

基于情感与主题的网络仇恨言论检测研究

陈卓^{1,2}, 张茂元^{2,3}, 郑萌⁴

(1. 华中师范大学人工智能与智慧学习湖北省重点实验室; 2. 华中师范大学计算机学院;
3. 国家语言资源监测与研究网络媒体中心; 4. 华中师范大学人工智能教育学部, 湖北 武汉 430079)

摘要: 随着社交媒体的发展, 隐性仇恨言论因其间接表达和模糊性成为检测难点。传统方法虽在显性仇恨言论检测中表现良好, 但在识别隐含歧视和敌意方面能力有限, 尤其在主题与文本关系建模上存在不足。为此, 提出一种基于方面情感特征提取的师生模型(ESTD)。该模型采用T5作为教师模型, RoBERTa作为学生模型, 通过知识蒸馏优化软标签分布, 提升隐性仇恨言论的检测能力。学生模型结合主题信息提取方面情感分析(ABSA)特征, 并使用DF-Net双通道结构融合全局语义与局部情感信息, 通过跨通道注意力机制提高情感特征提取精度。实验结果显示, ESTD在多个数据集上的准确率和F1得分均优于现有方法, 特别是在处理隐性仇恨言论时表现出更高的鲁棒性和精度。该方法具备良好的扩展性, 适用于舆情监测等领域, 为复杂语义和情感建模提供了新思路。

关键词: 隐性仇恨言论; 方面情感; 师生框架; 知识蒸馏; 文本主题

DOI: 10.11907/jjdk.251026

扫描二维码阅读全文:

中图分类号: TP391.1

文献标识码: A

文章编号: 1672-7800(2025)012-0080-05

Research on Online Hate Speech Detection Based on Emotion and Topic

CHEN Zhuo^{1,2}, ZHANG Maoyuan^{2,3}, ZHENG Meng⁴

(1. Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University; 2. School of Computer Science, Central China Normal University; 3. National Language Resources Monitoring and Research Network Media Center; 4. Faculty of Artificial Intelligence Education, Central China Normal University, Wuhan 430079, China)

Abstract: With the rise of social media, detecting implicit hate speech has become challenging due to its indirect and ambiguous nature. Traditional methods excel at detecting explicit hate speech but struggle with subtle discrimination, especially in modeling the relationship between topics and text. This paper proposes a teacher-student model (ESTD) that uses T5 as the teacher and RoBERTa as the student, applying knowledge distillation to improve implicit hate speech detection. The student model extracts aspect-based sentiment analysis (ABSA) features with topic information, using DF-Net's dual-channel structure to combine global semantics and local sentiment, enhancing feature extraction through cross-channel attention. Experimental results show ESTD outperforms existing methods, with higher robustness and precision in detecting implicit hate speech. This approach is scalable and can be applied to public opinion monitoring, offering new insights into complex semantic and sentiment modeling.

Key Words: latent hate speech; aspect-based sentiment; teacher-student framework; knowledge distillation; text topic

0 引言

近年来, 社交媒体快速发展, 其不仅促进了公众参与讨论, 也成为仇恨言论传播的重要渠道^[1]。相比显性仇恨言论, 隐性仇恨言论因其间接表达或文化背景的模糊

性, 使得其难以被传统基于规则或显性特征的检测方法捕捉^[2]。如表1例句, 虽然表面没有明显的攻击性词汇, 却通过隐晦的方式传递排斥和敌意, 属于典型的隐性仇恨言论。这类言论长期未被检测可能对受害者造成严重心理伤害, 并加剧社区对立^[3]。

现有仇恨言论检测方法(如基于规则、词汇总表^[4]和深

收稿日期: 2025-01-22

基金项目: 国家语委“十四五”重点科研项目(ZDI145-94)

作者简介: 陈卓(2001-), 女, CCF学生会会员, 华中师范大学计算机学院硕士研究生, 研究方向为情感分析; 张茂元(1975-), 男, 博士, 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、人工智能与知识获取; 郑萌(1997-), 女, 华中师范大学人工智能教育学部博士研究生, 研究方向为情感分析。本文通讯作者: 张茂元。

度学习模型如 RoBERTa^[5]、T5^[6]) 在显性仇恨言论的检测中表现较好,但在处理隐藏性仇恨言论时效果有限。其原因在于这些方法通常依赖显性特征(如直接侮辱性词汇),难以捕捉隐含的歧视或敌意。如表 1 所示,隐藏性仇恨言论往往缺乏明显的攻击性词汇,但通过隐晦方式传递排斥情感,难以被现有模型准确识别。

此外,传统方法对主题和文本之间的复杂关系缺乏深入建模,仅简单拼接主题与文本内容进行训练。这种局限性使得模型难以挖掘主题中蕴含的深层信息,也无法充分利用情感信息指导上下文分析,导致在应对复杂情感或隐性攻击性文本时表现不佳,从而限制了其检测能力。

Table 1 Example sentences

表 1 案例句

句子	RoBERTa	T5	正确标签
Let those people go back to where they belong. (Target: Minorities)	正常	正常	仇恨
Because it's the white thing to do. (Target: Caucasian)	正常	正常	仇恨

为解决隐藏性仇恨言论难以识别和主题与文本内涵挖掘不足的问题,本文提出一种基于情感和主题的仇恨言论检测方法。该方法通过主题信息提取文本的方面词和情感特征,结合情感特征辅助仇恨言论检测,提升了模型识别能力。本文主要贡献包括:①提出基于方面情感特征提取的师生模型 ESTD,通过知识蒸馏优化软标签分布,有效提升了隐藏性仇恨言论的检测能力^[7];②提出 DF-Net 双通道融合方法,结合全局语义和局部情感特征,并通过跨通道注意力机制提升了模型在隐藏性仇恨言论中的情感特征提取精度。

实验结果表明,本文模型在仇恨言论检测任务中表现优异,特别是在处理隐藏性仇恨言论时,具有更高的准确性和鲁棒性。该模型具备良好的灵活性和适应性,适用于舆情监测和负面情绪识别等跨领域应用,提供了处理复杂语义和情感关系的新思路。

1 相关工作

1.1 仇恨言论检测

随着互联网的普及,仇恨言论的传播问题愈发严重,对社会和个体的负面影响引发了广泛关注。早期基于规则的方法主要通过预定义词汇表和语法规则检测显性仇恨言论,虽在直接攻击性内容识别上表现良好,但在处理隐性表达方面力不从心^[4]。这类言论往往通过隐喻、文化背景或间接暗示表达敌意,使得规则匹配方法难以奏效。

近年来,预训练语言模型的发展显著提升了仇恨言论检测的语义理解能力。这些模型通过大规模语料预训练,能够有效捕捉上下文信息^[8-9]。然而,对于隐性仇恨言论,传统模型依然面临困难,如未能深入挖掘情感与主题的深

层关联,难以识别潜在敌意和歧视意图。此外,部分生成式方法虽然在提升模型解释性方面取得进展,但其生成误差常导致不准确的判断,影响可靠性^[10]。

总之,现有仇恨言论检测技术在识别隐性表达、提升情感理解和模型解释性等方面仍存在不足。如何通过更深入的情感和主题建模以增强隐性仇恨言论检测,成为这一领域的关键研究方向。

1.2 方面情感分析

方面情感分析 (Aspect-Based Sentiment Analysis, AB-SA) 是一种细粒度情感分析任务,旨在预测文本中特定方面的情感极性。传统方法多依赖手动规则,难以适应复杂多样的语言表达。近年来,深度学习方法通过注意力机制和图卷积网络 (Graph Convolutional Networks, GCN) 等技术,显著提升了 AB-SA 的性能^[11]。其中,注意力机制可动态聚焦不同文本片段,而基于句法树的 GCN 能更好地建模方面词与上下文之间的联系^[12-13]。

ABSA 作为仇恨言论检测的辅助特征,能帮助模型捕捉特定主题的情感态度。例如,种族、性别或宗教等敏感话题中的隐性仇恨言论,可能隐藏敌意或歧视。引入 AB-SA 可提高情感关联捕捉精度,提升检测性能,同时增强模型可解释性。

2 模型设计

如图 1 所示,本文模型架构由教师模型和学生模型两部分组成。教师模型通过在大规模语料库上进行预训练,能够获取深层次的上下文语义表示,并生成高质量的预测输出,作为知识蒸馏的指导信号。具体而言,教师模型利用其强大的语义理解能力,为学生模型提供丰富的软标签,这些软标签包含了教师模型对输入文本的细腻理解和预测信息。

学生模型在学习教师模型传递知识的基础上,进一步整合了基于方面的情感特征。通过引入 ABSA 模块,学生模型能够识别并提取文本中与特定主题相关的细粒度情感特征。这些方面情感特征,使得学生模型不仅具备良好的语义理解能力,还能精准捕捉文本中潜在的情感态度和攻击性意图。

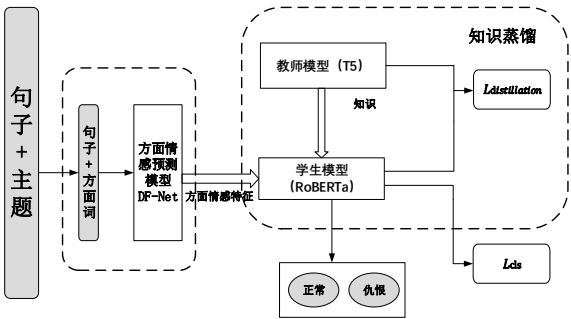


Fig. 1 ESTD model structure

图 1 ESTD 模型结构

2.1 主题建模与方面词提取

假设输入文本为 $T = \{t1, t2, \dots, tn\}$, 其中 $t1, t2, \dots, tn$ 为文本中的第 1, 2, ..., n 个词。通过主题建模方法, 提取与特定主题相关的方面词集合 $A = \{a1, a2, \dots, ak\}$, 其中 $a1, a2, \dots, ak$ 是从文本中抽取出的与主题密切相关的词。这些方面词为后续情感特征提取提供了指导性线索。

2.2 方面情感特征提取

为了充分利用全局语义和局部方面情感信息, 本文设计了一个动态跨通道注意力融合模型 (DF-Net)。该模型主要由 3 个部分组成: 全局语义通道、局部方面情感通道和跨通道注意力融合。DF-Net 模型架构如图 2 所示, 具体步骤如下:

(1) 全局与局部特征生成。通过 RoBERTa 模型^[5]提取全局语义特征, 得到全局表示 G' , 如式(1)所示。

$$G' = W_g \cdot \text{RoBERTa}(S, a) \quad (1)$$

其中, S 是输入句子, $a \in A$, A 表示方面词集合, W_g 是全局变换矩阵。

局部方面情感特征 F' 通过 BiLSTM 和 TextCNN 通道生成, 如式(2)所示。

$$F' = W_l \cdot \text{TextCNN}(\text{BiLSTM}(C)) \quad (2)$$

其中, C 是包含方面词集合 A 及其上下文的局部窗口, W_l 是局部特征的线性变换矩阵。

(2) 动态特征融合。

全局特征 G' 和局部特征 F' 通过跨通道注意力机制动态融合, 得到最终方面情感特征表示 F_{aspect} , 如式(3)所示。

$$F_{\text{aspect}} = \alpha \cdot G' + (1 - \alpha) \cdot F' \quad (3)$$

其中, 注意力权重 α 通过 Sigmoid 激活函数计算, 如式(4)所示。

$$\alpha = \text{sigmoid}(W_a(G' + F')) \quad (4)$$

W_a 是注意力权重计算的线性变换矩阵, 动态控制全局和局部特征的融合比例。

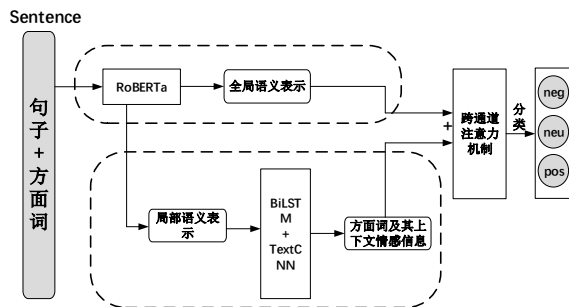


Fig. 2 DF-Net model structure

图2 DF-Net模型结构

2.3 教师模型

教师模型采用 T5 模型, 其编码器生成高质量的上下文语义表示^[6]。具体而言, 输入文本 $T = \{t1, t2, \dots, tn\}$ 通过 T5 编码器转化为隐层特征 h , 随后经过全连接层生成 logits 输出, 并通过 Softmax 归一化得到类别预测概率以用于提供高置信度的监督信号 $\hat{y}_t$, 如式(5)所示。

$$\hat{y}_t = \text{Softmax}(W \cdot h + b) \quad (5)$$

其中, W 是全连接层的权重矩阵, b 是偏置项, h 是 T5 编码器生成的隐层特征表示, 最终 $\hat{y}_t$ 输出作为高置信度的监督信号, 传递给学生模型加以训练。

2.4 学生模型

学生模型使用模型 RoBERTa, 并结合方面情感特征和教师模型输出的高置信度监督信号进行训练^[5]。具体过程如下:

(1) 特征选择与拼接。根据模型输出的类别置信度差异 $\text{prob_diff} = |p_1 - p_0|$ 判断是否需要补充方面情感特征, 如式(6)所示。

$$F_{\text{input}} = \begin{cases} G & \text{if prob_diff} \geq \varepsilon \\ \text{Concat}(F_{\text{aspect}}, G) & \text{if prob_diff} < \varepsilon \end{cases} \quad (6)$$

其中, ε 是置信度差异的阈值, Concat 表示特征拼接操作。

(2) 分类任务。将最终输入特征 F_{input} 输入到全连接层进行分类任务, 如式(7)所示。

$$\hat{y}_s = \text{Softmax}(W_c \cdot F_{\text{input}} + b_c) \quad (7)$$

其中, W_c 和 b_c 是分类器的权重和偏置, $\hat{y}_s$ 是学生模型的预测概率分布。

(3) 损失函数设计。学生模型的损失函数由分类损失和知识蒸馏损失两部分组成。其中, 分类损失结合真实标签 y_i 和预测值 $\hat{y}_s$, 使用二值交叉熵计算损失, 如式(8)所示。

$$L_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N [y_i * \log(\hat{y}_s) + (1 - y_i) * \log(1 - \hat{y}_s)] \quad (8)$$

知识蒸馏损失通过 KL 散度最小化教师模型预测值 $\hat{y}_t$ 与学生模型预测值 $\hat{y}_s$ 的差异, 如式(9)所示。

$$L_{\text{distillation}} = D_{\text{KL}}(\hat{y}_t || \hat{y}_s) \quad (9)$$

最终总损失是分类损失和蒸馏损失的加权和, 如式(10)所示。

$$L = \alpha L_{\text{cls}} + \beta L_{\text{distillation}} \quad (10)$$

其中, α 和 β 是损失的加权系数, 用于平衡分类损失和蒸馏损失的贡献。

3 数据集与实验设置

3.1 数据集

本文采用 3 个基准数据集——IHC、SBIC 和 ToxiGen, 对所提模型进行评估。IHC (Implicit Hate Speech Corpus) 来源于 Twitter, 包含 19 112 条推文, 主要用于检测隐含的仇恨言论^[14]。SBIC (Social Bias Inference Corpus) 是一个多层次的社会偏见和刻板印象分类数据集, 包含 150 000 条经过结构化标注的社交媒体帖子, 涵盖超过 34 000 条针对不同人口群体的偏见推断内容^[15]。ToxiGen 是一个大规模的机器生成数据集, 包含 274 186 条关于 13 个少数群体的

有毒和良性陈述^[16]。

3.2 实验设置

在训练过程中采用 AdamW 优化器,学生模型学习率设置为 $2e-5$,批量大小为 16,损失函数采用交叉熵损失 (Cross-Entropy Loss) 进行训练。在知识蒸馏阶段,引入了 Kullback-Leibler (KL) 散度,用于度量教师模型与学生模型输出之间的差异^[17]。模型训练总共进行了 10 个周期 (epochs),每个周期使用不同的学习率调度策略 (如线性衰减) 以避免过拟合。

4 实验结果分析

4.1 基线模型

实验中,选用了多种现代预训练语言模型作为基线模型,包括 BERT^[18]、HateBERT^[19]、ELECTRA^[20]、GPT-2^[21]、T5^[7] 和 TOXCL^[22],这些模型在自然语言处理任务中表现优异。BERT 基于 Transformer 架构,擅长双向上下文编码; HateBERT 是专为仇恨言论检测微调的 BERT 变体; ELECTRA 通过对抗训练提升了效率和性能; GPT-2 具备文本生成与理解能力; T5 采用统一的文本生成框架,适用于多任务处理; TOXCL 专注于仇恨言论检测与解释。这些基线模型为全面评估本文方法在隐性仇恨言论检测中的效果提供了基础。

4.2 结果分析

实验结果如表 2 所示。可以看出,ESTD 模型在仇恨言论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,这证明 ESTD 能够精准检测隐性仇恨言论,特别是那些表面中立但实际包含攻击性或歧视性意图的句子。ESTD 的核心优势在于有效应用主题信息和方面情感特征,通过主题与方面词的关联,精准捕捉文本的核心含义与情感特征。这种结合使模型能够识别通过隐性交织传递攻击性意图的复杂仇恨言论。

总而言之,ESTD 通过整合主题特征与方面情感特征,提升了对复杂语义的理解,在 IHC 和 SBIC 数据集上取得了显著性能提升,为隐性仇恨言论检测提供了新的思路。

Table 2 The overall performance of different models on the two datasets

表 2 不同模型在两个数据集上的整体效果						
Model	IHC		SBIC		ToxiGen	
	Acc	F1	Acc	F1	Acc	F1
BERT	76.98	76.16	87.83	87.73	-	75.30
HateBERT	76.67	75.83	87.21	87.30	-	73.10
ELECTRA	77.46*	75.87*	87.15*	87.03*	77.50*	77.33*
GPT-2	77.57	76.36	73.62	53.09	77.78	76.67
T5	70.71	62.95	87.55	87.48	-	79.50
TOXCL	78.29*	76.69*	87.64*	87.55*	81.55*	80.45*
ESTD	80.91	78.16	89.98	89.96	83.75	82.78

注:“*”表示作者根据论文中的方法而实现的模型效果,其他数据从已录用论文检索得出。

4.3 消融实验

分别对 ESTD 中的情感特征 (w/o SF)、方面情感特征 (w/o ASF) 和知识蒸馏 (w/o KL) 都进行消融实验。在每次消融中,仅去除一个模块并保持其余部分不变,以观察各模块在不同数据集 (IHC、SBIC、ToxiGen) 上的影响。表 3 展示了不同消融模型的准确率 (ACC) 和 F1 得分。

Table 3 The accuracy (ACC) and F1 scores of the ablation results on the two datasets

Model	IHC		SBIC		ToxiGen	
	Acc	F1	Acc	F1	Acc	F1
ESTD	80.91	78.16	89.98	89.96	83.75	82.78
w/o ASF	78.90	77.20	88.69	88.30	81.67	80.90
w/o SF	78.23	76.54	87.45	87.50	82.56	81.50
w/o KL	79.34	77.59	88.60	88.65	83.10	82.02

通过对 ESTD 模型进行消融实验,验证了情感特征 (Sentiment Feature, SF)、方面情感特征 (Aspect-based Sentiment Feature, ASF) 和知识蒸馏 (Knowledge Distillation, KL) 在隐藏性有毒言论检测中的重要性。实验结果表明,去除这 3 个模块中的任何一个都会导致模型在准确率和 F1 得分上的显著下降,尤其是 F1 得分的下降更为明显。具体而言,情感特征 (SF) 帮助模型捕捉文本中的情感倾向,方面情感特征 (ASF) 则提供了更精细的情感与主题的联系,从而提升了模型对潜在仇恨言论的识别能力。与此同时,知识蒸馏 (KL) 通过优化软标签分布,促进了教师模型和学生模型的协同学习,进一步提升了检测性能。

这些结果进一步证明了基于情感和主题联合建模及知识蒸馏对提高隐藏性仇恨言论检测精度的重要作用,为其未来在其他领域的应用和扩展提供了理论支持。

4.4 案例分析

为了更好地理解 ESTD 模型中情感特征 (SF) 和方面情感特征 (ASF) 的重要性,本文通过两个具体案例进行分析。表 4 中,符号①表示模型 BERT^[18]、符号②表示模型 TOXCL^[22]、符号③表示本文模型 ESTD。句子“Let those people go back to where they belong”和“Because it’s the white thing to do”均为隐藏性仇恨言论,但没有直接使用明显的攻击性语言。BERT 和 TOXCL 都未能准确捕捉到这种言论的仇恨倾向,而 ESTD 通过方面情感特征的加强,成功识别出其中的仇恨倾向。案例结果表明,ESTD 在仇恨言论检测中明显优于 BERT 和 TOXCL,尤其在识别隐藏性仇恨言论时,方面情感特征显著提高了模型准确性和敏感度,证明其对复杂情感分析任务具有重要意义。

4.5 偏见评估

由于仇恨言论检测涉及偏见问题,本文使用 IHC 和 SBIC 数据集对模型进行了偏见评估。将数据分为两类:目标为 none 和目标非 none 的数据,并计算这两类的准确率 (ACC) 和 F1 值,以评估模型是否存在偏见。

从表 5 中处理前的结果来看,模型在这两个数据集上

Table 4 The prediction results of some representative methods for the same sentence

表 4 一些代表性方法对同一句子的预测结果

句子	①	②	③	正确标签
Let those people go back to where they belong. (Target: Minorities)	×	×	√	仇恨
Because it's the white thing to do. (Target: Caucasian)	×	×	√	仇恨

的偏见问题较轻,表现差异不大,因而可以通过简单调整加以解决。本文采用了阈值优化方法,对于目标为 none 的类别降低阈值,其他类别提高阈值。调整后的结果(见表 5 后两行)表明,偏见问题得到了有效缓解,准确率和 F1 值在不同类别间更加平衡。

Table 5 Bias evaluation results

表 5 偏见评估结果

数据集	Targer:none		Target:other		All	
	Acc	F1	F1	F1	Acc	F1
IHC(处理前)	77.94	76.23	84.70	79.52	80.91	78.16
SBIC(处理前)	89.42	87.41	90.57	90.26	89.98	89.96
IHC(处理后)	79.74	77.90	81.01	78.20	80.91	78.16
SBIC(处理后)	89.92	88.59	89.90	89.87	89.98	89.96

5 结语

本文提出了 ESTD 模型,结合 T5 作为教师模型和 RoBERTa 作为学生模型,通过知识蒸馏优化软标签,显著提升了对隐性仇恨言论的检测能力。同时,设计了 DF-Net 双通道融合框架,整合全局语义特征与局部情感特征,并利用跨通道注意力机制实现动态融合,有效提高了情感提取精度。实验结果显示,ESTD 在多个数据集上表现优异,特别是在处理复杂语义结构和隐性仇恨言论时,比传统模型展现出更高的准确性和鲁棒性。其关键在于该模型核心方法中的方面情感特征提取帮助模型识别文本中的情感态度和细微变化,增强了对潜在仇恨言论的识别能力。通过挖掘情感与主题的深层关系,模型能更精准地捕捉隐含的敌意或歧视,从而提高检测准确度。得益于其核心方法具有通用性,也即情感与主题建模能够适应多种类型的分类任务,ESTD 不仅可用于仇恨言论检测,还具备广泛的跨领域应用潜力,为舆情监测等其他领域提供了新思路。此外,本文还未深入探讨模型的实时应用能力,未来研究将着重优化模型计算效率,提升其在实时检测中的性能,以适应实际部署需求。

参考文献:

- [1] YU X, BLANCO E, HONG L Z. Hate speech and counter speech detection: conversational context does matter [C]//North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5918–5930.
- [2] ELSHERIEF M, ZIEMS C, MUCHLINSKI D, et al. Latent hatred: a benchmark for understanding implicit hate speech [C]//Empirical Methods in

Natural Language Processing, 2021: 345–363.

- [3] FARRELL T, FERNANDEZ M, NOVOTNY J, et al. Exploring misogyny across the manosphere in reddit [C]//10th ACM Conference on Web Science, 2019: 87–96.
- [4] WIEGAND M, RUPPENHOFER J, SCHMIDT A, et al. Inducing a lexicon of abusive words —— a feature-based approach [C]//NAACL Conference on Human Language Technologies, 2018: 1046–1056.
- [5] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pre-training approach [DB/OL]. <https://arxiv.org/abs/1907.11692>.
- [6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(1): 1–67.
- [7] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. <https://arxiv.org/abs/1503.02531>.
- [8] WANG H H, ZHU Y. Fusion mechanism of irony of hate speech detection [J]. Computer Applications, 2024, 44(4): 1065–1071.
王海涵, 朱焱. 融合反讽机制的攻击性言论检测 [J]. 计算机应用, 2024, 44(4): 1065–1071.
- [9] KENNEDY B, JIN X S, MOSTAFAZADEH D A, et al. Contextualizing hate speech classifiers with post-hoc explanation [C]//58th Annual Meeting of the Association for Computational Linguistics, 2020: 5435–5442.
- [10] HUANG F, KWAK H, AN J S. Chain of explanation: new prompting method to generate higher quality natural language explanation for implicit hate speech [DB/OL]. <https://arxiv.org/abs/2209.04889>.
- [11] ZHENG C, YANG N. Aspect-level sentiment analysis based on syntactic, semantic and emotional know-ledge [J]. Computer Science, 2025, 52(7): 218–225.
郑诚, 杨楠. 基于句法、语义和情感知识的方面级情感分析 [J]. 计算机科学, 2025, 52(7): 218–225.
- [12] WANG Y Q, HUANG M L, ZHU X Y, et al. Attention-based LSTM for aspect-level sentiment classification [C]//Conference on Empirical Methods in Natural Language Processing, 2016: 606–615.
- [13] LIANG B, SU H, GUI L, et al. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks [J]. Knowledge-Based Systems, 2021, 235: 10–43.
- [14] ELSHERIEF M, ZIEMS C, MUCHLINSKI D, et al. Latent hatred: a benchmark for understanding implicit hate speech [C]//Conference on Empirical Methods in Natural Language Processing, 2021: 345–363.
- [15] SAP M, GABRIEL S, QIN L, et al. Social bias frames: reasoning about social and power implications of language [C]//58th Annual Meeting of the Association for Computational Linguistics, 2020: 5477–5490.
- [16] HARTVIGSEN T, GABRIEL S, PALANGI H, et al. ToxiGen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 3309–3326.
- [17] CSISZÁR I. I-divergence geometry of probability distributions and minimization problems [J]. The Annals of Probability, 1975, 76(6): 146–158.
- [18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171–4186.
- [19] CASELLI T, BASILE V, MITROVIĆ J, et al. HateBERT: retraining BERT for abusive language detection in English [C]//5th Workshop on Online Abuse and Harms, 2021: 17–25.
- [20] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators [DB/OL]. <https://arxiv.org/abs/2003.10555>.
- [21] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [DB/OL]. <https://arxiv.org/abs/1901.05027>.
- [22] HOANG N, DO X L, DO D A, et al. ToXCL: a unified framework for toxic speech detection and explanation [C]//NAACL Conference: Human Language Technologies, 2024: 6460–6472.

(责任编辑:孙娟)