

基于情感增强机制的大语言模型虚假新闻检测

冉广煜, 肖克晶

北京印刷学院信息工程学院, 北京

收稿日期: 2025年12月29日; 录用日期: 2026年1月26日; 发布日期: 2026年2月5日

摘要

为解决现有新闻文本虚假检测方法仅依赖语义特征、忽视情感特征, 导致复杂内容检测准确度低的问题, 提出一种基于情感增强机制的大语言模型虚假新闻检测方法 (Sentiment-Enhanced Large Language Model for Fake News Detection, SELLM-FND)。该方法先对新闻文本进行情感分析以提取情感特征, 再通过大语言模型融合文本与情感特征完成检测。在WELFake_Dataset_Edited数据集上的实验显示, 该方法准确率达0.929, 检测性能优于以往基于文本的虚假新闻检测方法。

关键词

虚假新闻检测, 大语言模型, 情感增强, 情感特征提取

False News Detection of Large Language Model Based on Emotion Enhancement Mechanism

Guangyu Ran, Kejing Xiao

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: December 29, 2025; accepted: January 26, 2026; published: February 5, 2026

Abstract

In order to solve the problem that the existing false news detection methods only rely on semantic features and ignore emotional features, which leads to the low accuracy of complex content detection, a sentient-enhanced large language model for false news detection (SELLM-FND) based on emotional enhancement mechanism is proposed. This method firstly analyzes the news text to extract emotional features, and then completes the detection by fusing the text and emotional features through the large language model. Experiments on WELFake_Dataset_Edited data set show that the accuracy of this method is 0.929, and the detection performance is better than the previous text-

based false news detection methods.

Keywords

False News Detection, Large Language Models, Emotion Enhancement, Emotional Feature Extraction

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在数字时代, 信息获取变得十分便捷, 但信息的可信度却无法得到保证。互联网的开放性使得信息能够无限制传播, 进而导致虚假信息扩散[1]。虚假信息误导公众舆论, 引发信任危机并扰乱社会秩序。为解决这些问题, 近年来学界提出了自动虚假新闻检测方法, 并采用新型、更强大的机器学习模型来识别虚假新闻及评估新闻危害性[2]。

自从使用机器学习模型识别虚假新闻(以下简称虚假新闻检测)这一任务被提出后, 研究者们已提出多种方法, 旨在采用非人工方法快速识别虚假新闻, 防止用户沦为社交媒体上快速扩散的误导性信息的受害者[3]。根据检测特征与逻辑维度, 现有方法可大致划分为“基于内容的检测方法”“基于传播的检测方法”与“基于社交上下文的检测方法”三类[4], 其中基于内容的虚假新闻检测是指仅依赖新闻文本(或其他多媒体内容)自身属性判断新闻真实性的方法。

由于“新闻”以文本内容为主的特性, 基于文本内容的虚假新闻检测是自动虚假新闻检测领域的核心技术方向之一, 该类方法无需依赖图像、音频或社交传播数据等其他信息, 只从新闻文本本身提取可量化特征, 通过机器学习或深度学习模型实现虚假新闻的自动识别。

基于文本内容的虚假新闻检测的核心是准确地提取文本内容的特征, 例如词汇频率、句法结构、事实引用数量等显性特征, 以及语义一致性、上下文依赖、潜在立场等隐形特征[5]-[8]。现有方法多聚焦于语义特征, 却忽视了情感特征在虚假新闻检测中的重要作用。从心理学角度看, 虚假新闻常通过激发受众强烈情绪影响其认知判断; 从传播学角度, 情感是虚假新闻快速传播的重要驱动力。基于此, 本文提出基于情感增强机制的大语言模型虚假新闻检测方法, 具体流程如下: 在模型训练与微调阶段, 先训练用于情感分析的子模型, 再利用该部分模型为新闻数据集自动添加情感标注, 然后利用该数据集通过低秩适配(Low-Rank Adaptation, LoRA)微调策略微调出可以通过文本语义和情感标签检测虚假新闻的核心部分。在部署阶段, 该方法先对新闻的文本进行情感分析提取其中的情感特征, 再对提取出来的情感特征和新闻文本进行整合与检测, 实现更高精度的虚假新闻检测。

2. 相关工作

2.1. 基于文本的虚假新闻检测

基于文本的虚假新闻检测是识别新闻危害性的关键工具, 主要发挥两方面作用: 1) 作为独立方法直接检测虚假新闻; 2) 作为可信度评估、网络传播模式分析等其他检测方法的增强手段[9][10], 或多模态虚假新闻检测中的单模态处理模块。

目前该领域已有较多研究成果: Philogene 等人采用双向长短期记忆(BiLSTM)模型开展虚假新闻检测[11]; Hrishikesh Telang 等人运用 LSTM 变种模型——门控循环单元(GRU)完成检测任务[12]; Ye-Chan

Ahn 等首次将预训练模型 BERT 应用于该任务, 通过 WordPiece 模型对数据集进行分词, 经微调后得到虚假新闻检测模型[6]。Bibek 等人聚焦情感与情绪特征, 将新闻情感划分为“积极”“中立”“消极”三类, 扩展 LIAR 数据集并提出融合情感分析的检测方法[7]; Chunyuan Yuan 等人利用预训练 BERT 模型对评论进行情感倾向标注, 获取弱标注数据[8]; Mahmood 等人采用双 BERT 模型并行处理新闻标题与正文, 以提升语义捕捉能力及检测准确率[13]。当前, 以 BERT 为核心的方法仍是该领域主流。此外, 近两年大模型逐步应用于基于文本的虚假新闻检测, 如 Beizhe Hu 等人利用大模型的解释能力生成分析依据以辅助检测[14], Ke Jing 等人则直接借助大模型的推理能力实现语义增强, 进而完成检测[15]。

通过对以上研究的分析, 可以看出过往研究多聚焦于文本的显性特征(如词汇频率、句法结构)的提取, 情感分析环节主要采用“积极-中立-消极”的三分类方法, 未能充分挖掘细粒度情感特征对检测性能的提升作用, 且缺乏对情感特征作用机制的理论分析。

2.2. 细粒度情感分析

情感分析的核心任务是识别文本情感并按极性划分为“积极”“中立”“消极”三类[16]。随着网络用户生成内容激增, 自动情感分析的研究关注度持续提升, 且向更深层次发展, 不仅需识别文本情感倾向(如正面、负面), 还需进一步确定情感所指向的具体方面或属性, 即细粒度情感分析。

其主要研究方向包括两方面: 1) 从文本维度出发, 将单段文本划分为多个部分, 各部分对应不同情感倾向, 该方向通常称为方面级情感分析; 2) 从情感判定维度出发, 将情感极性判断转化为情感类型识别, 使情感分析结果从三类扩展至更多类别, 该方向通常称为多标签情感分类[17]。

在文本分类任务中, 细粒度情感分析技术已广泛应用, 且普遍认为其相较于三分类情感分析方法, 更有助于提升文本分类准确度[18]-[20]。与此同时, 在过往针对虚假新闻检测的研究实践中, 研究者发现采用三分类情感分析方法实现情感信息增强, 能够显著提高虚假新闻检测的成功率[7][8]。基于上述研究结论可推知, 若能进一步提升情感分析技术本身的精度, 深入分析不同情感类型在虚假新闻中的分布特征及作用, 将对虚假新闻检测成功率的提升产生更积极作用。

2.3. 大语言模型与微调

大语言模型(Large Language Models, LLM)是基于深度学习技术、以海量文本数据为基础的人工智能模型, 核心能力包括理解与生成人类语言, 同时具备一定的逻辑推理、知识存储及跨任务适配能力[21]。其具体优势体现为: 拥有强大的复杂语境处理能力, 可深度挖掘文本潜藏的隐含信息, 为信息分析提供全面支撑; 具备更高效的跨领域迁移能力, 能在不同场景任务中快速适配, 降低场景切换的技术适配成本; 相较于传统模型, 可解释性更强, 该特性直接提升检测结果的可信度, 为技术应用可靠性提供重要保障[22]。基于以上优点, 近两年大模型在包括虚假新闻检测在内文本分类任务中得到了广泛应用。

模型微调是指在特定任务或领域数据上进一步训练模型, 使其更好适配目标场景的技术手段, 是大模型从通用走向专业的关键途径。经微调后, 大模型在特定任务中的表现显著优于通用状态。

针对虚假新闻检测任务, LLM 主要微调策略分为两类: 全参数微调(更新预训练模型所有参数)与参数高效微调(Parameter-Efficient Fine-Tuning, PEFT, 仅调整模型少量参数)。相较于全参数微调, PEFT 可根据需求调整 0.5%~1%比例的参数, 具有资源需求低、原模型结构保留度高、过拟合风险低、性价比高等优势[23]。

2.4. 基于提示词的情感增强

提示词工程在解锁大语言模型潜能上具有重要作用。该方法通过设计提示指令指导模型响应。确保

响应的相关性、连贯性和准确性。提示工程无需微调模型参数, 可与下游任务无缝衔接[24]。

2021年, Reynolds 和 McDonell 首次在《Prompt programming for large language models: Beyond the few-shot paradigm》中提出了提示词工程的概念[25]。经过多年的发展, 如今的提示词工程方法已经发展出了多种技术路径: Kojima 等人提出零样本思维链(CoT), 通过引导模型逐步推理提升复杂任务表现[26]; Yao 等人扩展出思维树(ToT)和推理与行动(ReAct), 分别优化了复杂规划任务和推理-行动协同场景[27][28]; Besta 等人提出思维图(GoT), 以图结构融合多推理路径增强可解释性[29]; Zhou 等人提出自动提示工程(APE), 实现指令的自动生成与筛选[30]; Lewis 等人提出检索增强生成(RAG), 通过外部知识库提升回答准确性与实时性, 这些技术共同推动提示词工程成为解锁大语言模型潜能的核心手段[31]。

2.5. 基于情感增强机制的大语言模型虚假新闻检测

针对过往的检测方法缺少对情感特征的深度挖掘的问题, 本文将细粒度情感分析应用在虚假新闻检测领域, 结合大语言模型与模型微调, 设计了一个拥有情感增强机制的用于虚假新闻检测的大语言模型。

3. 本文方法

3.1. 任务描述

本文将假新闻检测任务定义为一个二分类任务, 给定一个新闻的文本信息 T , 使其输出预测结果 y , 即需要设计函数 f , 使其满足 $y = f(T)$ 。为了实现情感增强, 本文还需要对新闻文本 T 进行情感分析, 其中情感分析过程定义为函数 g_2 , 情感分析的结果设为 Emo , 即有 $Emo = g_2(T)$ 。同时将情感分析结果作为后续新闻检测的依据之一, 即重新设计一个新的函数 f_2 , 可得:

$$y = f(T) = f_2(T, g_2(T)) \tag{1}$$

3.2. 模型设计

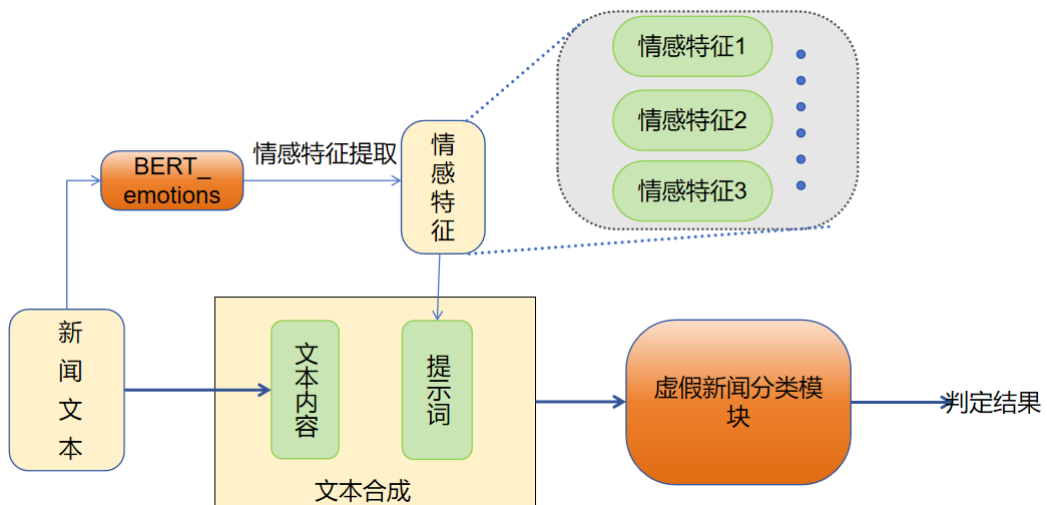


Figure 1. Sentiment-enhanced LLM for fake news detection model
图 1. 基于情感增强机制的大语言模型虚假新闻检测模型

根据上一节中对于任务的分析, 本节设计了一个基于情感增强机制的大语言模型虚假新闻检测方法, 具体包括以下几个模块: (1) 基于 BERT 的情感分析模块; (2) 利用大语言模型结合新闻文本和所提取的情感特征的虚假新闻分类模型; (3) 文本合成模块。如图 1。

基于情感增强机制的大语言模型虚假新闻检测方法的模型由情感处理模块、文本合成模块与虚假新闻检测模块组成, 情感分析模块生成新闻文本的情感特征, 并将情感标注与语义分析的结果进行文本合成, 输入特征融合模块, 经过分类器分类后输出判断结果。

3.2.1. 情感标注

本文所提出的 SELLM-FND 模型, 首先需要对新闻的文本进行情感分析提取其中的情感特征, 情感分析部分使用 BERT 模型作为基础。通用 BERT 模型基于双向多层 Transformer 编码器架构构建。其中 BERT-Base 模型的架构包含 12 个编码器, 每个编码器由 8 层组成: 4 层多头自注意力层(multi-head self-attention layers)和 4 层前馈网络层(feed forward layers)。基于双向 Transformer 架构, 通过自注意力机制能够同时关注句子中前后文的所有单词, 生成上下文敏感的词嵌入。

由于 Transformer 的自注意力机制, BERT 能通过替换合适的输入和输出来适配多种下游任务。对于情感分析中常见的一词多义、反讽、否定结构等现象, BERT 能根据具体语境动态调整单词的表示, 准确区分其情感倾向。

本文通过对预训练模型 BERT 进行微调, 获得具备情感分析能力的 BERTemotions 模型。为使该模型区别于现有仅能进行极性判断的情感分析模型, 本文使用情感数据集(见 4.1 节)对其进一步训练, 使其能够识别十一种情感, 并输出由这些情感倾向组成的情感向量。为在充分体现文本情感倾向的同时减少模型计算量, 我们进一步优化模型, 使其仅输出情感倾向最强烈的三种情感所组成的情感向量。

3.2.2. 数据集加工

为支撑本文研究, 实验所需数据集需要满足“同时包含情感标签与新闻真实性标签”的高质量标注要求, 然而, 一方面, 虚假新闻检测任务的数据集往往仅标注“真实/虚假”二元标签或简单的情感标签, 缺乏细粒度情感标注(如愤怒、恐惧、喜悦); 另一方面, 情感分析专用数据集虽包含丰富的情感类别标注, 却未关联新闻文本的真实性标签, 无法直接满足“基于情感特征预测新闻真实性”的实验设计目标。

针对已有数据集不满足本文的研究目标的问题, 本文采用在训练中途对数据集进行二次加工的方式, 将训练任务一分为二, 并将数据集预处理穿插在二者之间, 具体流程分为三步: 第一步, 对预训练模型 BERT 进行微调, 使其具备情感分析能力; 第二步, 基于微调后的模型为新闻数据集(见 4.1 节)添加情感标注; 第三步, 利用完成情感标注的新数据集, 对模型剩余部分展开训练。

3.2.3. 大模型及微调

Wael Etaiwi 等人比较了 GPT-4 和 DeepSeek-V3 在主题分类等五个领域上的能力, 发现 DeepSeek 在语义分析和主题分类上优于 GPT-4 [32], 综合对比 GPT-4, DeepSeek-R1 和 DeepSeek-V3 的部署需求(如表 1), 本实验最终选用经新闻分类任务微调后的 DeepSeek-R1 模型(DeepSeek-R1-Distill-Qwen-7B-News-Classifer)作为基础模型(以下以基础模型代指) [33], 该基础模型已经具备新闻文本分类的基础能力, 通过进一步微调可以进行虚假新闻检测功能, 使其能够接受文本内容和情感倾向的输入, 并输出新闻真实性判断。

Table 1. Comparison of DeepSeek-V3, DeepSeek-R1, and GPT-4 models

表 1. DeepSeek-V3、DeepSeek-R1、GPT-4 模型对比

对比维度	DeepSeek-V3	DeepSeek-R1	GPT-4
参数辆	6710 亿总参数, 单 token 激活 370 亿参数	参数规模 100~300 亿	参数规模约 1.8 万亿 (行业推测)
上下文窗口长度	推测 16k~32k tokens (可稳定 处理长新闻文本、多段落信息)	8k~16k tokens (长文本覆盖能力有限)	标准版 8k tokens, GPT-4 Turbo 支持 128k tokens (适配超长新闻报道)

续表

长文本处理能力	5%~8%, 长文本逻辑链捕捉效率高	上下文衰减 20%+, 易丢失跨段落事实关联, 长文本可信度判断稳定性弱	128k 窗口衰减 10%~15%, 需高性能硬件, 成本高
事实矛盾识别 (F1 分数)	84.6%	65%	76%
长文本虚假信息甄别(召回率)	62.5%+	60%	28.6%
技术领域事实判断(F1 分数)	72.7%	70%	88.2%
事实冲突整体识别(F1 分数)	69.3%	65%	64%

典型的神经网络包含大量执行矩阵乘法的密集层, 这些层中的权重矩阵通常允许具有满秩。然而, Edward Hu 等人[28]的研究表明, 在模型适配特定下游任务时, 预训练语言模型存在“内在维度较低”的特性——即便通过低维参数重参数化, 模型仍能高效学习。对于预训练权重矩阵 $W_0 \in \mathbb{R}^{d \times k}$, 适配时不直接更新 W_0 (冻结), 而是引入低秩分解矩阵 $A \in \mathbb{R}^{k \times r}$ (随机高斯初始化)和 $B \in \mathbb{R}^{d \times r}$ (零初始化), 权重更新表示为:

$$W_0 + \Delta W = W_0 + BA \quad (2)$$

若原始前向传播输出为 $h = W_0 x$, 则引入 LoRA 后的前向传播可表示为:

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad (3)$$

低秩适配(LoRA)这一高效的适配策略既不会引入推理延迟, 也不会缩短输入序列长度, 同时还能保持模型性能。重要的是, 在作为服务部署时, LoRA 可通过共享模型的绝大部分参数, 实现快速的任务切换[34]。

3.2.4. 基于提示词的情感增强机制

本文设计情感增强机制是基于提示词工程的情感增强机制, 核心是通过优化提示指令实现情感特征与新闻语义特征的高效融合, 引导基础模型聚焦情感与文本语义的关联信息, 强化虚假新闻检测判别力。该机制无需改动模型底层结构, 仅通过指令优化激活模型对情感特征的捕捉能力, 与 LoRA 微调形成协同, 提升检测精度。

结合情感标注结果与虚假新闻检测任务的需求, 本文使用的模板如下“结合新闻文本及其情感特征判断是否为虚假新闻。新闻文本: [文本内容]; 细粒度情感特征: [情感类型 1, 置信度 1]、[情感类型 2, 置信度 2]、[情感类型 3, 置信度 3]。输出: 判定结果。”

4. 实验设计

4.1. 数据集

4.1.1. 数据集概述

本实验使用 sem_eval_2018 数据集训练情感分析模型[35], 使用 WELFake_Dataset_Edited 数据集作为基准数据集进行模型微调和测试[36]。

Affect in Tweets (推文中的情感数据集)是 SemEval-2018_Task_1 中使用的数据集, 该数据集围绕推文的情感与情绪分析构建, 包含可用于不同子任务的五种子数据集, 本文使用的是其中的情感分类任务子数据集。一共包含 22,458 条数据, 其中 12,677 条训练集, 2150 条验证集, 7631 条测试集, 每一条数据

内容为: 给定一条推文, 判断其属于 11 种特定情感中的 1 种或多种(多标签分类, 即一条推文可同时标注多种情感), 用于识别发推者的具体情感类型组合。

WELFake_Dataset_Edited 是托管在 Hugging Face 平台上的一个新闻文本分类数据集, 核心用途是用于训练和评估“虚假新闻检测”相关的机器学习模型。共包括 72,133 条数据, 其中 35,024 条真实新闻、37,109 条虚假新闻。原始数据未划分训练集与验证集, 每一条数据内容为: 给定一条推文及其标题(部分推文标题部分为 null), 判断其是否为虚假新闻, 如果是则 label 标签为 1。

4.1.2. 数据集预处理

在对实验涉及的数据集完成常规数据清洗基础上, 需针对 WELFake_Dataset_Edited 数据集进一步开展切分与二次处理, 具体流程分为两步:

1) 数据集初次切分: 将 WELFake_Dataset_Edited 数据集按 7:3 的比例划分为两部分, 其中 70% 作为训练/验证集, 用于后续模型训练与参数调优; 30% 作为独立测试集, 用于最终验证模型泛化性能。

2) 数据集二次处理与再切分: 将上述步骤得到的训练/验证集输入已训练完成的 BERTemotions 模型, 由模型为该部分数据自动标记情感标签, 生成包含新闻文本、真实性标签及情感标签的新数据集, 命名为 NewsEmotions。随后, 将 NewsEmotions 数据集按 6:1 的比例再次划分为训练集与验证集, 以适配后续虚假新闻检测模型的训练与验证需求。

4.2. 模型训练设计

本文的模型训练设计步骤如下:

1) 利用 Affect in Tweets 数据集(见 4.1)对 BERT 模型进行微调训练, 预训练的 BERT 模型文本编码维度为 768 维, 实验设置 batch size 为 16, 注意力头数 h 为 12, 训练轮次为 3 轮并采用提前停止策略以防止模型过拟合。使用 ReLU 激活函数, 为了得到模型最优参数, 使用 Adam 作为优化器, 将微调后的模型作为 SELLM-FND 的情感分析模块。

2) 用 SELLM-FND 情感分析模块批量处理 WELFake_Dataset_Edited 数据集(见 4.1)获得包含情感标签的数据集 NewsEmotions。该数据集包含三个核心字段: 新闻文本(标题 + 摘要)、情感标签集合(多标签)、真实性标签, 为下一步融合情感特征的虚假新闻检测模型训练提供了关键数据支撑。

3) 使用包含情感标签的数据集 NewsEmotions 微调基础模型(见 3.2.2), 该模型微调采用 LoRA 微调(LORA 的秩为 16, 缩放因子为 32, 参数更新总量为 0.05), 实验设置 batch size 为 8, 且每累积 4 次小批次梯度后再更新一次模型参数, 注意力头数 h 为 32 头, 训练轮次为 3 轮并采用提前停止策略以防止模型过拟合。使用 ReLU 激活函数, 为了得到模型最优参数, 使用 Adam 作为优化器。将微调后的模型作为 SE-FND 检测模块。

4.3. 对比实验结果

我们将根据 4.2 中设计的模型(SELLM-FND)进行训练和微调。并对比该模型与现有的虚假新闻检测方法, 包括: 基于 BERT 模型微调的虚假新闻检测模型[6]、用未引入新闻情感标签微调的 DeepSeek-R1 大模型[32]、只基于细粒度情感分析进行新闻检测的模型 FOREAL [37]、基于传统情感分析方法的虚假新闻检测模型 EmoSentBERT [7], 基于 LLM-GAN 提示框架的可解释虚假新闻检测模型 LLM-GAN [38]。测试使用的数据集为 WELFake_Dataset_Edited 数据集的测试集部分。

本文评估指标包括: 准确率、精度、召回率以及 F1 指数。准确率是最直观的用于衡量预测假新闻和真实假新闻之间的相似性的指标, 利用精度、召回率以及 F1 指数为假新闻检测提供整体预测性能, 在假新闻检测中, 这四个指标结合使用, 能更全面地反映模型的实际效果。

Table 2. Comparison of experimental results of various detection methods
表 2. 各种检测方法的实验结果对比

模型	简述	Accuracy	Precision	召回度	F1 指数
BERT_detection	基于 BERT 的纯文本的检测模型, 仅基于语义进行检测, 属于较早期的新闻检测方法。	0.765	0.857	0.833	0.845
FOREAL	基于情感向量的检测模型, 根据细粒度情感分析的结果进行判断	0.700	0.788	0.839	0.813
EmoSentBERT	结合了情感维度和语义维度的预训练检测模型	0.886	0.912	0.912	0.912
LLM-GAN	基于对抗提示机制的虚假新闻检测大模型	0.916	0.922	0.913	0.917
DeepSeek-R1_detection	在基础模型上进行直接微调的检测模型	0.925	0.943	0.943	0.943
SELLM-FND	本文设计的检测模型	0.929	0.952	0.944	0.948

对比结果如表 2 所示, 仅基于细粒度情感分析的 FOREAL 模型与仅基于 BERT 微调的 BERT_detection 模型检测效果最差, 推测原因是二者均依赖单一维度特征(FOREAL 依赖情感特征, BERT_detection 依赖语义特征), 特征表征的全面性不足, 导致检测性能受限。相比之下, 融合情感维度与语义维度双特征的 EmoSentBERT 模型, 检测效果显著优于上述两类单维度模型, 验证了多维度特征融合对提升虚假新闻检测性能的有效性。

此外, 未引入情感变量的两类大模型方法(LLM-GAN 与 DeepSeek-R1_detection), 性能均优于传统预训练融合模型。其中, 基于对抗提示机制的 LLM-GAN 模型, 凭借大模型的逻辑推理与对抗训练优势, 实现了 0.916 的准确率, 但未融入情感特征, 其性能仍落后于引入情感增强的模型; DeepSeek-R1_detection 模型(基于基础模型 DeepSeek-R1-Distill-Qwen-7B-News-Classifer 针对虚假新闻检测任务进一步微调所得)虽未引入情感分析模块, 但其检测效果仍优于 FOREAL、BERT_detection 与 EmoSentBERT 三类模型, 这体现出了基础大模型在语义理解与任务适配方面的天然优势。

本文设计的 SELLM-FND 模型在 DeepSeek-R1_detection 的基础上增加了情感增强机制, 其检测性能在所有评估指标中均表现最佳, 各项指标均高于其他对比模型, 且在精度与 F1 指数上的优势尤为突出。这表明, SELLM-FND 模型在虚假新闻检测任务中具备更精准的预测能力与更均衡的综合性能, 同时验证了情感增强机制对提升大模型虚假新闻检测效果的积极作用。

实验结果说明, 单一的语义维度和情感维度的检测模型相比于能够结合两种维度的模型都相对落后, 而单一的基于情感维度的模型检测也落后于基于语义的基础 BERT 模型检测, 大模型的简单微调后的检测结果相比普通的预训练模型更好, 基于情感增强机制的大语言模型虚假新闻检测方法, 无论是比预训练模型还是未使用情感增强机制的大模型性能都更好。

4.4. 消融实验

为验证 SELLM-FND 模型各核心模块对虚假新闻检测性能的贡献度, 明确不同模块的作用价值, 本研究设计了 2 组消融实验。所有实验均基于 WELFake_Dataset_Edited 数据集的测试集开展, 沿用准确率、精度、F1 指数作为核心评估指标, 同时引入性能下降率与模块贡献度量化模块价值, 实验结果如表 3 所示。

情感模块是性能提升的核心贡献源: 消融组 1 移除情感模块后, F1 指数显著下降 2.95%, 模块贡献度高达 38.2%, 为所有模块中最高。这一结果充分证明, BERTemotions 情感模块提取的细粒度情感特征是 SELLM-FND 模型优于对比模型的核心因素。该模块提供的“情感类型 + 置信度”双重信息, 能够精准捕捉虚假新闻常用的情感煽动、情绪误导等操纵策略, 有效弥补了单一语义特征在情感维度信息捕捉

上的不足, 为虚假新闻的精准识别提供了重要支撑。

跨注意力融合强化特征关联性: 消融组 2 用“语义 + 情感直接拼接”替代跨注意力融合机制后, F1 指数下降 1.58%, 贡献度为 20.4%。这表明, 简单的特征拼接无法实现情感特征与语义特征的深度交互, 易导致特征冗余或关键关联信息丢失; 而跨注意力机制能够动态挖掘两类特征间的内在关联, 根据任务需求自适应分配特征权重, 使融合后的特征更具判别力, 从而提升模型的检测性能。

Table 3. Comparison of ablation experiment results

表 3. 消融实验结果对比

实验组别	模型配置(基于 SELLM-FND 基准)	准确率 (Accuracy)	精度 (Precision)	F1 指数	性能 下降率(%)	模块 贡献度(%)
基准组 (SELLM-FND)	情感模块(BERTemotions) + 跨注意力融合 + LoRA 微调	0.929	0.952	0.948	-	-
消融组 1 (无情感模块)	移除 BERTemotions, 仅用 DeepSeek-R1 + LoRA 微调	0.901	0.918	0.920	2.95	38.2
消融组 2 (无跨注意力)	保留情感模块, 采用“语义 + 情感直接拼接”	0.915	0.935	0.933	1.58	20.4

注: 模块贡献度 = 该组 F1 下降率/所有消融组总 F1 下降率。

综上, SELLM-FND 模型的两大核心模块(情感模块、跨注意力融合)均对性能提升产生关键作用, 其中情感模块的核心价值尤为突出, 跨注意力融合保障了特征融合质量, 二者与 LoRA 微调策略协同作用, 共同实现了虚假新闻检测性能的最优表现。

5. 总结

为了解决过往检测方法对于情感特征利用不足的问题, 本文提出了一种基于情感增强机制的大语言模型虚假新闻检测方法, 并基于该方法设计了 SELLM-FND 模型, 通过结合文本的语义信息和情感信息的方法, 在基于文本的虚假新闻检测任务中实现了对现有主流方法的超越, 验证了其有效性和优越性。这证明大模型在虚假新闻检测领域比小模型更优秀, 也证明了合理使用细粒度情感分析的方法在虚假新闻检测领域能够有效提高模型检测准确度和稳定性。

面对现有新闻数据集缺少情感标注的问题, 本文提出的“先训练情感分析模型, 再利用该模型为数据集添加情感标注, 进而辅助核心模型训练”的方案, 不失为一种有效的解决办法。

但需注意, 单一语义维度或情感维度的检测模型性能均落后于双维度融合模型, 而本文方法仍属于基于文本的单一模态检测方法, 在面对包含文本、图像、音频和视频的多媒体新闻时的能力尚未得到验证。但即使能力可能有所不足, SELLM-FND 模型作为基于文本的虚假新闻检测模型也可作为多模态虚假新闻检测中的文本模态处理方法, 加入到多模态的虚假新闻检测中。

基金项目

北京市教育委员会科研计划项目资助(KM202410015002);
北京印刷学院博士启动资金(27170123034、27170124026)。

参考文献

- [1] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H. (2017) Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19, 22-36. <https://doi.org/10.1145/3137597.3137600>
- [2] Capuano, N., Fenza, G., Loia, V. and Nota, F.D. (2023) Content-Based Fake News Detection with Machine and Deep

- Learning: A Systematic Review. *Neurocomputing*, **530**, 91-103. <https://doi.org/10.1016/j.neucom.2023.02.005>
- [3] Hirlekar, V.V. and Kumar, A. (2020) Natural Language Processing Based Online Fake News Detection Challenges—A Detailed Review. 2020 *5th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, 10-12 June 2020, 748-754. <https://doi.org/10.1109/icc48766.2020.9137915>
- [4] Shen, Y., Liu, Q., Guo, N., Yuan, J. and Yang, Y. (2023) Fake News Detection on Social Networks: A Survey. *Applied Sciences*, **13**, Article 11877. <https://doi.org/10.3390/app132111877>
- [5] Phan, H.T., Nguyen, N.T. and Hwang, D. (2023) Fake News Detection: A Survey of Graph Neural Network Methods. *Applied Soft Computing*, **139**, Article ID: 110235. <https://doi.org/10.1016/j.asoc.2023.110235>
- [6] Ahn, Y. and Jeong, C. (2019) Natural Language Contents Evaluation System for Detecting Fake News Using Deep Learning. 2019 *16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Chonburi, 10-12 July 2019, 289-92. <https://doi.org/10.1109/jcsse.2019.8864171>
- [7] Upadhayay, B. and Behzadan, V. (2020) Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification. 2020 *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Arlington, 9-10 November 2020, 1-6. <https://doi.org/10.1109/isi49825.2020.9280528>
- [8] Yuan, C., Qian, W., Ma, Q., Zhou, W. and Hu, S. (2021) SRLF: A Stance-Aware Reinforcement Learning Framework for Content-Based Rumor Detection on Social Media. 2021 *International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, 18-22 July 2021, 1-8. <https://doi.org/10.1109/ijcnn52387.2021.9533444>
- [9] Zhang, X. and Ghorbani, A.A. (2020) An Overview of Online Fake News: Characterization, Detection, and Discussion. *Information Processing & Management*, **57**, Article ID: 102025. <https://doi.org/10.1016/j.ipm.2019.03.004>
- [10] Nasir, J.A., Khan, O.S. and Varlamis, I. (2021) Fake News Detection: A Hybrid CNN-RNN Based Deep Learning Approach. *International Journal of Information Management Data Insights*, **1**, Article ID: 100007. <https://doi.org/10.1016/j.jiime.2020.100007>
- [11] Dimpas, P.K., Po, R.V. and Sabellano, M.J. (2020) Filipino and English Clickbait Detection Using a Long Short Term Memory Recurrent Neural Network. 2017 *International Conference on Asian Language Processing (IALP)*, Singapore, 5-7 December 2017, 276-280.
- [12] Telang, H., More, S., Modi, Y. and Kurup, L. (2019) Anempirical Analysis of Classification Models for Detection of Fake News Articles. 2019 *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, 20-22 February 2019, 1-7. <https://doi.org/10.1109/icecct.2019.8869504>
- [13] Farokhian, M., Rafe, V. and Veisi, H. (2023) Fake News Detection Using Dual BERT Deep Neural Networks. *Multimedia Tools and Applications*, **83**, 43831-43848. <https://doi.org/10.1007/s11042-023-17115-w>
- [14] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., et al. (2024) Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 22105-22113. <https://doi.org/10.1609/aaai.v38i20.30214>
- [15] Ke, J. (2024) An Implicit Semantic Enhanced Fine-Grained Fake News Detection Method Based on Large Language Models. *Journal of Computer Research and Development*, **61**, 1250-1260.
- [16] Hussein, D.M.E.M. (2018) A Survey on Sentiment Analysis Challenges. *Journal of King Saud University—Engineering Sciences*, **30**, 330-338. <https://doi.org/10.1016/j.jksues.2016.04.002>
- [17] Kalbhor, S. and Goyal, D. (2023). Survey on ABSA Based on Machine Learning, Deep Learning and Transfer Learning Approach. *AIP Conference Proceedings*, **2782**, Article ID: 020041. <https://doi.org/10.1063/5.0154549>
- [18] Zhong, B. (2024) Fine-grained Sentiment Analysis Using Multidimensional Feature Fusion and GCN. *Journal of Information and Telecommunication*, **9**, 91-112. <https://doi.org/10.1080/24751839.2024.2386785>
- [19] Nkhata, G., Gauch, S. and Anjum, U. (2025) Fine-Tuning BERT with Bidirectional LSTM for Fine-Grained Movie Reviews Sentiment Analysis. arXiv: 2502.20682.
- [20] Teng, J., He, H. and HU, G. (2025) A Fine-Grained Sentiment Recognition Method for Online Government-Public Interaction Texts Based on Large Language Models. *International Conference on Artificial Intelligence and Machine Learning Research (CAIMLR 2024)*, Singapore, 28-29 September 2024. <https://doi.org/10.1117/12.3058735>
- [21] Gu, J.W. (2025) A Survey on LLM-As-A-Judge. arXiv: 2411.15594.
- [22] Boissonneault, D. and Hensen, E. (2024) Fake News Detection with Large Language Models on the LIAR Dataset. <https://doi.org/10.21203/rs.3.rs-4465815/v1>
- [23] Han, Z.Y. (2024) Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv: 2403.14608.
- [24] 王东清, 芦飞, 张炳会, 等. 大语言模型中提示词工程综述[J]. 计算机系统应用, 2025, 34(1): 1-10.
- [25] Reynolds, L. and McDonnell, K. (2021) Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, 8-13 May

-
- 2021, 1-7. <https://doi.org/10.1145/3411763.3451760>
- [26] Kojima, T., *et al.* (2022) Large Language Models are Zero-Shot Reasoners. arXiv: 2205.11916.
- [27] Yao, S.Y., *et al.* (2023) ReAct: Synergizing Reasoning and Acting in Language Models. arXiv: 2210.03629.
- [28] Yao, S.Y., *et al.* (2023) Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv: 2305.10601.
- [29] Besta, M., *et al.* (2023) Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv: 2308.09687.
- [30] Zhou, Y.C., Muresanu, A., Han, Z.W., *et al.* (2023) Large Language Models Are Human-Level Prompt Engineers. International Conference on Learning Representations. <https://iclr.cc/virtual/2023/10850>
- [31] Lewis, P., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv: 2005.11401.
- [32] Etaiwi, W. and Alhijawi, B. (2025) Comparative Evaluation of ChatGPT and DeepSeek across Key NLP Tasks: Strengths, Weaknesses, and Domain-Specific Performance. *Array*, **27**, Article ID: 100478. <https://doi.org/10.1016/j.array.2025.100478>
- [33] Real-Jiakai (2024) Deepseek-R1-Distill-Qwen-7B-News-Classifer. Hugging Face. <https://huggingface.co/real-jiakai/Deepseek-R1-Distill-Qwen-7B-News-Classifer>
- [34] Hu, E. (2021) LoRA: Low-Rank Adaptation of Large Language Models. arXiv: 2106.09685.
- [35] SemEvalWorkshop (2024) Sem_eval_2018_task_1 Dataset. Hugging Face. https://huggingface.co/datasets/SemEvalWorkshop/sem_eval_2018_task_1
- [36] Summitsky (2024) WELFake_Dataset_Edited Dataset. Hugging Face. https://huggingface.co/datasets/Summitsky/WELFake_Dataset_Edited
- [37] Kolev, V., Weiss, G. and Spanakis, G. (2022) FOREAL: Roberta Model for Fake News Detection Based on Emotions. *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, 3-5 February 2022, 429-440. <https://doi.org/10.5220/0010873900003116>
- [38] Wang, Y., Gu, Z., Zhang, S., Zheng, S., Wang, T., Li, T., *et al.* (2025) LLM-GAN: Constructing Generative Adversarial Network through Large Language Models for Explainable Fake News Detection. *ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, 6-11 April 2025, 1-5. <https://doi.org/10.1109/icassp49660.2025.10889048>