴力, 尤其是在情感强度高或话题可辨性强的语句上表现突出。其核心是将话题、情感与时间联合嵌入统一潜在空间, 并以非线性映射获得蕴含协同演化语义的高阶表征, 生成用于检测的判别性特征; 同时依靠情感与所属话题的精确对齐显著提升语义一致性与识别准确率。总之, 该方法通过统一建模“话题—情感—时间”并学习协同演化表征, 显著增强了对复杂语义的刻画与推断能力, 在两个数据集上均取得显著性能增益。

3.4.2 实例分析

为验证本文模型在情感和话题特征对齐方面的有效性, 从数据集中随机抽取若干样本进行定性分析。表 4 给出了不同模型在这些样本上的预测结果。在例一中, RoBERTa 可能受到关键词干扰, 误判为“网暴”, 而 GPT-5 与本文模型正确判为“网暴”; 在例二中, 其余两个模型均误判为“网暴”, 本文模型捕捉到整体正向情感及文本语义为非攻击性意图, 因而正确判为“非网暴”; 在例三中, 其余两个模型均误判为“非网暴”, 本文模型识别到“地域”这一话题及其隐含语义线索, 从而正确判为“网暴”。

Table 3 Experimental results of various models on the CHNCI dataset

表 3 各模型在 CHNCI 数据集上的实验结果 (%)

模型	600/30/0/4/Cot			1 000/40/0/4/Cot			1 600/50/0/4/Cot		
	准确率	召回率	F1 分数	准确率	召回率	F1 分数	准确率	召回率	F1 分数
Bert	63.10	70.89	66.77	67.12	72.52	69.81	70.12	75.03	72.49
HateBert	62.66	71.39	66.74	68.50	73.05	70.70	70.77	73.69	72.20
Conprompt	65.01	71.79	68.23	72.23	74.55	73.37	73.87	73.82	73.88
P-tuning	63.21	71.23	66.98	68.72	73.06	71.06	71.37	73.85	73.61
KPT	61.83	68.76	65.11	70.77	75.76	73.18	73.99	77.39	75.65
KPT++	69.68	75.10	72.29	70.72	76.95	73.24	72.65	71.21	74.58
Llama3-8B(Zero-shot)	73.06	74.94	73.99	73.06	74.94	73.99	73.06	74.94	73.99
Qwen-7B(Zero-shot)	75.86	76.72	77.74	75.86	76.72	77.74	75.86	76.72	77.74
Llama3-8B(Few-shot)	75.11	76.95	76.02	75.11	76.95	76.02	75.11	76.95	76.02
Qwen-7B(Few-shot)	77.12	80.16	78.61	77.12	80.16	78.61	77.12	80.16	78.61
Llama3-8B(CoT)	74.93	75.07	75.00	74.93	75.07	75.00	74.93	75.07	75.00
Qwen-7B(CoT)	77.03	80.01	78.49	77.03	80.01	78.49	77.03	80.01	78.49
Co-Evolution Model	78.62	78.47	78.53	82.49	82.19	82.34	84.64	83.12	84.57

Table 4 Analysis results of cyberbullying cases

表 4 网络语言暴力案例分析结果

句子	正确标签	RoBERTa	GPT-5	本文模型
很难不种草!!!	非网暴	网暴	非网暴	非网暴
这位太太您的作品很优秀…但如果太太是嗑 zmz 的话,也没必要发来恶心角色推	非网暴	网暴	网暴	非网暴
大量北京人当年主动给八国联军带路	网暴	非网暴	非网暴	网暴

为验证所提网络暴力预警机制的有效性,选取一个网络暴力事件,分析其网络暴力比例的时间演化情况,揭示其生命周期的潜伏期、扩散期、控制期与平稳期 4 个阶段,对预警机制的有效性进行系统评估^[24]。如图 2 所示,在潜伏期(0~10 h),网络暴力比例维持在 20%~30%,风险较低;在扩散期(10~35 h),比例持续上升,由 30% 增至 70% 以上,按预警规则捕获“持续上升”,并触发“单点上升”;在控制期(35~42 h),比例显著提升并接近 90% 峰值,超出动态阈值,被标记为“暴发点”,表明舆情与负面情绪达顶、治理难度加大;在平稳期(42 h 至事件结束),比例快速回落至小于 20%,关注与风险逐步消退。实验结果表明,所提预警机制能准确刻画阶段演化并提前识别高风险时点,具备有效干预价值。

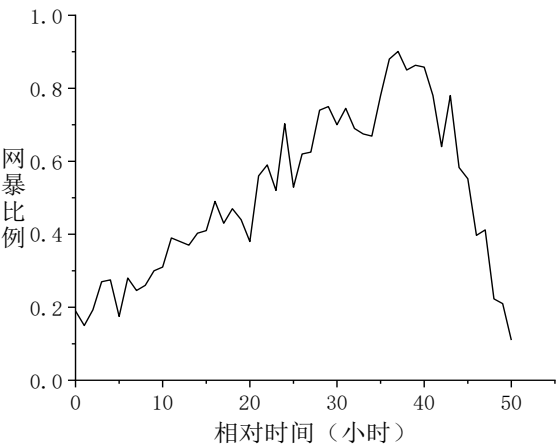


Fig. 2 Cyberbullying ratio over relative time
图 2 网络语言暴力比例随相对时间变化图

4 结语

本文通过引入时间、话题与情感维度的协同特征融合机制,构建了一种网络语言暴力动向检测与预警模型,其基于迁移学习完成三维信息的跨域映射与标签对齐/补全,在数据层面确保了协同建模的一致性与完整性,并为跨平台和跨语种的协同演化建模与预警扩展提供了可复用的基础。实验结果显示,所提模型在两个数据集上表现优异,特别是在标签稀疏的情况下表现出更强的鲁棒性,这是由于其能精准捕获协同演化特征,增强了演化动向检测能力;所设计的预警机制通过严格定义暴发点、持续上升和单点上升三类预警类型,实现了对演化生命周期的全面覆盖,具有良好的灵敏性。该模型具备广泛的跨领域应用潜力,可为舆情监测等提供新手段。未来可进一步考虑引入图像、视频、音频等多模态信息融合分析网络暴力行为,以挖掘更多隐含的网络暴力特征行为;还可构建同时具备会话粒度与时间戳的英文网暴数据集,以用于跨语言评测。

参考文献:

[1] PATCHIN J W, HINDUJA S. Measuring cyberbullying: implications for research[J]. Aggression and Violent Behavior, 2015, 23: 69-74.
[2] ELSAFOURY F, KATSIKIANNIS S, PERVEZ Z, et al. When the timeline meets the pipeline: a survey on automated cyberbullying detection [J]. IEEE Access, 2021, 9: 103541-103563.
[3] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020,

- 33: 1877–1901.
- [4] PARK J H, FUNG P. One-step and two-step classification for abusive language detection on Twitter [DB/OL]. <https://arxiv.org/abs/1706.01206>.
- [5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171–4186.
- [6] CASELLI T, BASILE V, MITROVIĆ J, et al. HateBERT: retraining BERT for abusive language detection in English [DB/OL]. <https://arxiv.org/abs/2010.12472>.
- [7] ZHANG Y, ZHONG T, YI T, et al. Domain-enhanced prompt learning for Chinese implicit hate speech detection [J]. IEEE Access, 2024, 12: 13773–13782.
- [8] GHORBANPOUR F, DEMENTIEVA D, FRASER A. Can prompting LLMs unlock hate speech detection across languages? a zero-shot and few-shot study [DB/OL]. <https://arxiv.org/abs/2505.06149>.
- [9] MIN C, LIN H, LI X, et al. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective [J]. Information Fusion, 2023, 96: 214–223.
- [10] SHEN C, SONG Z, HE P, et al. Online rumors during the COVID-19 pandemic: co-evolution of themes and emotions [J]. Frontiers in Public Health, 2024, 12: 1375731.
- [11] YANG Q, CHEN Y, XU Z, et al. SCCD: a session-based dataset for Chinese cyberbullying detection [C]//Proceedings of the 31st International Conference on Computational Linguistics, 2025: 9533–9545.
- [12] ZHU Y, ZOU X, WU X. Chinese cyberbullying detection: dataset, method, and validation [DB/OL]. <https://arxiv.org/abs/2505.20654>.
- [13] DENG J, ZHOU J, SUN H, et al. COLD: a benchmark for Chinese offensive language detection [DB/OL]. <https://arxiv.org/abs/2201.06025>.
- [14] LIU Z, LIN W, SHI Y, et al. A robustly optimized BERT pre-training approach with post-training [C]//China National Conference on Chinese Computational Linguistics, 2021: 471–484.
- [15] KIM Y. Convolutional neural networks for sentence classification [DB/OL]. <https://arxiv.org/abs/1408.5882>.
- [16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735–1780.
- [17] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report [DB/OL]. <https://arxiv.org/abs/2412.19437>.
- [18] LIU X, ZHENG Y, DU Z, et al. GPT understands, too [J]. AI Open, 2024, 5: 208–215.
- [19] HU S, DING N, WANG H, et al. Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification [DB/OL]. <https://arxiv.org/abs/2108.02035>.
- [20] NI S, KAO H Y. KPT++: refined knowledgeable prompt tuning for few-shot text classification [J]. Knowledge-Based Systems, 2023, 274: 110647.
- [21] KIM Y, PARK S, NAMGOONG Y, et al. ConPrompt: pre-training a language model with machine-generated data for implicit hate speech detection [C]//Findings of the Association for Computational Linguistics: EMNLP 2023, 2023: 10964–10980.
- [22] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models [DB/OL]. <https://arxiv.org/abs/2407.21783>.
- [23] BAI J, BAI S, CHU Y, et al. Qwen technical report [DB/OL]. <https://arxiv.org/abs/2505.09388>.
- [24] LIU Z, WU X. Structural analysis of the evolution mechanism of online public opinion and its development stages based on machine learning and social network analysis [J]. International Journal of Computational Intelligence Systems, 2023, 16(1): 99.

(责任编辑:尹晨茹)