文章编号:1671-9352(2026)03-0044-10 DOI:10.6040/j.issn.1671-9352.1.2024.044

基于预训练模型的仇恨言论检测

林原[1],张亚[1],于蒙[1],许侃[2*],林鸿飞[2]
(1.大连理工大学公共管理学院, 辽宁 大连 116024; 2.大连理工大学电子信息与电气工程学部, 辽宁 大连 116024)

摘要:为准确检测和识别仇恨言论,通过微调大语言模型对数据集样本进行扩充与平衡,并基于预训练模型 RoBERTa 构建 RoBERTa-Attention-GRU-TextCNN 模型,将深度学习强大的特征捕获和提取能力应用到文本序列数据的分析、挖掘中。首先通过 RoBERTa 模型对文本数据进行特征提取;然后利用自注意机制获取单词间的依赖关系;最后将获取到的特征矩阵输入到 GRU-TextCNN 层中以捕捉更深层次的语义信息和局部特征。使用 TweetEval 提供的 2 个公开的数据集来评估模型效果,实验结果表明,该模型相较于传统的仇恨言论检测模型具有更好的检测效果。

关键词:大语言模型;仇恨检测;RoBERTa;预训练模型;RoBERTa-Attention-GRU-TextCNN
中图分类号:TP391 **文献标志码**:A
引用格式:林原,张亚,于蒙,等. 基于预训练模型的仇恨言论检测[J]. 山东大学学报(理学版),2026,61(3):44-53.

Hate speech detection based on pre-trained models

LIN Yuan[1], ZHANG Ya[1], YU Meng[1], XU Kan[2*], LIN Hongfei[2]
(1. School of Public Administration and Policy, Dalian University of Technology, Dalian 116024, Liaoning, China; 2. Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, Liaoning, China)

Abstract: To accurately detect and identify hate speech, the dataset samples are expanded and balanced by fine-tuning the large language model. The RoBERTa-Attention-GRU-TextCNN model is constructed based on the pre-training model RoBERTa, leveraging the powerful feature capture and extraction capabilities of deep learning for the analysis and mining of text sequence data. Firstly, the RoBERTa model is used to extract features from the text data; then, the self-attention mechanism is used to obtain the dependencies between words; finally, the acquired feature matrix is input into the GRU-TextCNN layer to capture deeper semantic information and local features. Two publicly available datasets provided by TweetEval are used to evaluate the model effect, and the experimental results show that the model has a better detection effect compared to the traditional hate speech detection model.

Key words: large language model; hate detection; RoBERTa; pre-trained model; RoBERTa-Attention-GRU-TextCNN

0 引言

随着社交媒体的普及和信息传播技术的快速发展,人们可以在 Facebook、Twitter、微博、小红书等社交平台较自由地表达自己的观点和看法。然而,由于这些接触是非现实的,情绪的表达往往比日常生活中夸张、激烈,因此导致大量恶性的仇恨言论产生。在 2021 年 7 月的联合国大会上,鉴于全球范围内“仇恨言论呈迅猛增长与广泛蔓延的态势”,与会各方对此表达了深切忧虑,并一致通过旨在“增进宗教与文化间的对话与和谐,有效遏制仇恨言论蔓延”的决议[1]。

社交网络平台上存在的大量仇恨言论多为文本信息,能够通过自动检测和识别技术进行检测识别,一定程度上可以防止仇恨言论进一步传播和蔓延,危害网络空间的秩序和安全。在文本情感分析领域,仇恨言论

收稿日期:2024-10-09; **网络出版时间**:2025-04-09
基金项目:国家自然科学基金资助项目(61976036); 国家社会科学基金资助项目(20BTQ074)
第一作者:林原(1983—),男,副教授,博士,研究方向为信息检索、排序学习、数字治理. E-mail:zhlin@dlut.edu.cn
*通信作者:许侃(1981—),男,高级工程师,博士,研究方向为信息检索. E-mail:xukan@dlut.edu.cn

的检测与识别占据重要地位。这一任务可通过 2 种方法实现,即基于机器学习技术和深度学习技术的情感分析策略。前者的核心在于通过对已知情感标签的文本集合进行有监督的训练,构建出一个能够准确区分文本情感极性的分类模型。而后者要先将文本转化为词向量等数值化特征,然后将其输入到神经网络模型中,利用神经网络的强大能力捕获文本中潜在的细微特征,最终实现情感的精确分类。但传统的文本向量化方法,例如 One-hot 编码和 TF-IDF,存在数据稀疏性和无法有效捕捉语义信息的问题,生成的高维稀疏向量增加模型训练的内存需求的同时也限制算法对文本内在语义的深入理解。而且这些静态的词向量表示无法充分解决一词多义的问题,还限制模型在复杂语境下的表达能力。预训练模型可以有效解决数据稀疏性、语义信息捕捉不足以及一词多义处理困难等问题,从而增强模型对文本细微情感差异的识别能力。

基于上述背景,本研究聚焦于仇恨言论的有效检测与识别,旨在构建一种基于预训练模型的仇恨言论检测模型,将预训练语言模型与深度学习模型进行深度融合,利用其在大规模语料上学习到的丰富语义信息和上下文相关性,为文本提供更为丰富和精细的向量表示,同时将深度学习强大的特征捕获和提取能力应用到文本序列数据的分析、挖掘中,有效提高仇恨言论检测的准确度。

1　相关工作与技术

1.1　仇恨言论检测概述

仇恨言论检测是一项重要研究。它通过计算机技术对社交网络平台上的仇恨语句进行自动识别和分析,不但可以营造健康积极的网络环境,而且有利于社会和谐稳定及民主价值、公平正义的实现。传统的仇恨言论检测主要使用机器学习模型,其中支持向量机(support vector machine, SVM)、逻辑回归(logistic regre-ssion, LR)、朴素贝叶斯(naive Bayes, NB)等机器学习算法被广泛应用。深度学习模型因其在模型可移植性和劳动成本方面的优势,逐渐被用于仇恨言论检测研究中,其中循环神经网络(recurrent neural network, RNN)在处理文本数据时具有独特优势,并表现出良好性能,因此受到了特别关注。近年来,自注意力机制(self-attention mechanism)被逐渐应用到自然语言处理(natural language processing, NLP)工作中,它使模型能够在序列内部加权并关注到上下文关系,同时支持并行处理,从而大大提高了模型效率。

1.2　国内外学者研究现状

1.2.1　基于机器学习的仇恨言论检测研究

就仇恨言论检测任务而言,传统的机器学习方法在该领域内被广泛应用。Ting 等[2]通过 WEKA 和朴素贝叶斯算法对 Facebook 中的仇恨言论进行检测。Mehdad 等[3]提取文本的 N-gram、字符级别和情感特征并使用 SVM 对仇恨言论进行检测。Del Vigna 等[4]将句子中的单词的情感特征作为主要特征,并使用 SVM 和 LSTM 算法来判断一个句子是否属于仇恨言论。Rodriguez 等[5]从 Facebook 构建了一个仇恨言论的数据集,并提出包括负面情感词及负面表情符号在内的情感特征集合。Briliani 等[6]使用 K 近邻分类方法检测 Instagram 评论上的仇恨言论,准确率达到了 98.13%。Das 等[7]分别使用 LR、NB、KNN、DT、RF 和 SVM 算法对 Twitter 上的仇恨言论进行检测,实验结果表明,SVM、DT 和 RF 的表现优于其他所有模型。

1.2.2　基于深度学习的仇恨言论检测研究

经典的深度学习模型在文本情感分析的研究中已被广泛应用并取得显著的研究成果,但是深度学习情感分析模型的表现高度依赖于训练数据的质量和规模,其中低质量或者数量有限的数据集往往导致模型性能不佳和过拟合问题。为克服数据集的质量和规模带来的挑战,研究者们引入预训练模型。Vaswani 等[8]提出 Transformer 模型,该模型使用注意力机制和多头注意力机制,能够更好地捕捉序列之间的关系。Devlin 等[9]基于 Transformer 模型的 Encoder 部分,提出 BERT 模型,该模型在 NLP 的多项任务中都表现出色。Liu 等[10]提出基于 BERT 的改进模型 RoBERTa,该模型采用更多的训练数据、更大的批量大小以及更长的训练时间,显著提高模型性能。

特别是在仇恨言论检测任务上,深度学习技术已经展现出显著的成效。Zhang 等[11]提出一种结合 CNN 和 GRU 的深度神经网络模型,该模型能够学习更高级的语义特征表示,从而实现对 Twitter 上的仇恨言论自动分类。Kshirsagar 等[12]提出一种可变换的词嵌入模型,该模型可对一般的在线仇恨言论进行分类,特别是在种族主义和性别歧视言论分类方面具有更好的性能。Watanabe 等[13]提出一种通过 unigrams 和句法特征

来识别 Twitter 上的仇恨言论和攻击性言论的方法,并在二分类和三分类数据集上进行了验证。Patihullah 等[14]首先通过 Word2vec 提取特征,然后使用 GRU 模型来检测印度尼西亚语中的仇恨言论,实验最佳准确率达到 92.96%。Tekiröğlü 等[15]构建一个大规模数据集,并且在识别仇恨言论的模型中使用预训练语言模型 GPT-2。Mozafari 等[16]使用 BERT 和不同的嵌入层的组合来进行仇恨和攻击性言论检测,均取得了良好的效果。Albadi 等[17]分别使用仇恨言论词嵌入技术和 BERT 预训练模型进行仇恨言论检测,并比较了二者的性能,验证预训练模型在仇恨言论检测任务上的有效性。Azhari 等[18]使用 RoBERTa 方法检测印度尼西亚艺人 Instagram 贴子评论区的仇恨言论,并设置全预处理和非全预处理两种测试场景。实验结果表明,非全预处理的平均准确率高于全预处理,RoBERTa 在不使用全预处理时具有很好的预测注释性能。王琰慧等[19]提出一种基于谐音干扰词替换的中文仇恨言论检测方法,通过 N-gram 提取干扰词候选项,对相应谐音干扰词进行替换,使用 RoBERTa-wmm-ext 得到语义特征以实现仇恨言论检测任务。刘旭东等[20]结合图卷积网络的多模态仇恨迷因识别方法,在 2 个数据集上的准确率分别达到 76.03%和73.9%,优于现有的 SOTA 模型。

1.3 相关技术

1.3.1 RoBERTa 模型

RoBERTa 模型是在 BERT 模型的基础上进行改进的预训练模型,旨在提升语言表示的质量和模型的鲁棒性。类似于 BERT,RoBERTa 同样基于多层 Transformer 编码器架构,集成自注意力机制与前馈神经网络,但 RoBERTa 在预训练及微调阶段进行了若干改进以优化模型性能。在预训练阶段,RoBERTa 采用更大的批量大小和更长的序列长度,有助于模型更好地学习语言的上下文信息。同时 RoBERTa 采用动态掩码技术,即在每次迭代中随机遮蔽部分 token,而非固定遮蔽连续位置的 MASK token,使模型能够更加灵活地理解文本的上下文信息。此外 RoBERTa 还利用更多的训练数据与更长的预训练时长,进一步增强模型的泛化性能与语言表征能力。

1.3.2 大语言模型及其微调方法

随着 NLP 领域中预训练模型的发展,大语言模型(large language model, LLM)展现出卓越的性能。本文使用阿里云研发的通义千问大语言模型系列中的 Qwen-1_8B-Chat[21]开源大模型。它是一个经过大规模数据预训练的语言模型,具有强大的自然语言理解能力,能够处理复杂的文本数据、理解语境以及捕捉语言的细微特征。同时,Qwen-1_8B-Chat 模型还具备文本生成的能力,可以通过模型生成与训练数据风格一致但内容新颖的样本,从而为有效扩充数据集并提升样本多样性提供可能。此外 Qwen-1_8B-Chat 模型支持灵活的微调,允许使用者根据具体需求调整模型参数,能够在有限的计算资源下达到较好的性能表现。大语言模型在多领域性能卓越,但由于其高昂的运算成本,仅通过大语言模型实现对大规模文本的检测是不可行的。因此,使用大语言模型扩充与平衡数据集,进而训练小型的仇恨言论检测模型,也是实现文本分类的可行方案。

大语言模型的微调方法包括有监督微调(supervised fine tuning, SFT)和基于人类反馈的强化学习微调(reinforcement learning with human feedback, RLHF)等。本文所使用的 Chat 型模型是在 Base 模型基础上通过微调得到的,根据官方的基准评估,有监督微调能够显著提升模型性能,因此本文采用有监督微调的方法。微调过程本质上是一种模型的自适应过程,涉及更新预训练得到的 Base 模型中的所有参数。然而,对于大型语言模型而言,直接更新所有参数的计算需求极高,实验表明这一过程无法仅通过单个 GPU 完成。此外,完全微调后的模型参数量庞大,对存储空间的需求也显著增加。因此本文采用一种低秩自适应方法 LoRA(low-rank adaptation, LoRA)[22]以实现高效的微调。相较于在原始大型语言模型上添加适配器的方法[23]和通过优化前缀以改进连续提示生成的方法[24],LoRA 方法避免了因适配器顺序执行而丧失并行计算优势丧失以及模型推理时出现的显著延迟问题。通过使用低秩矩阵来表示原始参数的变化,LoRA 大幅提高了微调的效率,并显著降低了微调所需的硬件门槛,使得微调过程能够在单个 GPU 上完成,从而为低成本高效微调提供了可能性。

LoRA 的基本原理如图 1 所示。对于大小为 $d\times k$ 的一个矩阵,用大小分别为 $d\times r$ 和 $r\times k$ 的 2 个矩阵相乘的形式来表示。由于 r 远小于 d 和 k,因此存储这 2 个矩阵所用空间远小于直接存储原矩阵所需空间。在实施 LoRA 微调策略时,具体操作是在网络架构中引入一个旁路结构,如图 1 所示。在采用 LoRA 进行微调训

练的过程中,原始网络的参数 $\boldsymbol{W}$ 将被固定不变,仅对右侧的路分支进行训练。通过这种方式,实际参与训练的参数数量远远少于全参数的训练方法,所需的显存资源消耗大致等同于模型推理阶段的资源消耗。微调结束后,调用微调后的 Qwen-1_8B-Chat 大模型,根据正负样本比例生成新的正负样本,之后将大模型生成的新样本与原有数据集合并得到扩充与平衡后的数据集。

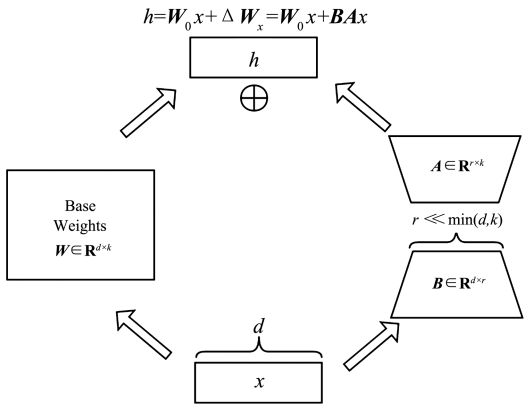

图 1　LoRA 原理

Fig.1　LoRA principle

2　模型描述

通过使用 LoRA 微调 Qwen-1_8B-Chat 实现数据集的扩充与平衡,并基于预训练模型 RoBERTa 构建 RoBERTa-Attention-GRU-TextCNN(RAGT)英文仇恨言论检测模型。该模型共有 3 层,分别是嵌入向量表示层、语义信息提取层和情感计算层。首先通过 RoBERTa 模型得到数据集中句子级的语义特征表示,然后利用自注意机制获取单词间的依赖关系,将经过自注意力机制处理后的特征矩阵输入至 GRU-TextCNN 组合层以深入挖掘更精细的语义信息和局部特征,最后使用 Softmax 分类器对评论语句进行情感分类,判断其是否属于仇恨言论。本文的整体模型架构如图 2 所示。

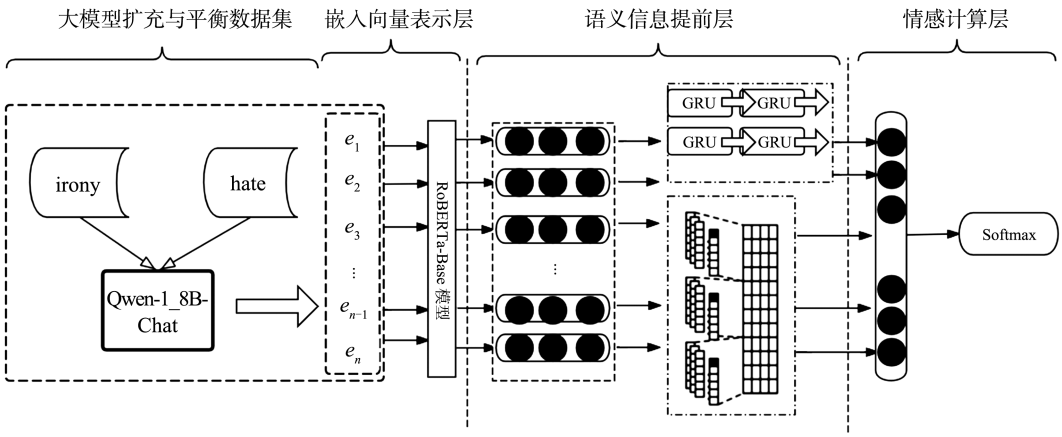

图 2　RAGT 整体模型架构

Fig.2　RAGT overall model architecture

2.1　LoRA 微调大模型以扩充与平衡数据集

使用 LoRA 微调 Qwen-1_8B-Chat,首先要构建微调数据集,本文利用原始数据集的训练集生成对应的微调数据集。微调数据集将训练集中的情感标签作为输入值,将其对应的英文文本作为输出值,微调数据集的格式如图 3 所示。

```
{
    "id":"<训练数据 id>",
    "conversations": [
            {
              "from":"user","value":"<标签输入>"
            },
            {
              "from":"assistant","value":" <文本输出>"
            }
      ]
}
```

图 3　微调数据集格式示例

Fig.3　Example of fine-tuning the dataset format

微调数据集生成后使用 LoRA 算法微调 Qwen-1_8B-Chat 模型,大模型微调及输出见分别为

$$\theta_{\text{fine-tuned}} = \arg\min_{\theta} \sum_{D_{\text{train}}} \ell(f_\theta(x), \langle \text{text} \rangle), \tag{1}$$

$$\text{Qwen\_Output} = f_{\theta_{\text{fine-tuned}}}(\langle \text{labels} \rangle), \tag{2}$$

其中:θ 为模型参数;D_{train}为原始数据集的训练集;ℓ 为损失函数;〈text〉为文本信息;$\theta_{\text{fine-tuned}}$为微调后的模型参数;〈labels〉为情感标签信息;$f_\theta(x)$为模型的预测值,即要生成的文本内容;Qwen_Output 为微调后的大模型输出。

2.2 嵌入向量表示层

将文本数据输入到 RoBERTa 模型中,由 RoBERTa 对文本特征进行提取,获得句子级的语义特征表示。本文调用 Huggingface 中的 RoBERTa-Base 英文预训练模型,它可以将文本数据转化成需要的向量形式,具体过程如图 4 所示。

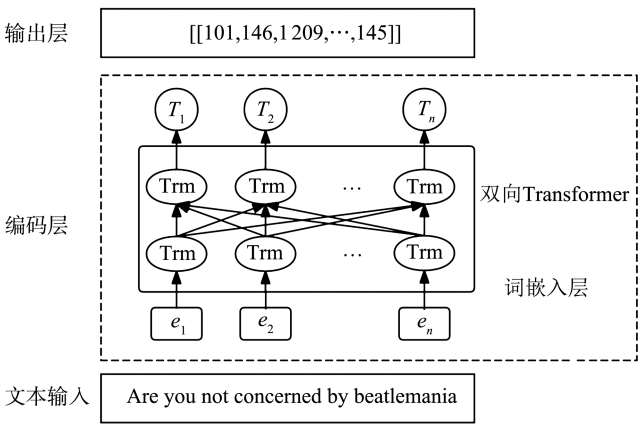

图 4 RoBERTa 词向量化

Fig.4 RoBERTa word vectorization

由图 4 可知,编码过程包括 3 个部分,即文本输入、编码层和输出层。对于一条预处理后的文本语句$E=\{e_1,e_2,\cdots,e_m\}$,首先通过构建语句 E 的 input_ids 和 attention_mask,作为 roberta-base 模型的输入,得到字级别的词向量集 $\boldsymbol{T}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_m\}$;然后,编码层的多个双向 Transform 编码器对每条文本语句进行编码,进一步提取句子级语义特征;最终将编码后得到的词向量进行拼接,得到语句 E 的词嵌入矩阵 $\boldsymbol{R}_{m\times n}$,并将其作为特征矩阵输入到模型下一层。向量拼接方式为

$$\boldsymbol{R}_{m\times n} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_2 \oplus \cdots \oplus \boldsymbol{T}_m, \tag{3}$$

其中:$\oplus$为向量拼接运算符;m 为单条语句的最大长度,由于评论信息多为短文本,因此本文设置 m 为 60;n 为词向量的维度,因为 RoBERTa-Base 模型的词向量维度为 768,所以 n 的值为 768。

2.3 语义信息提取层

从 roberta-base 模型得到词嵌入矩阵 $\boldsymbol{R}_{m\times n}$后,将其中的每个向量输入到自注意力机制,利用自注意力机制获取单词间的依赖关系从而动态生成新的词表示向量。自注意力机制的结构如图 5 所示。

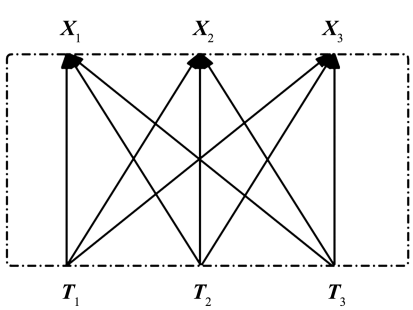

图 5 自注意力机制结构

Fig.5 Structure of self-attention mechanism

自注意力机制考虑到每个单词之间的相互依赖关系,并通过权重量化计算依赖程度。例如,在生成新的词向量 $\boldsymbol{X}_1$ 时,$\boldsymbol{T}_1$ 将与输入序列中的 $\boldsymbol{T}_1$、$\boldsymbol{T}_2$、$\boldsymbol{T}_3$ 依次计算依赖权重,然后采用加权求和的方式得到向量 $\boldsymbol{X}_1$。具体计算过程如下:

$$\boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{X} + b_Q, \tag{4}$$

$$\boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{X} + b_K, \tag{5}$$

$$\boldsymbol{V} = \boldsymbol{W}_V\boldsymbol{X} + b_V, \tag{6}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_Q}}\right)\boldsymbol{V}, \tag{7}$$

其中, $\boldsymbol{Q}$ 为查询矩阵,$\boldsymbol{K}$ 为关键字矩阵,$\boldsymbol{V}$ 为值矩阵,d_Q 为查询向量维度。由此,经计算得到新的特征矩阵 $\boldsymbol{R}^*_{m\times n}$。

通过自注意机制处理后得到新的特征矩阵 $\boldsymbol{R}^*_{m\times n}$,继续输入到 GRU-TextCNN 层中,进行全局特征和局部特征提取。通过 GRU 模型捕获长距离的语义信息,进行全局特征提取。特征矩阵 $\boldsymbol{R}^*_{m\times n}$ 经过 GRU 模型后,序列整体特征被进一步提取,挖掘到更加深层次的语义信息,矩阵表示变为 $Y_{m\times n}$。同时,对特征矩阵 $\boldsymbol{R}^*_{m\times n}$,使用 TextCNN 模型实施一维卷积操作,并通过采用多类型卷积核来捕获文本的局部特征。本文选择窗口大小分别为 3、4、5 的 3 种不同尺寸的卷积核,每种尺寸各配置 100 个。卷积操作后应用 ReLU 激活函数,同时采用最大池化法缩小特征参数矩阵,保留显著特征,并将池化后的特征向量进行拼接操作。经上述操作后,生成新的特征矩阵 $\boldsymbol{Z}_{m\times n}$。

2.4　情感计算层

情感计算层将经过 GRU-TextCNN 模型得到的 2 个特征矩阵 $\boldsymbol{Y}_{m\times n}$ 和 $\boldsymbol{Z}_{m\times n}$ 进行拼接,通过 Softmax 分类器对仇恨言论进行识别和分类。为增强模型的泛化能力并预防过拟合现象,引入 Dropout 层以随机选择性地屏蔽一定比例的神经元,本文实验将随机忽略概率设定为 0.5,然后通过 2 个线性层压缩向量维度,最后将情感特征向量输送到分类 Softmax 中,根据归一化后的值来判断情感极性,越接近 0 表示是仇恨言论的可能性越大。

情感计算层将拼接后得到的特征向量通过 Softmax 分类器计算语句 E 在情感标签中的概率向量 $\boldsymbol{O}=[\boldsymbol{o}_1,\boldsymbol{o}_2]$,选择最大值所代表的情感标签作为最终输出的情感标签。Softmax 分类器将输入向量的第 i 项转化为概率

$$P_i=\frac{e^i}{\sum_{j=1}^{K}e^j},\quad i=1,2,\cdots,K, \tag{8}$$

其中 K 为情感分类类别数。本文采用反向传播算法对模型进行迭代优化,更新各层网络的权重参数,选择交叉熵作为损失函数,在训练过程中致力于将交叉熵损失降至最低。计算过程如下:

$$l_{\text{loss}}=-\sum_{i=1}^{D}\sum_{j=1}^{K}y_i^j\log\hat{y}_i^j, \tag{9}$$

其中,D 为训练集大小,K 为情感分类类别数,y 为真实的情感类别,$\hat{y}$ 为预测的情感类别。

3　实验

3.1　数据集

使用公开的推特数据集 Irony detection[25] 和 Hate speech detection[26] 对英文仇恨言论检测模型进行训练与验证。原始数据集样本数量及比例见表 1。这 2 个数据集都包含 2 列数据,即推特评论文本(text)和对应的情感标签(label)。由于数据集正负样本比例失衡,因此本文实验首先使用微调数据集对 Qwen-1_8B-Chat 进行微调,其中大模型输入样本如图 6 所示。

表 1　原始数据集样本数量及比例

Table 1　Sample size and proportion of the original dataset

数据集	正样本数	负样本数	正负样本比
Irony detection	2 212	2 389	0.926
Hate speech detection	5 462	7 500	0.728

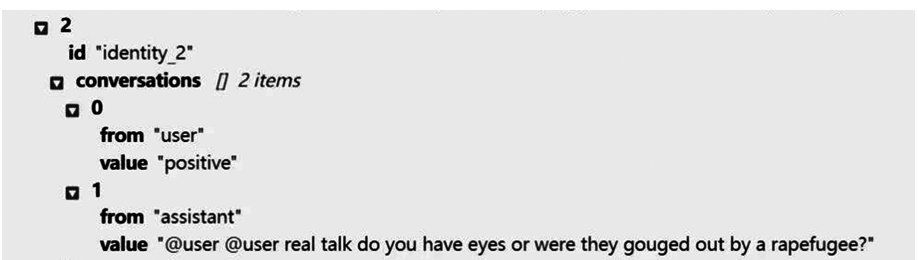

图 6　大模型输入样本示例

Fig.6　Example of a large model input sample

使用微调后的大语言模型对数据集样本进行扩充,根据原始数据集的构成情况,本文选择扩充大量正样本和少量负样本的扩充策略,分别扩充 2 个原始数据集。本文通过实验发现微调后的大模型确实能够生成训练集中未出现的样本,甚至会生成与训练集样本风格完全不同的样本。本文认为这种情况是由于大语言模型在巨大高质量语料库上经过充分预训练,这些不同风格的样本可能来源于大语言模型最初的预训练语料库。

使用微调后的大模型扩充数据集生成的样本如图 7 所示。考虑到样本重复问题,将扩充后的样本与原始数据集合并并去重,随机选择一定数量的正负样本生成扩充与平衡后的数据集,扩充与平衡后的数据集样本数量及比例见表 2。由表 2 可知,扩充后的数据集正负样本各占 50%,数据集的正负样本达到了平衡状态,实现了数据集的扩充与平衡。实验分别选取扩充与平衡后的 2 个数据集的 70%作为训练集,30%作为测试集进行模型训练与评估。

@user Lol I'm scared.
I can't believe I forgot to clear my browser cache #so-sorry
@user @user so we all just shut down when some people call Obama an evil genocide rat, except that's not how freedom works!!
The holiday season can be tiring so here's something to cheer you up #Thanksgiving2015
Love these rainy days ðð¤
@user when is the next one in Sydney? #xbbtasymptom
@user . @user I just want to be happy, blessed and healthy..
When people you care about act like you're important to them but are in fact not that important at all.
@user @user @user you know how little the libs really know about math right?
Look! It's raining cats and dogs! #UHARINOTHURST #SnowSnowSnow
@user @user He's number one...and I'm #16....I should be wearing that for my birthday tomorrow...gosh where do I put it?
a #good 2 have a dance with your n Fast food, expensive... but fun!

图 7 微调后的大模型输出样本示例
Fig.7 Example of a fine-tuned sample output from a large model

表 2 扩充后数据集样本数量及比例
Table 2 Example of a fine-tuned sample output from a large model

数据集	正样本数/个	负样本数/个	正负样本比
Irony detection	2 500	2 500	1
Hate speech detection	7 350	7 350	1

3.2 实验配置

依据研究的实际需求,本文选取的大语言模型为阿里云开发的通义千问开源大语言模型 Qwen-1_8B-Chat,即参数量为 18 亿的 Chat 型模型。实验在单个 RTX 4090 GPU 上训练模型,对大语言模型的微调采用的优化器为 AdamW 优化器。

对于 RAGT 仇恨检测模型,按照上述模型结构进行搭建。实验使用的操作系统为 Windows10,Python 版本为 3.9.7,搭配深度学习框架 Pytorch,版本为 2.1.2,使用的 CUDA 版本为 12.1。调用开源社区 Huggingface 中的 RoBERTa-Base 模型,选择 AdamW 优化器,使用交叉熵损失函数,通过权重衰减 weigh_decay 参数控制各阶段预训练任务和微调时正则化强度,防止过拟合。训练轮次 num_epoch 设为 10,训练集批次大小 train_batch_size 设为 32,测试集批次大小 test_batch_size 设为 32,学习率设为 1e-05,调节正则项的权重衰减参数 weight_decay 设为 0.01,随机失活 dropout 概率设为 0.5,数据集样本输入到 RoBERTa-Base 模型中文本的最大截取长度 max_length 设为 60。

3.3 实验结果及分析

3.3.1 对比实验

本文选取 4 组模型与 RAGT 模型进行对比,采用不同的词向量方法和分类模型进行组合,对比模型如下:

1) GloVe-LSTM 模型:将长短期记忆网络应用到文本分类任务中,通过 GloVe 模型生成每个词对应的词向量,并将其输入到 LSTM 模型中。

2) GloVe-GRU 模型:将门控循环单元网络应用到文本分类任务中,通过 GloVe 模型生成每个词对应的词向量,并将其输入到 GRU 模型中。

3) BERT 模型:采用英文预训练模型 BERT 将评论语句向量化,然后使用一层全连接层完成对仇恨言论的分类。

4) RoBERTa 模型:采用英文预训练模型 RoBERTa 将评论语句向量化,然后使用一层全连接层完成对

仇恨言论的分类。

3.3.2　实验结果分析

分别在 2 个英文数据集上进行实验,将准确率(accuracy, Acc)、精确率(precision, Pre)、*F*1 值作为衡量模型效果的指标。具体实验结果如表 3 所示。

表 3　模型实验结果对比

Table 3　Comparison of model experimental results　　单位:%

模型	Irony detection			Hate speech detection		
	Acc	Pre	*F*1	Acc	Pre	*F*1
GloVe-LSTM	63.43	67.92	65.52	69.25	74.67	72.04
GloVe-GRU	63.58	66.38	65.63	69.93	72.66	71.75
BERT	64.64	64.94	58.69	74.71	72.98	69.50
RoBERTa	70.41	69.83	69.30	75.75	75.36	68.94
RAGT	**73.02**	**76.71**	**73.73**	**80.43**	**80.67**	**81.23**

分析 Irony detection 数据集上实验结果可知,以 GloVe 模型作为词嵌入向量的 LSTM 和 GRU 模型的各项评价指标均较低,与 RoBERTa 相比还存在一定差距。这一方面是因为评论文本结构不完整、逻辑混乱的自身特点以及数据量太少,模型难以捕捉到有用的语义信息;另一方面与 LSTM 和 GRU 这两种循环神经网络架构在处理长距离依赖关系时的固有局限性有关,它们难以捕捉到讽刺和仇恨言论中复杂的语义和情感变化。BERT 模型在 Irony detection 上的表现不如 LSTM 和 GRU,可能是因为在处理讽刺这种需要理解语境和微妙语义的任务时,BERT 其预训练任务与讽刺检测任务之间的差异较大,导致其优势没有得到充分发挥。这表明仅改变词向量转换方法不一定能提升仇恨言论检测的效果。RoBERTa 模型的性能优于 BERT,是因为 RoBERTa 模型在 BERT 模型的基础上进行了改进,其采用更大的批量大小、更长的序列长度和动态掩码技术,进一步增强模型的泛化性能,这也是本文选取它作为底层模型的原因之一。而本文提出的 RAGT 模型,在准确率、精确率和 *F*1 值上都明显优于其他模型,相较于 RoBERTa 模型,各项指标分别提升 2.61%、6.88%、4.43%,具有良好的仇恨言论检测性能。

分析 Hate detection 数据集上实验结果可知,以 GloVe 模型作为词嵌入向量的 LSTM 和 GRU 模型的各项评价指标随着样本量的增多有所提升,部分指标甚至超过了 BERT、RoBERTa 模型。这说明样本量对深度学习模型的影响较大,数据集样本的增多会显著提高模型性能,尤其是在处理类别不平衡或样本量较少的任务时,数据量的增加可以显著提高模型的泛化能力。BERT 与 RoBERTa 的模型效果差距不是太大,这是因为其都采用了 Transformer 架构,能够有效处理长距离依赖关系,且在预训练阶段已经接触到了大量多样化的数据,为它们在不同评价指标上的表现提供了坚实的基础,二者在不同的评价指标上各占优势,符合样本量适中时使用二者进行分类的一般效果。相较于其他模型,本文提出的 RAGT 模型具有独特的架构设计,能更好地捕捉到仇恨言论的复杂特征,在准确率、精确率和 *F*1 值指标上表现更出色,与 RoBERTa 模型相比在这些指标上分别提高 4.68%、5.31%、12.29%,具备良好的仇恨言论检测性能。

3.4　消融实验

为验证模型各个模块的有效性,本文设置以下消融实验以进行对比分析。

1) RAGT-NL:不使用大语言模型扩充与平衡数据集的 RAGT,即该方法在原始数据集上进行实验。

2) RoBERTa-GRU:去掉 TextCNN 模型和自注意力机制,只使用 RoBERTa 和 GRU 模型,在扩充与平衡后的数据集上进行实验。

3) RoBERTa-TextCNN:去掉 GRU 和自注意力机制,只使用 RoBERTa 和 TextCNN 模型,在扩充与平衡后的数据集上进行实验。

4) RoBERTa-Attention:去掉 GRU 和 TextCNN 模型,只使用 RoBERTa 和自注意力机制,在扩充与平衡后的数据集上进行实验。

5) RoBERTa-GRU-TextCNN:去掉自注意力机制,只使用 GRU 和 TextCNN 模型,在扩充与平衡后的数据集上进行实验。

在仇恨言论检测任务上进行消融实验,实验结果见表 4,验证了大语言模型扩充与平衡数据集、自注意力机制和 GRU-TextCNN 模型 3 个模块对仇恨言论检测的必要性。

表 4 消融实验结果

Table 4 Ablation test results

单位:%

模型	Acc	Pre	*F*1
RAGT-NL	79.24	78.18	77.23
RoBERTa-GRU	79.46	75.04	79.63
RoBERTa-TextCNN	79.16	76.13	78.15
RoBERTa-Attention	78.95	76.48	80.03
RoBERTa-GRU-TextCNN	79.48	78.54	79.94
RAGT	**80.43**	**80.67**	**81.23**

对消融实验结果的分析如下。

1) RAGT-NL:相较于使用大语言模型扩充与平衡数据集的 RAGT 模型,效果稍逊一筹。这表明在仇恨检测任务中,正负样本的平衡程度对实验结果有着一定影响,使用大语言模型平衡与扩充数据集是有效的解决方法之一。

2) RoBERTa-GRU:去掉 TextCNN 模型和自注意力机制,Pre 下降了 5.63%,速度较快这表明 GRU 模型在处理序列数据时虽然能够有效捕获文本中的时序特征,但可能在某些情况下对文本的局部特征捕捉存在不足。

3) RoBERTa-TextCNN:去掉 GRU 和自注意力机制,*F*1 下降了 3.08%,下降速度较快。这表明 TextCNN 模型在捕获文本的局部特征方面表现出色,尤其是在短文本中,能够识别出关键的词汇和短语,但无法有效捕获文本中的时序特征。

4) RoBERTa-Attention:GRU 和 TextCNN 模型被去掉后,Acc 和 Pre 下降较快,分别下降了 1.48%、4.19%,模型性能下降。这表明对于短文本分类,捕捉更深层次的语义信息和局部关键特征是有必要的。

5) RoBERTa-GRU-TextCNN:不使用自注意力机制后,模型性能也有所下降。这表明利用自注意力机制获取单词间的依赖关系也会对短文本仇恨言论检测产生一定影响。

4 结论

本文通过微调 Qwen-1_8B-Chat 扩充与平衡数据集,并基于预训练模型 RoBERTa,构建了一个融合深度学习模型组合的 RAGT 模型,将深度学习强大的特征捕获和提取能力应用到文本序列数据的分析、挖掘中。实验结果表明,该模型相较于传统的深度学习模型具有更好的检测效果。

本文模型充分探究利用预训练模型实现短文本仇恨检测任务高准确率与高效率的方法,并对所提出模型相对于传统策略的优越性进行了验证与分析,取得了令人满意的实验结果。未来的研究工作可以在以下几个方面进行进一步探索:首先,扩充情感类别,使模型能够识别和分类更多种类的情感;其次,进行长文本和中文文本的分类,提升模型的精确度;最后,通过模型创新,使其能够对多模态数据(如文本、图像、视频等)进行仇恨检测,从而拓宽模型的应用范围,进一步提高模型的泛化能力。

参考文献:

[1] 联合国大会.促进宗教间和文化间对话与容忍,打击仇恨言论[EB/OL].(2021-07-22)[2024-10-09]. https://documents.un.org/doc/undoc/gen/n21/200/60/pdf/n2120060.pdf.
United Nations General Assembly. Promoting interreligious and intercultural dialogue and tolerance to combat hate speech[EB/OL].(2021-07-22)[2024-10-09]. https://documents.un.org/doc/undoc/gen/n21/200/60/pdf/n2120060.pdf.

[2] TING I H, CHI H M, WU J S, et al. An approach for hate groups detection in facebook[C]// Proceedings of the 3rd International Workshop on Intelligent Data Analysis and Management. Dordrecht: Springer, 2013:101-106.

[3] MEHDAD Y, TETREAULT J. Do characters abuse more than words? [C]// Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles: ACL, 2016:299-303.

[4] DEL VIGNA F, CIMINO A, DELL'ORLETTA F, et al. Hate me, hate me not: hate speech detection on Facebook[C]// Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). Venice: CEUR, 2017:86-95.

[5] RODRIGUEZ A, ARGUETA C, CHEN Y L. Automatic detection of hate speech on facebook using sentiment and emotion analysis[C]// Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Okinawa: IEEE, 2019:169-174.

[6] BRILIANI A, IRAWAN B, SETIANINGSIH C. Hate speech detection in indonesian language on instagram comment section

using *K*-nearest neighbor classification method[C]//Proceedings of the 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). Bali: IEEE, 2019:98-104.

[7] DAS S, BHATTACHARYYA K, SARKAR S. Performance analysis of logistic regression, naïve Bayes, KNN, decision tree, random forest and SVM on hate speech detection from twitter[J]. International Research Journal of Innovations in Engineering and Technology, 2023, 7(3):24-28.

[8] VASWANI A, SHAZEEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. New York: Curran Associates, Inc., 2017:5999-6009.

[9] 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, Minneapolis: ACL, 2019:4171-4186.

[10] LIU Z, LIN W, SHI Y, et al. A robustly optimized BERT pre-training approach with post-training[C]//Proceedings of the 20th China National Conference on Chinese Computational Linguistics. Hohhot: Springer, 2021:471-484.

[11] ZHANG Z, ROBINSON D, TEPPER J. Detecting hate speech on Twitter using a convolution-gru based deep neural network [C]//Proceedings of the Semantic Web: 15th International Conference. Heraklion: Springer, 2018:745-760.

[12] KSHIRSAGAR R, CUKUVAC T, MCKEOWN K, et al. Predictive embeddings for hate speech detection on Twitter[C]//Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels: ACL, 2018:26-32.

[13] WATANABE H, BOUAZIZI M, OHTSUKI T. Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection[J]. IEEE Access, 2018, 6:13825-13835.

[14] PATIHULLAH J, WINARKO E. Hate speech detection for Indonesia tweets using word embedding and gated recurrent unit [J]. Indonesian Journal of Computing and Cybernetics Systems, 2019, 13(1):43-52.

[15] TEKÖRĠLÜ S S, CHUNG Y L, GUERINI M. Generating counter narratives against online hate speech: data and strategies[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Hohhot: ACL, 2020:1177-1190.

[16] MOZAFARI M, FARAHBAKHSH R, CRESPI N. A BERT-based transfer learning approach for hate speech detection in online social media[C]//Proceedings of the Eighth International Conference on Complex Networks and Their Applications. Lisbon: Springer, 2020:928-940.

[17] ALBADI N, KURDI M, MISHRA S. Are they our brothers? Analysis and detection of religious hate speech in the arabic twittersphere[C]//Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona: IEEE, 2018:69-76.

[18] AZHARI A A, SIBARONI Y, PRASETIYOWATI S S. Detection of Indonesian hate speech in the comments column of indone-sian artists' instagram using the RoBERTa method[J]. Jurnal Ilmiah Penelitian dan Pembelajaran Informatika, 2023, 8(3):764-773.

[19] 王琰慧,王小龙,张顺香,等. 基于谐音干扰词替换的中文仇恨言论检测方法[J]. 应用科技,2024,51(3):72-81.
WANG Yanhui, WANG Xiaolong, ZHANG Shunxiang, et al. A Chinese hate speech detection method based on homophonic interference word replacement[J]. Applied Science and Technology, 2024, 51(3):72-81.

[20] 刘旭东,杨亮,张冬瑜,等. 结合图卷积网络的多模态仇恨迷因识别研究[J]. 重庆理工大学学报(自然科学),2024,38(1):169-179.
LIU Xudong, YANG Liang, ZHANG Dongyu, et al. Research on multimodal hate meme recognition based on graph convolutional network[J]. Journal of Chongqing University of Technology (Natural Science), 2024, 38(1):169-179.

[21] BAI Jinze, BAI Shuai, CHU Yunfei, et al. Qwen technical report[EB/OL].(2023-09-28)[2024-10-09]. https://doi.org/10.48550/arXiv.2309.16609.

[22] HU E J, SHEN Y L, WALLIS P, et al. Lora: low-rank adaptation of large language models[EB/OL]. (2021-06-17)[2024-10-09]. https://doi.org/10.48550/arXiv.2106.09685.

[23] HOULSBY N, GIURGIU A, JAŚTREBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of the 36th International conference on machine learning. Long Beach: PMLR, 2019:2790-2799.

[24] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL, 2021:4582-4597. https://aclanthology.org/2021.acl-long.353/.

[25] VAN HEE C, LEFEVER E, HOSTE V. Semeval-2018 task 3: irony detection in english tweets[C]//Proceedings of the 12th International Workshop on Semantic Evaluation. New Orleans: ACL, 2018:39-50.

[26] BASILE V, BOSCO C, FERSINI E, et al. Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter[C]//Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis: ACL, 2019:54-63.

(编辑:李艺)