

# 基于多元评论和三重视角的多模态 虚假新闻检测

虞永毅<sup>1)</sup> 肖聪<sup>2)</sup> 王明文<sup>1)</sup> 黄琪<sup>1)</sup> 罗文兵<sup>1)</sup> 朱莹婷<sup>3)</sup>

<sup>1)</sup>(江西师范大学人工智能学院 南昌 330022)

<sup>2)</sup>(江西师范大学财务处 南昌 330022)

<sup>3)</sup>(南昌师范学院数学与信息科学学院 南昌 330032)

**摘要** 随着互联网和新媒体技术的发展,虚假新闻的传播变得更为迅速和广泛。现有检测方法主要存在两种问题:(1)大多数利用了生成评论的检测方法只引入单一来源或类型的评论数据作为训练与测试样本,未能充分覆盖真实场景下评论的多元化特征导致模型泛化能力较弱;(2)现有借助了大语言模型的检测方法则存在提取的新闻语义信息容易受到幻觉影响的局限性。为此,本文提出基于多元评论和三重视角的多模态虚假新闻检测方法,该方法通过动态评论交互融合模块来减轻大语言模型生成评论中的幻觉,并聚焦真实评论的重要信息,结合两者的有效内容来辅助虚假新闻检测任务。同时使用自我审查、同行评估和综合评测的三重视角来有效降低大语言模型在使用时产生的部分幻觉,新闻信息分析融合模块利用了CLIP模型来分别对齐新闻文本、图像和大语言模型新闻分析信息来得到文本、图像模态信息,再通过交叉注意力机制实现了文本模态和图像模态的跨模态语义特征融合。在Weibo、Weibo21、GossipCop等公开的真实数据集中的实验结果显示,准确率分别达到95.1%、93.1%和88.3%,比最新基线方法分别提高5%、1.4%和2.7%,比大语言模型基线方法分别提高13.2%、15.2%和16.6%。

**关键词** 虚假新闻检测;大语言模型;多视角;多模态;语义融合

中图法分类号 TP18 DOI号 10.11897/SP.J.1016.2026.00591

## Multimodal Fake News Detection Based on Multiple Comments and Triple Perspectives

YU Yong-Yi<sup>1)</sup> XIAO Cong<sup>2)</sup> WANG Ming-Wen<sup>1)</sup> HUANG Qi<sup>1)</sup> LUO Wen-Bing<sup>1)</sup> ZHU Ying-Ting<sup>3)</sup>

<sup>1)</sup>(School of Artificial Intelligence, Jiangxi Normal University, Nanchang 330022)

<sup>2)</sup>(Finance Office, Jiangxi Normal University, Nanchang 330022)

<sup>3)</sup>(School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032)

**Abstract** With the rapid development of the Internet and new media technologies, the dissemination of news posts has become faster and more widespread than ever before. While these advancements provide people with unprecedented access to information, they have also introduced numerous hidden dangers

收稿日期:2025-07-01;在线发布日期:2025-11-12。本课题得到国家自然科学基金(62266023、62466028)、江西省自然科学基金(20242BAB20045)、江西省高等教育学会2023年度学会课题(ZX1-B-001)、江西省教育厅科学技术研究项目(GJJ2401903)的部分资助。虞永毅,硕士研究生,主要研究领域为自然语言处理、虚假新闻检测。E-mail: yuyongyi@jxnu.edu.cn。肖聪,硕士,主要研究领域为自然语言处理、虚假新闻检测。王明文(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息抽取、信息检索、数据挖掘。E-mail: mwwang@jxnu.edu.cn。黄琪,博士,讲师,主要研究领域为社交网络分析、虚假新闻检测、情感分析。罗文兵,博士,高级实验师,主要研究领域为教育知识图谱、自然语言处理、个性化推荐。朱莹婷,硕士,讲师,主要研究领域为自然语言处理、计算机教育。

and challenges. A significant issue is the proliferation of unverified or deliberately fabricated false information, often created by individuals or groups with ulterior motives. Such misleading content is frequently mixed with legitimate social news, leading to widespread misinformation and causing substantial negative impacts on society. Consequently, accurately and efficiently identifying false news has become a critical and urgent task. Existing false news detection methods using large language models either predict the underlying false news by generating comments or analyze the context consistency of news content. Although these methods have significantly improved the detection efficiency of false news by utilizing text, comments, and other information and provided valuable experience for subsequent research, they still have some unsolved problems. For instance, most of the detection methods that utilize generated comments only introduce a single source or type of comment data as training and testing samples, failing to fully cover the diversified features of comments in real scenarios, which leads to weak generalization ability of the models. Moreover, compared with the weak generalization ability and difficulty in dealing with complex scenarios of traditional methods, the existing detection methods relying on large language models have the limitation that the extracted news semantic information is prone to be affected by hallucinations. In fact, both the comments generated by large language models and real comments have their own advantages. The combination of the two types of comment information can, to a certain extent, make up for each other's shortcomings. Therefore, this paper proposes a multi-modal false news detection method based on multi-angle comments and three perspectives. This method designs a dynamic comment interaction fusion module, which includes an adaptive comment aggregator that can effectively mitigate the hallucinations in the comments generated by large language models, and a self-attention mechanism that focuses on the important information in real comments. By combining the effective content of both, it assists in fake news detection task. A single perspective is always limited, while multiple perspectives often provide a more comprehensive view of the problem. Therefore, inspired by year-end summaries and background checks, this paper uses self-review, peer assessment, and comprehensive evaluation as three perspectives to effectively reduce the hallucinations produced by large language models during use. The news information analysis and fusion module utilizes the CLIP model to align the news text, images and the news analysis information generated by large language models to obtain separated feature representations for the text and image modalities, and then achieves cross-modal semantic feature fusion between the text and image modalities through a cross-attention mechanism. Experimental results on public real datasets such as Weibo, Weibo21, and GossipCop show that the accuracy rates reach 95.1%, 93.1%, and 88.3%, respectively, which are 5%, 1.4%, and 2.7% higher than the latest baseline methods, and 13.2%, 15.2%, and 16.6% higher than the large language model baseline methods.

**Key words** fake news detection; large language model; multi-view; multimodal; semantic fusion

## 1 引 言

互联网与新媒体技术的蓬勃发展为信息传播构建了多元化平台与交互式交流通道,但是在这一过程中不可避免地滋生了大量虚假信息。依托互联网的即时传输特性和跨地域覆盖能力,虚假新闻及其衍生危害得以迅速传播,加剧了社会认知偏差,

进一步增加了社会治理的不稳定因素,甚至可能会因此扰乱经济秩序,威胁国家安全。例如,一篇有关于美国白宫爆炸的虚假新闻帖子在社交媒体上引起了恐慌,导致道琼斯指数在短短两分钟内下跌100点<sup>[1]</sup>。

虚假新闻是指那些被制作出来用于传播不实信息或误导性信息的新闻,它可能是凭空捏造的,

抑或是对已有事实进行夸大或歪曲<sup>[2]</sup>。这些新闻往往借助煽动性标题、断章取义的内容和伪造的信息误导民众，并且由于大语言模型的出现，为不法分子制作虚假新闻提供了更多的技术手段，内容更加丰富多样，危害性也更大<sup>[3-4]</sup>。因此，开展虚假新闻检测技术的研究，不仅能够有效维护网络空间的安全有序与和谐稳定，还可以为政府部门、各类企业以及社会民众提供精准可靠的信息验证工具。

传统的检测方法主要依赖于从不同模态输入中提取的内容和风格特征：Nan 等人<sup>[5]</sup>提出从新闻评论中提取知识并将其参数化后注入检测模型来提高模型在不同语境中的泛化能力。Cui 等人<sup>[6]</sup>通过深入探讨用户情绪表达与新闻可信度的关联性机制，来挖掘用户评论中的情感极性、情绪强度等深层特征，有效地构建了虚假新闻与真实内容的情感区分模型。亓鹏等人<sup>[7]</sup>则依靠模型中隐含的事实知识与提取的实体对象来建模虚假新闻的多模态深层语义信息。

近几年的大语言模型(Large Language Models)凭借着庞大的语料库成为了新一代人工智能技术的代表<sup>[8-10]</sup>。它展现出的卓越跨任务泛化能力和知识推理整合能力能够为虚假新闻检测提供强劲动力<sup>[11]</sup>。Hu 等人<sup>[12]</sup>通过设计一个自适应原理指导网络来让小语言模型(Small Language Models)有选择性地从 LLMs 的网络中获取新闻分析的见解与核查。Nan 等人<sup>[13]</sup>则研究出了一个生成式的反馈检测框架 GenFEND 通过让大语言模型作为用户模拟器和评论生成器来弥补那些本该由沉默用户提出的评论，然后提取评论信息用于虚假新闻帖子的预测。

尽管上述模型在部分任务中展现出了良好的性能，但仍然存在以下不足之处：1) 现有基于大语言模型生成评论的方法还存在一定的局限性，只利用了单一评论，未能充分考虑多元化的评论信息，这使得检测模型的泛化性能还有提升的空间。2) 大语言模型缺少特定任务的知识并且会产生一部分幻觉，所以无法完全取代传统的检测模型来直接判断虚假新闻<sup>[12]</sup>，同时幻觉会影响大语言模型对新闻的分析结果，所以如何有效缓解或减轻幻觉对虚假新闻检测模型带来的影响变得越来越重要。

针对上述两个问题，本文提出一种基于多元评论和三重视角的多模态虚假新闻检测方法。该方法首先通过提示工程(prompt engineering)让大语言模型来模拟用户对新闻内容生成特定的评论，结合

新闻数据集自带的真实评论通过 BERT<sup>[14]</sup>来编码文本语义特征和评论语义特征，将它们输入一个动态评论交互融合模块，再通过交叉注意力机制得到文本与评论的语义交互特征。然后继续使用大语言模型，运用它丰富的背景知识和上下文解析能力进行新闻纠错和分析，得到相应的新闻信息分析报告。同时使用 CLIP<sup>[15]</sup>和交叉注意力机制与 Swin-T<sup>[16]</sup>编码的图像特征和 BERT 编码的文本特征进行跨模态语义融合来提高虚假新闻检测的准确率。本文的贡献主要有：

(1) 提出了一种融合多元评论和增强跨模态语义理解的虚假新闻检测框架。相比于使用单一评论的模型，该方法既保留了拥有人类思维和真情实感的真实评论又拥有比较完整和规范的生成评论，有效提升了基于以评论数据为主的虚假新闻检测的泛化能力和准确率；

(2) 提出了三重视角的检测方法。受工作报告和背景审查的启发提出了自我审查、同行评估和综合评测的三重视角检测方法，并且提取大语言模型对上下文的背景知识审核与新闻内容分析用于图文模态的深层语义，既可以减轻一部分幻觉问题，也可以挖掘新闻内容的深层信息；

(3) 基于三个公开基准数据集的系统性实验评估显示，本研究所提出的方法在关键性能指标(如准确率、精确度、召回率、F1 分数)上均取得了较好的结果，并且超过了现有主流的大语言基准模型。

## 2 相关工作

随着移动社交平台的增多，海量掺杂着虚假信息的新闻帖子在各个社交网络上传播，为社会发展埋下了不稳定的因素。虚假新闻检测作为一个二元分类任务，从技术角度上来区分，可以分为基于传统方法的虚假新闻检测研究和基于大语言模型方法的虚假新闻检测研究。

### 2.1 基于传统方法的虚假新闻检测研究

早期的方法主要通过研究者人工设计的特征和样式，在此基础上运用机器学习思想来进行虚假新闻甄别<sup>[17]</sup>。比如 Zhao 等人<sup>[18]</sup>提出了一种基于决策树分类器的检测模型，通过从文本信息中提取质疑类和更正类短语特征，然后分析特征之后整合归类，随后基于统计特征构建决策树分类模型来进行虚假新闻检测。Ze 等人<sup>[19]</sup>提出了一个新闻评论生成的“阅读-参与-评论过程”，并使用阅读网络和生成

网络来形式化该过程。然而,依靠单一模态的检测容易因为缺失多源信息印证而被“局部真实”内容蒙蔽,易导致模型过拟合。

多模态虚假新闻检测通过整合文本、图像、音视频等多元数据,从跨模态一致性、互补性等角度挖掘深层虚假信息线索来提高检测的准确率。在评论特征提取方面,以 Transformers<sup>[8]</sup>架构为基础的预训练语言模型(如 BERT、RoBERTa<sup>[19]</sup>等)展现出卓越的语义表征能力;在图像特征提取方面,深度卷积神经网络(如 VGG19<sup>[20]</sup>、ResNet<sup>[21]</sup>等)则能够有效捕获图像的多层次视觉特征。Zhang 等人<sup>[22]</sup>创造性地提出了跨模态注意力融合框架,该方法通过建立文本-图像双向注意力权重矩阵,实现了帖子内部图文特征的动态对齐与交互式融合。传统方法的检测面临知识面太过单一、对虚假新闻的反应太慢等问题。

## 2.2 基于大语言模型方法的虚假新闻检测研究

大语言模型(LLMs)凭借强大的语义理解、上下文推理和多任务的泛化能力,为虚假新闻检测提供了新的思路。大语言模型的优势在于可解析新闻文本中的行文逻辑矛盾、隐含的政治性偏见以及生成式 AI(Artificial Intelligence)伪造的虚假内容,可以实现深层次的语义关联挖掘;而以 CLIP、LLaVA<sup>[23]</sup>等为代表的多模态大语言模型通过对齐文本、音频、图像等模态的语义空间来执行跨模态一致性验证任务。Xuan 等人<sup>[24]</sup>提出了一个大视觉语言模型(LEMMA)利用 LVLM<sup>[25]</sup>的直觉推理能力,以外部知识来强化它们的虚假新闻检测能力。Wan 等人<sup>[26]</sup>则制作了一个大模型专家网络 DELL 专门用来研究理解各个方面新闻的专家,通过大语言模型强大的知识储备来判断新闻帖子的真实性。柯婧等人<sup>[27]</sup>提出了利用大语言模型来挖掘虚假新闻的隐含语义信息并按主次事件来进行分类。除了文本分析能力,大语言模型还具备了文本生成的能力,甚至在效果上不输于人工生成的文本。Liu 等人<sup>[28]</sup>引入了一个基于大模型的虚假新闻传播仿真框架(FPS)来详细研究虚假新闻传播的趋势和控制。Ignat 等人<sup>[29]</sup>制作了一个数据集 Maide-up,其中包括了 10000 条真实的酒店评论和 10000 条用大模型生成的虚假酒店评论,从三个维度来探索几种用于酒店评论欺骗检测模型的有效性。Wang 等人<sup>[30]</sup>从社会心理学角度制作了一个全面基于大模型的理论框架(LLM-

Fake Theory)让大模型自动生成虚假新闻,并由此创建了一个机器生成的虚假新闻数据集(Mega-Fake)来为大模型时代的虚假新闻检测和治理提供自己的经验。上述方法通过有效利用大语言模型技术,不仅提高了虚假新闻检测的效率,还给后续的研究提供了可借鉴的思路。

## 3 模型

### 3.1 问题定义

虚假新闻检测的核心任务通常被建模为区分真实新闻和虚假新闻的二分类问题<sup>[31]</sup>。给定一组新闻  $N = \{D_1, D_2, \dots, D_n\}$ , 其中每条新闻  $D_i$  包含文本  $T_i$ 、评论  $C_i$  和图像  $I_i$ , 在训练集中标注其真实性标签  $y_i$ ,  $y_i \in \{0, 1\}$  (1 表示为真实新闻, 0 表示为虚假新闻)。多模态虚假新闻检测的目标是训练一个分类模型  $f$ , 使其能够基于文本、评论和视觉信息的联合表示  $D_i = \{T_i, C_i, I_i\}$  预测新闻的真实性  $y_i$ 。

### 3.2 模型框架

基于多元评论和三重视角的多模态虚假新闻检测方法的整体架构如图 1 所示, 该模型主要由数据处理模块、动态评论交互融合模块、新闻信息分析融合模块和分类器模块组成。下面详细介绍各个模块。

### 3.3 数据处理模块

为了能够克服传统虚假新闻检测方法中只依靠单一评论的局限性, 该模块通过大语言模型为每条新闻生成评论和多角度评估, 为后续的检测模块生成了高质量、多视角和经过初步校验的增强数据。

#### 3.3.1 大语言模型生成评论

首先使用智谱清言大语言模型(GLM-4-flash)<sup>[32]</sup>通过提示为每个新闻样本生成 8 个模拟用户评论。考虑到方案的成熟性, 在这里参考了 GenFEND<sup>[13]</sup>的提示模板, 把性别定为男性和女性; 年龄分别为 18 岁以下、18 到 29 岁、30 到 49 岁、50 到 64 岁和 64 岁以上; 受教育程度定为未接受过高等教育、正在接受高等教育和已经完成高等教育。由于用户对新闻内容的立场<sup>[33]</sup>以及宗教信仰<sup>[34]</sup>也会影响虚假新闻的判断, 所以将这两个因素也一同加入到模拟用户的属性中。图 2 为生成评论的提示模板。

#### 3.3.2 自我审查

大语言模型除了优秀的文本生成能力, 还有不俗的上下文逻辑分析能力和知识参考能力。在得到

相应的生成评论后继续使用智谱清言大语言模型对新闻内容和生成评论做第一轮的基础分析，得到

一些浅显的语义关系和关联信息，为下一步的分析做好准备。图3为自我审查的提示模板。

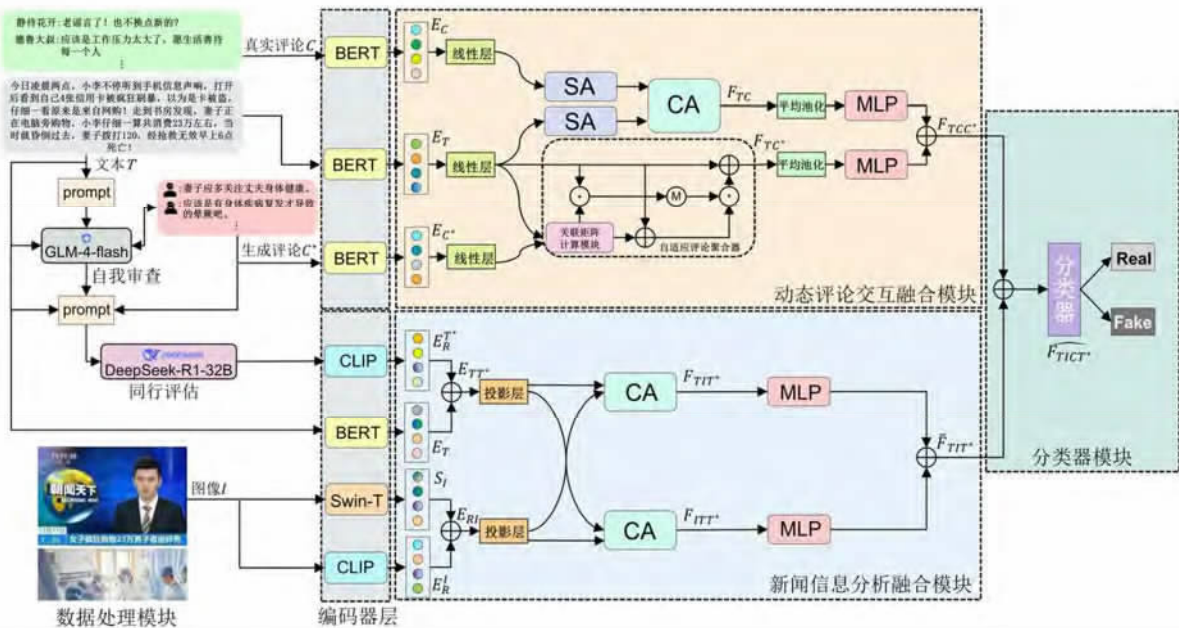


图1 系统模型图

**Prompt1: 生成评论**

你是一位[性别]用户，年龄[年龄]，受教育程度为[受教育程度]，宗教信仰为[宗教信仰]，对新闻内容的立场是[对新闻内容的立场]，请针对以下新闻内容发表一条不超过20字的评论：{text}

图2 生成评论的提示模板

**Prompt2: 自我审查**

这是新闻内容：{text}，和相应新闻的8个生成评论：{gen\_comments}。现在请你作为一个真假新闻分析专家来根据新闻内容和你自己生成的评论内容对这则新闻进行分析，可以根据新闻的来源、时间、背景、逻辑或常识来分析，不需要给出判断结果，只需要给出你的分析，字数限制在50字以内。

图3 自我审查的提示模板

### 3.3.3 同行评估

在虚假新闻检测领域，加入了大语言模型的检测模型相比于传统方法的检测模型，展现出了更强大的深层次语义理解和知识推理能力<sup>[35]</sup>。大语言模型偶尔会产生与用户要求输出内容不一致的部分，或与先前提供的内容产生矛盾，或与现实世界的知识相悖，这被称为幻觉问题<sup>[36]</sup>。幻觉的出现会严重影响大语言模型输出的内容，最终造成模型检测准确性的下降。因此，多重视角的评估检测是十分有必要的。通过大语言模型对自己输出内容的检查可以解决一部分最低级的幻觉，而同行大语言模型的二次独立评估，可以指出一些更难以发觉的幻觉。DeepSeek<sup>[37]</sup>大语言模型作为近年来国产大模型领域中一颗冉冉升起的新星，以更低的计算成本和更高

水准的处理性能逐渐受到人们的关注。在该模块中使用的是本地部署的 DeepSeek-R1-32B，拥有近320 亿参数，在分析虚假新闻逻辑、核查上下文一致性、减轻幻觉等方面存在一定优势。图4是同行评估的提示模板，在对新闻内容、生成评论和现有分析的审核评估后得到相应的分析报告，作为新闻分析数据，用于下一步综合评测的输入。

**Prompt3: 同行评估**

**System\_prompt:** 你是一个资深新闻事实核查专家，需要完成：

1. 结合新闻内容和用户评论分析信息一致性
2. 评估现有分析的合理性和潜在问题
3. 给出专业、中立、逻辑严谨的分析报告
4. 严格保持250字以内

**User\_prompt:** 新闻内容：{text} 生成评论：{gen\_comments} 现有分析：{analysis}

请按照以下框架分析：

1. 信息交叉验证（新闻与评论的一致性）
2. 现有分析优缺点（指出合理性和潜在偏见）
3. 你的专业意见（保持客观中立）

图4 同行评估的提示模板

### 3.4 动态评论交互融合模块

为了能够有效聚合两类评论的核心语义，并减轻生成评论中幻觉带来的不利影响，通过该模块来对新闻文本、真实用户评论和生成评论进行深度融合，提炼出最能揭示新闻真实性的综合特征表示。

#### 3.4.1 文本与真实评论特征融合

给定新闻文本  $T = [t_1, t_2, \dots, t_n]$  和真实评论  $C = [c_1, c_2, \dots, c_n]$  以及生成评论  $C^* = [c_1^*, c_2^*, \dots, c_n^*]$ ，其中  $n$  表示新闻文本和评论的长度，使用基于Transformer 架构的 BERT 编码器编码后的文本特

征向量和评论特征向量为  $\mathbf{E}_T = [T_f^1, T_f^2, \dots, T_f^n]$ 、 $\mathbf{E}_C = [C_f^1, C_f^2, \dots, C_f^n]$  和  $\mathbf{E}_C^* = [C_f^{*1}, C_f^{*2}, \dots, C_f^{*n}]$ ， $T_f^i \in R^{d_i}$  和  $C_f^i \in R^{d_i}$  表示文本和评论中第  $i$  个词的嵌入， $d_i$  是具体词嵌入的维度表示。将它们放入一个线性层 (Linear Layer) 降到共同的维度用来进行下一步操作。

由于自注意力机制可以捕捉上下文的长距离依赖关系<sup>[8]</sup>，通过计算序列中所有词元之间的权重，让模型更加“关注”那些信息量最大的部分(如新闻中的关键性证据、矛盾点等信息)。对新闻文本做自注意力操作可以提取其关键实体与核心主张，忽略大量的冗余信息；而对评论信息做自注意力操作则可以捕捉一些情绪极端化的用词，降低其权重，提高一些理性分析评论的权重。此外，面对一些虚假新闻，如果存在大部分质疑评论，而只有少部分支持评论，那么通过自注意力机制可以有效聚焦评论中的关键信息、建模群体一致性从而减少因为评论导致模型误判的问题。因此，我们将自注意力机制引入到 SA 层，如图 5 所示。具体首先将文本和真实评论的向量  $\mathbf{E}_T$  和  $\mathbf{E}_C$  分别输入到 SA 层中，以文本特征向量  $\mathbf{E}_T$  作为查询向量  $\mathbf{Q}$ 、键向量  $\mathbf{K}$  和值向量  $\mathbf{V}$ ，评论特征向量  $\mathbf{E}_C$  作为查询向量  $\mathbf{Q}$ 、键向量  $\mathbf{K}$  和值向量  $\mathbf{V}$ ，通过计算  $\mathbf{Q}$  和  $\mathbf{K}$  的相似度分数并应用 softmax 函数归一化，得到相应注意力权重系数，再使用这些权重系数加权求和值向量  $\mathbf{V}$ ，得到加权特征向量  $\text{Atten}(\mathbf{E}_T, \mathbf{E}_T)$  和  $\text{Atten}(\mathbf{E}_C, \mathbf{E}_C)$ 。

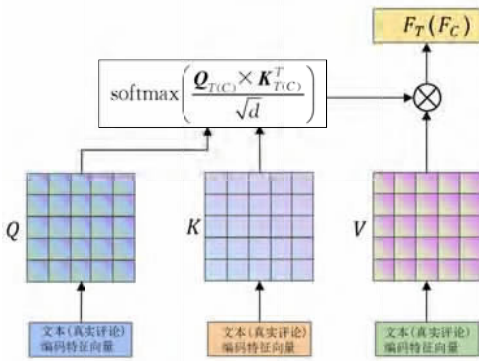


图 5 SA 层

$$\mathbf{Q}_T = \mathbf{E}_T \times \mathbf{W}_T^Q, \mathbf{K}_T = \mathbf{E}_T \times \mathbf{W}_T^K, \mathbf{V}_T = \mathbf{E}_T \times \mathbf{W}_T^V \quad (1)$$

$$\mathbf{Q}_C = \mathbf{E}_C \times \mathbf{W}_C^Q, \mathbf{K}_C = \mathbf{E}_C \times \mathbf{W}_C^K, \mathbf{V}_C = \mathbf{E}_C \times \mathbf{W}_C^V \quad (2)$$

$$\text{Atten}(\mathbf{E}_T, \mathbf{E}_T) = \text{softmax}\left(\frac{\mathbf{Q}_T \times \mathbf{K}_T^T}{\sqrt{d}}\right) \times \mathbf{V}_T \quad (3)$$

$$\text{Atten}(\mathbf{E}_C, \mathbf{E}_C) = \text{softmax}\left(\frac{\mathbf{Q}_C \times \mathbf{K}_C^T}{\sqrt{d}}\right) \times \mathbf{V}_C \quad (4)$$

$$\mathbf{F}_T = \text{SA}(\mathbf{E}_T, \mathbf{E}_T) \quad (5)$$

$$\mathbf{F}_C = \text{SA}(\mathbf{E}_C, \mathbf{E}_C) \quad (6)$$

$\mathbf{W}_T^Q, \mathbf{W}_T^K, \mathbf{W}_T^V, \mathbf{W}_C^Q, \mathbf{W}_C^K, \mathbf{W}_C^V \in R^{d_m \times d_n}$ ，并且  $\mathbf{W}_T^Q, \mathbf{W}_T^K, \mathbf{W}_T^V, \mathbf{W}_C^Q, \mathbf{W}_C^K, \mathbf{W}_C^V$  均为权重映射矩阵， $d_m$  表示自注意力 SA 层预先设定的维度， $n$  是注意力头的个数， $d_n = d_m / n$ ， $\mathbf{F}_T$  和  $\mathbf{F}_C$  分别为文本 SA 层和评论 SA 层的输出内容表示。

在虚假新闻检测领域，交叉注意力机制通过建立文本与真实评论之间的深度交互关系，能显著提升模型对隐蔽性虚假内容的识别能力<sup>[38]</sup>，因此本文构建了一个基于交叉注意力机制的 CA 层。在 CA 层里对文本和真实评论使用交叉注意力操作是通过文本来挖掘真实评论中的关键信息，同时解决部分因为不同评论而导致的真假新闻误判问题，如忽略与新闻无关的评论；或者积极寻找与新闻内容相关的评论部分、权威机构与专业领域人士在评论时所提出的坚实证据(如官方链接、数据来源等)。如图 6 所示，具体是先将文本的加权特征向量  $\mathbf{F}_T$  作为查询向量  $\mathbf{Q}$ ，将真实评论的加权特征向量  $\mathbf{F}_C$  同时作为键向量  $\mathbf{K}$  和值向量  $\mathbf{V}$ 。然后，计算  $\mathbf{Q}$  和  $\mathbf{K}$  之间的相似度分数再应用 softmax 函数将其归一化为权重系数，最后，使用这些权重系数加权求和值向量  $\mathbf{V}$ ，其中加权特征向量为  $\text{Atten}(\mathbf{E}_T, \mathbf{E}_C)$ ， $\mathbf{F}_{TC}$  为文本与真实评论特征融合向量。

$$\mathbf{Q}_T = \mathbf{F}_T \times \mathbf{W}_T^Q, \mathbf{K}_C = \mathbf{F}_C \times \mathbf{W}_C^K, \mathbf{V}_C = \mathbf{F}_C \times \mathbf{W}_C^V \quad (7)$$

$$\text{Atten}(\mathbf{E}_T, \mathbf{E}_C) = \text{softmax}\left(\frac{\mathbf{Q}_T \times \mathbf{K}_C^T}{\sqrt{d}}\right) \times \mathbf{V}_C \quad (8)$$

$$\mathbf{F}_{TC} = \text{CA}(\mathbf{E}_T, \mathbf{E}_C) \quad (9)$$

$\mathbf{W}_T^Q, \mathbf{W}_C^K, \mathbf{W}_C^V \in R^{d_m \times d_n}$ ， $\mathbf{W}_T^Q, \mathbf{W}_C^K, \mathbf{W}_C^V$  均为权重映射矩阵， $d_m$  表示交叉注意力 CA 层预先设定的维度， $n$  是注意力头的个数， $d_n = d_m / n$ ， $\mathbf{F}_{TC}$  是 CA 层的输出内容表示。

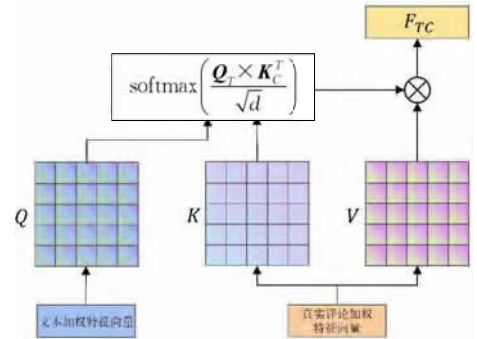


图 6 CA 层

### 3.4.2 文本与生成评论特征融合

为了解决大语言模型在生成评论时所产生的幻觉问题，引入了 Su 等人<sup>[39]</sup>提出的 DAAD 模型中

的自适应评论聚合器，如图7所示。具体包括关联矩阵计算模块、评论信息聚合层、动态门控调节机制和残差融合网络。首先通过可学习的参数矩阵  $W_T$  和  $W_C$  构建文本特征  $E_T$  与评论特征  $E_{C^*}$  的双语义映射空间，生成维度为  $n \times m$  的动态关联矩阵：

$$W_{TC^*} = (W_{C^*} \cdot E_{C^*}) \times (W_T \cdot E_T)^T \quad (10)$$

该矩阵用于量化文本与生成评论间的语义关联强度，识别评论中的关键信息片段。以关联矩阵  $W_{TC^*}$  作为注意力权重，对评论特征  $E_{C^*}$  进行加权聚合：

$$\widetilde{E}_{C^*} = \text{softmax} \left( \frac{W_{TC^*}^T}{\sqrt{d}} \right) \times E_{C^*} \quad (11)$$

通过使用  $\text{softmax}$  函数归一化来稳定梯度，实现评论信息的选择性聚合。构建元素级激活门控向量  $M_i$ ，通过逐元素相乘抑制幻觉影响：

$$M_i = [\delta(t_1 \odot \widetilde{c}_1^*), \delta(t_2 \odot \widetilde{c}_2^*), \dots, \delta(t_n \odot \widetilde{c}_n^*)] \quad (12)$$

其中， $\delta$  为  $ReLU$  激活函数，用以保留文本与评论的正向语义交互特征。通过跨层连接保留原始文本特征，输出最终融合向量：

$$F_{TC^*} = W_f(M_i \cdot (E_T \odot \widetilde{E}_{C^*})) + E_T \quad (13)$$

该操作通过权重矩阵  $W_f$  实现特征加权，并通过残差连接避免信息丢失，合理引导模块的输出，保证其不被评论信息所主导。 $d_m$  是预设的维度， $F_{TC^*}$  表示文本与生成评论特征融合向量， $\odot$  表示按元素相乘。

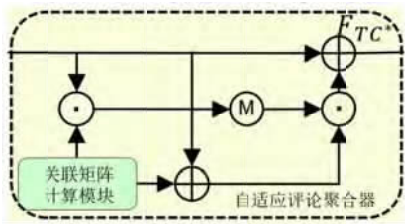


图7 自适应评论聚合器

### 3.4.3 真实评论与生成评论特征融合

为了能够聚合两种评论的关键信息，提高模型的泛化能力，将文本与真实评论特征融合向量  $F_{TC}$  和文本与生成评论特征融合向量  $F_{TC^*}$  放入  $MLP$  中进行降维和特征聚合，其中  $MLP$  包含线性层、归一化层和  $ReLU$  激活函数，在放入  $MLP$  之前先对两个融合向量进行平均池化操作以保留全局信息、降低计算成本。然后对输出的两个融合向量进行拼接得到评论间特征融合表示  $F_{TCC^*}$ ， $mean$  为平均池化操作：

$$F_{TCC^*} = MLP(\text{mean}(F_{TC})) \oplus MLP(\text{mean}(F_{TC^*})) \quad (14)$$

## 3.5 新闻信息分析融合模块

作为整个模型的多模态信息集成与一致性校验中心，该模块通过深度分析新闻的文本内容、新闻分析数据及图像信息之间的关联来捕捉潜在的新闻矛盾点，从而识别虚假新闻。

### 3.5.1 多模态信息提取

考虑到标准的 Transformer 架构在处理二维图像数据时存在计算复杂度高、内存消耗大等劣势。这里采用了 Swin-T 作为图像编码器，给定图像特征数据  $I = [I_1, I_2, \dots, I_n] \in R^{\omega \times h}$ ， $I_n$  表示第  $n$  张图像特征，其中  $\omega$  和  $h$  分别表示图像数据的宽度和高度，通过 Swin-T 编码后得到图像空间域特征  $S_i = [I_1^s, I_2^s, \dots, I_n^s]$ ， $I_i^s \in R^{d_s}$  表示第  $i$  张图像的空间域特征向量。其中  $d_s$  代表图像特征的嵌入维度，该维度决定了特征向量的语义表示能力， $s$  对应了 Swin-T 模型中图像分块(patch)的数量，反映了图像空间特征提取的粒度。

为了能够使得提取的新闻分析数据不光能和文本比对，还可以与图像比对，充分发挥出多模态检测的优势，需要借助 CLIP 预训练模型的深度语义对齐能力。它可以通过对比学习机制将新闻分析数据与图像映射至共同的语义空间，实现跨模态语义关联的深度捕捉。考虑到 CLIP 的训练目标是全局对齐，其更多关注图像与文本之间的共性，而忽略了每个模态的细节特征。如在图像中，CLIP 更多关注整体场景理解，可能错过细微的篡改痕迹；在文本中，CLIP 更关注关键词匹配，忽略了复杂的语义逻辑。因此，需要 Swin-T 来捕获图像的局部特征增强对图像的真伪鉴别，而使用 BERT 可以捕获文本的深层语义、语法关系，有助于识别虚假文本中的逻辑矛盾和事实错误。具体先使用 CLIP 模型编码新闻分析数据和图像数据，得到新闻分析内容编码特征向量  $E_R^{T^*} \in R^{d_c}$  和图像编码特征向量  $E_R^I \in R^{d_c}$ ， $d_c$  为 CLIP 模型的嵌入维度。通过拼接新闻分析内容编码特征向量和文本编码特征向量、图像编码特征向量和图像空间域特征向量可以增加多样化信息，确保新闻的各类信息数据可以有效地加入到模型中，这就是综合评测的具体意义。

$$E_{TT^*} = E_R^{T^*} \oplus E_T \quad (15)$$

$$E_{RI} = E_R^I \oplus S_i \quad (16)$$

其中， $E_{TT^*}$  为图文模态文本向量， $E_{RI}$  为图文模态图像向量。

### 3.5.2 跨模态特征融合

为了识别跨模态的语义一致性, 构建新闻分析数据与图像特征间的交互关系, 捕捉深层的跨模态语义关联, 这里使用双向的交叉注意力机制在 CA 层中实现跨模态的特征融合, 如图 8 所示。具体是先将图文模态文本向量和图文模态图像向量分别放入我们构建的投影层(Projection Layer)中映射到同一维度, 其中投影层由两层的线性层、两层的批量归一化层、两层的 ReLU 激活函数组成, 如图 9 所示。

$$\mathbf{E}_{TT^*} = \text{Projection}(\mathbf{E}_{TT^*}) \quad (17)$$

$$\mathbf{E}_{RI} = \text{Projection}(\mathbf{E}_{RI}) \quad (18)$$

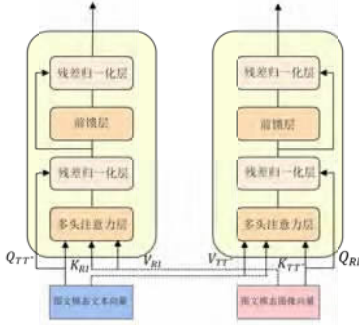


图 8 CA 层



图 9 投影层

然后以图文模态文本向量  $\mathbf{E}_{TT^*}$  为查询向量  $\mathbf{Q}$ , 再将图文模态图像向量  $\mathbf{E}_{RI}$  同时作为键向量  $\mathbf{K}$  和值向量  $\mathbf{V}$ , 计算  $\mathbf{Q}$  和  $\mathbf{K}$  之间的相似度分数再应用  $\text{softmax}$  函数将其归一化为权重系数后, 使用这些权重系数加权求和值向量  $\mathbf{V}$ , 其中加权特征向量为  $\text{Atten}(\mathbf{E}_{TT^*}, \mathbf{E}_{RI})$ 。再以图文模态图像向量  $\mathbf{E}_{RI}$  为查询向量  $\mathbf{Q}$ , 图文模态文本向量  $\mathbf{E}_{TT^*}$  同时作为键向量  $\mathbf{K}$  和值向量  $\mathbf{V}$ , 通过计算  $\mathbf{Q}$  和  $\mathbf{K}$  之间的相似度分数后应用  $\text{softmax}$  函数将其归一化得到权重系数, 使用这些权重系数加权求和值向量  $\mathbf{V}$ , 其中加权特征向量为  $\text{Atten}(\mathbf{E}_{RI}, \mathbf{E}_{TT^*})$ 。

$$\begin{aligned} \mathbf{Q}_{TT^*} &= \mathbf{E}_{TT^*} \times \mathbf{W}_{TT^*}^Q, \mathbf{K}_{RI} = \mathbf{E}_{RI} \times \mathbf{W}_{RI}^K, \\ \mathbf{V}_{RI} &= \mathbf{E}_{RI} \times \mathbf{W}_{RI}^V \end{aligned} \quad (19)$$

$$\begin{aligned} \mathbf{Q}_{RI} &= \mathbf{E}_{RI} \times \mathbf{W}_{RI}^Q, \mathbf{K}_{TT^*} = \mathbf{E}_{TT^*} \times \mathbf{W}_{TT^*}^K, \\ \mathbf{V}_{TT^*} &= \mathbf{E}_{TT^*} \times \mathbf{W}_{TT^*}^V \end{aligned} \quad (20)$$

$$\text{Atten}(\mathbf{E}_{TT^*}, \mathbf{E}_{RI}) = \text{softmax} \left( \frac{\mathbf{Q}_{TT^*} \times \mathbf{K}_{RI}^T}{\sqrt{d}} \right) \times \mathbf{V}_{RI} \quad (21)$$

$$\text{Atten}(\mathbf{E}_{RI}, \mathbf{E}_{TT^*}) = \text{softmax} \left( \frac{\mathbf{Q}_{RI} \times \mathbf{K}_{TT^*}^T}{\sqrt{d}} \right) \times \mathbf{V}_{TT^*} \quad (22)$$

$$\mathbf{F}_{TT^*} = \text{CA}(\mathbf{E}_{TT^*}, \mathbf{E}_{RI}) \quad (23)$$

$$\mathbf{F}_{TT^*} = \text{CA}(\mathbf{E}_{RI}, \mathbf{E}_{TT^*}) \quad (24)$$

$\mathbf{W}_{TT^*}^Q, \mathbf{W}_{RI}^K, \mathbf{W}_{RI}^V, \mathbf{W}_{TT^*}^Q, \mathbf{W}_{TT^*}^K, \mathbf{W}_{TT^*}^V \in \mathbb{R}^{d_m \times d_n}$ ,  $\mathbf{W}_{TT^*}^Q, \mathbf{W}_{RI}^K, \mathbf{W}_{RI}^V, \mathbf{W}_{TT^*}^Q, \mathbf{W}_{TT^*}^K, \mathbf{W}_{TT^*}^V$  分别为权重映射矩阵,  $d_m$  表示交叉注意力 CA 层预先设定的维度,  $n$  是注意力头的个数,  $d_n = d_m/n$ ,  $\mathbf{F}_{TT^*}$  为图文跨模态文本表示,  $\mathbf{F}_{TT^*}$  为图文跨模态图像表示。

将图文跨模态文本表示和图文跨模态图像表示分别放入 MLP 用来降维, 提取最关键的特征, MLP 包含线性层、归一化层和 ReLU 激活函数。拼接两个跨模态表示得到图文跨模态融合表示:

$$\tilde{\mathbf{F}}_{TT^*} = \text{MLP}(\mathbf{F}_{TT^*}) \oplus \text{MLP}(\mathbf{F}_{TT^*}) \quad (25)$$

### 3.6 分类器模块

分类器由集成  $\text{softmax}$  激活函数的全连接层 FC 所构成, 其中包括三层线性层、两层批量归一化层和两层的 ReLU 激活函数, 级联评论间特征融合表示  $\mathbf{F}_{TCC}$  和图文跨模态融合表示  $\tilde{\mathbf{F}}_{TT^*}$ , 并将其输入分类器以获得新闻样本的预测结果。y 是样本的真实标签,  $\hat{y}$  是模型预测标签, 选择最小化交叉熵函数作为损失函数,  $\mathcal{L}$  代表训练损失。

$$\widehat{\mathbf{F}}_{TTC^*} = \mathbf{F}_{TCC} \oplus \tilde{\mathbf{F}}_{TT^*} \quad (26)$$

$$\hat{y} = \text{FC}(\widehat{\mathbf{F}}_{TTC^*}) \quad (27)$$

$$\mathcal{L} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \quad (28)$$

## 4 实验结果与分析

### 4.1 实验设置

本文选择了 Weibo21<sup>[40]</sup>, Weibo<sup>[41]</sup>, GossipCop<sup>[42]</sup> 三个公开数据集进行实验, 其新闻样本所包含的文本、评论和图像均采集自真实社交平台。在实验开始前对数据集进行处理, 删除了缺少文本、评论或图片的样本, 包括一些因为包含敏感信息导致大语言模型无法生成评论或无法分析的样本。中文数据集的文本限制在 200token, 因为在相同的 token 下, 中文比英文表达的意思更丰富, 所以英文数据集的文本限制在 300token。每个新闻样本的真实评论为真实社交用户所发表, 数量不均等, 但限制单个评论的最大长度为 200token, 为每个样本生成的评论数为 8 个, 因此生成评论的总数要多于真实评论的数量。数据集的统计信息如表 1 所示, 评论统计信息如表 2 所示。

表 1 数据集统计信息

数据集	Fake	True	Overall
Weibo21	1922	2592	4514
Weibo	2805	3058	5863
GossipCop	3868	4331	8199

表2 评论统计信息

数据集	生成评论	真实评论	Overall
Weibo21	36112	20504	56616
Weibo	46904	34609	81513
GossipCop	65592	61684	127276

## 4.2 实验参数设置与评估指标

实验采用 8:2 的比例划分训练集与测试集。模型参数配置方面, BERT、Swin-T 和 CLIP 均采用默认维度设定, 且不进行参数微调。在数据集编码方面, 选择“bert-base-chinese”对中文数据集进行编码; 对英文数据集则使用“bert-base-uncased”模型来编码。CLIP 模型没有使用中文数据集微调, 因此对两个中文数据集中的 CLIP 输入部分使用智谱清言大语言模型(GLM-4-flash)翻译成英文。SA 层和 CA 层中预先设定的共同维数  $d_m=512$ , 注意力头的个数为  $n=8$ 。使用大语言模型生成评论时为了保证评论的多样性和丰富性, 采样温度设置为 0.7, 核采样数设置为 0.9; 在分析新闻时为了保证准确性和逻辑性, 聚焦核心语义, 采样温度设置为 0.1, 核采样数设置为 0.8。对三个数据集所使用的提示模板是一致的, 英文数据集则使用智谱清言大语言模型(GLM-4-flash)把提示翻译成英文。实验使用 Adam 优化器进行优化, 学习率设置为  $5 \times 10^{-6}$ , batchsize 的大小为 32, epoch 为 50 轮。实验平台为配备 Intel Core i7-11700 处理器及 NVIDIA GeForce RTX 3090 显卡(24 GB 显存)的服务器。实验结果的评估标准包括 Accuracy、Precision、Recall 和 F1-score。

## 4.3 对比模型

SAFE<sup>[43]</sup>: 通过深入挖掘新闻文本和所对应的视觉信息之间的关联特性, 构建跨模态关联分析模型来进行更全面、准确地识别新闻中可能出现的虚假信息。

SpotFake<sup>[44]</sup>: 提出了一种基于特征串联的多模态融合框架, 先分别提取新闻文本和对应图像或视频等多模态信息, 再通过级联文本与视觉模态的特征向量形成融合表示后实现虚假新闻检测。

SpotFake+<sup>[45]</sup>: SpotFake 的升级版, 通过跨模态迁移学习策略, 深度挖掘新闻文本与关联图像之间的语义相关性, 相较于之前版本, 该版本通过迁移学习增强了对语义深层关联的捕捉能力。

DistilBert<sup>[46]</sup>: 重点探究文本内容与社交网络拓扑结构之间的潜在关联性, 为虚假新闻检测提供了

从“内容语义-传播结构”双维度交叉验证的新范式, 展现了超越单一模态分析的检测优势。

CAFE<sup>[47]</sup>: 提出了一种基于歧义感知的多模态虚假新闻检测框架, 动态聚合单模态特征和跨模态交互信息进行虚假新闻检测。

MCAN<sup>[48]</sup>: 提出了一种创新的多模态协同注意力框架, 通过并行的文本编码器、空间域视觉编码器和频域视觉编码器实现特征提取, 并采用跨模态注意力机制进行特征融合。

FND-CLIP<sup>[49]</sup>: 基于多模态 CLIP 框架, 分别提取视觉和文本的深度嵌入表示, 并设计模态感知注意力机制实现特征的自适应动态适配。

EANN<sup>[50]</sup>: 使用迁移学习的思想运用对抗神经网络把在事件中学习的特征迁移到对虚假新闻的鉴别中。

BMR<sup>[51]</sup>: 采用 CLIP 模型获取图像-文本的联合嵌入空间表征, 通过可学习的跨模态注意力网络实现特征重加权和层级融合。

LogicDM<sup>[52]</sup>: 提出一种基于逻辑的神经网络模型, 集成了可解释的逻辑子句和参数化逻辑符号。

MRHFR<sup>[53]</sup>: 提出了一种结合新闻文本和图像信息并通过融合策略与用户评论的方法来检测虚假新闻。

FSRU<sup>[54]</sup>: 采用双重对比学习, 利用了虚假新闻的异常频率来鉴别虚假新闻。

AKA-Fake<sup>[55]</sup>: 采用将新闻文本内容和图像内容生成关联子图, 然后通过异构图学习的方式来整合新闻内容与外部知识。

CroMe<sup>[56]</sup>: 提出了利用跨模态三变换和度量学习来捕获模态内和模态间的语义关系, 再利用图像编码器和语言模型来组合跨模态的融合表示。

GPT-3.5-turbo<sup>[9]</sup>: 是 OpenAI 公司在 2022 年推出的 GPT 系列大语言模型, 基于 Transformer 架构, 拥有 1750 亿参数, 具备强大的自然语言理解与生成能力。

GLM-4-flashx<sup>[32]</sup>: 由北京智谱华章科技有限公司于 2023 年推出的 GLM 系列大模型, 是 GLM-4-flash 的改进型, 汇聚了超过 10TB 的多语言数据集资源, 支持处理长达 128KB 的上下文推理, 具备强大的推理性能与多语言能力。

DeepSeek-V3<sup>[37]</sup>: 由杭州深度求索人工智能基础技术研究有限公司在 2024 年 12 月 26 日发布, 基于创新的专家混合(MOE)架构, 共有 671B 个参数,

在长文本处理和数学推理等多个领域都展示了顶尖的性能。

Qwen2.5-32B<sup>[57]</sup>: 是阿里通义千问团队于 2024 年 9 月推出的开源 AI 大模型, 采用了结合 MQA 和 GQA 的混合注意力机制, 在自然语言、文本生成等方面都有了显著的提升。

Llama3-8B<sup>[58]</sup>: 由 Meta 公司在 2024 年推出的开源高效语言模型, 基于改进型的 Transformer 架构, 擅长多语言支持与高效推理。

#### 4.4 实验结果分析

为了能够确保基线结果的权威性与稳定性, 本文中的深度学习基线模型实验结果均来自文献原文; 大语言基准模型的实验结果则由本地部署或

API 调用的方式来得到。AKA-Fake 基线模型采用了混合指标不区分真实和虚假新闻, 所以在这里只采用了 *Accuracy* 指标。实验结果显示, 所提方法在多个数据集场景下均展现出显著的性能优势。如表 3 所示, 在 Weibo 数据集上的准确率达到 95.1%, 比最新的 FSRU 模型提升 5%。在 Weibo21 数据集上准确率达到 93.1%, 比最新的 CroMe 模型高 1.4%。在 GossipCop 数据集上准确率达到 88.3%, 比最新的 AKA-Fake 模型高出 2.7%。由于我们的数据集来自现实世界, 而现实世界中虚假新闻的数量是分布不均的, 这也导致了我们的样本产生了一定的不平衡, 进而影响了最终的结果。

表 3 数据集上不同方法的实验结果对比

数据集	Method	Pub.' Year	Accuracy	Fake News			Real News		
				Precision	Recall	F1-score	Precision	Recall	F1-score
GossipCop	EANN <sup>[50]</sup>	SIGKDD' 2018	0.864	0.702	0.518	0.594	0.887	0.956	0.920
	SpotFake+ <sup>[45]</sup>	AAAI' 2020	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	SAFE <sup>[43]</sup>	PAKDD' 2020	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	DistilBert <sup>[46]</sup>	arXiv' 2021	0.857	0.805	0.527	0.637	0.866	0.960	0.911
	CAFE <sup>[47]</sup>	WWW' 2022	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	BMR <sup>[51]</sup>	AAAI' 2023	<b>0.895</b>	0.752	0.639	<b>0.691</b>	<b>0.920</b>	<b>0.965</b>	<b>0.936</b>
	AKA-Fake <sup>[55]</sup>	AAAI' 2024	0.856	—	—	—	—	—	—
<b>Ours</b>	—	—	<b>0.883</b>	<b>0.895</b>	<b>0.853</b>	0.872	0.873	0.910	0.890
Weibo21	EANN <sup>[50]</sup>	SIGKDD' 2018	0.870	0.902	0.825	0.862	0.841	0.912	0.875
	SpotFake <sup>[44]</sup>	AAAI' 2020	0.851	<b>0.953</b>	0.733	0.828	0.786	<b>0.964</b>	0.866
	CAFE <sup>[47]</sup>	WWW' 2022	0.872	0.851	0.830	0.840	0.889	0.903	0.896
	BMR <sup>[51]</sup>	AAAI' 2023	0.929	0.908	<b>0.947</b>	0.927	<b>0.946</b>	0.906	0.925
	CroMe <sup>[56]</sup>	arXiv' 2025	0.917	0.944	0.917	0.930	0.880	0.918	<b>0.930</b>
	<b>Ours</b>	—	—	<b>0.931</b>	0.902	0.908	<b>0.933</b>	0.932	0.929
Weibo	SAFE <sup>[43]</sup>	PAKDD' 2020	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	CAFE <sup>[47]</sup>	WWW' 2020	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	SpotFake <sup>[44]</sup>	AAAI' 2020	0.892	0.902	<b>0.964</b>	0.932	0.847	0.656	0.739
	MCAN <sup>[48]</sup>	ACL' 2021	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	LogicDM <sup>[52]</sup>	ACL' 2023	0.852	0.862	0.845	0.853	0.843	0.859	0.851
	FND-CLIP <sup>[49]</sup>	ICME' 2023	0.907	0.914	0.801	0.908	0.914	0.901	0.907
	MRHFR <sup>[53]</sup>	AAAI' 2023	0.907	0.939	0.869	0.903	0.879	0.931	0.904
	FSRU <sup>[54]</sup>	AAAI' 2024	0.901	0.922	0.892	0.906	0.879	0.913	0.895
<b>Ours</b>	—	—	<b>0.951</b>	<b>0.978</b>	0.921	<b>0.947</b>	<b>0.921</b>	<b>0.978</b>	<b>0.949</b>

此外, 我们在上述基线模型的基础上, 又增加了 5 个大语言模型作为基线, 其中包括 3 个国内大语言模型和 2 个国外大语言模型。我们采用了 Zero-Shot、Zero-Shot CoT、Few-Shot 和 Few-Shot CoT 四种提示方式, Zero-Shot 为直接要求大语言模型判断新闻, Zero-Shot CoT 以思维链的方式提示大语言模型来判断新闻, Few-Shot 则为少样本的提示方式, 样本数设为 4 个, 2 个真实新闻和 2 个虚假新

闻, Few-Shot CoT 为少样本加思维链的提示方式, 具体结果如表 4、表 5 和表 6 所示。

对比基线模型的优劣。SAFE 依靠挖掘文本和视觉的关联性, 但是对于文本的多样化信息没有很好地集成; SpotFake 仅简单拼接了文本和图像, 对图文语义的冲突没有充分考虑; 作为改进型的 SpotFake+ 依赖大量的标记数据训练, 对真实新闻的分类性能较弱; CAFE 模型的长距离依赖捕捉不

表 4 在 GossipCop 数据集上同大语言模型方法的实验结果对比

数据集	LLM	Method	Accuracy	Fake News			Real News		
				Precision	Recall	F1-score	Precision	Recall	F1-score
GossipCop	DeepSeek-V3 <sup>[37]</sup>	Zero-Shot	0.646	0.355	0.650	0.459	0.860	0.644	0.737
		Zero-Shot CoT	0.682	0.889	0.375	0.527	0.631	0.958	0.761
		Few-Shot	0.717	0.827	0.508	0.630	0.672	0.905	0.771
		Few-Shot CoT	0.716	0.852	0.479	0.613	0.667	0.926	0.776
	Qwen2.5-32B <sup>[57]</sup>	Zero-Shot	0.681	0.780	0.452	0.572	0.644	0.886	0.746
		Zero-Shot CoT	0.651	0.784	0.359	0.493	0.614	0.912	0.734
		Few-Shot	0.685	0.814	0.430	0.563	0.642	0.912	0.754
		Few-Shot CoT	0.686	0.765	0.483	0.592	0.653	0.868	0.745
	GLM-4-flashx <sup>[32]</sup>	Zero-Shot	0.699	0.787	0.500	0.611	0.661	0.878	0.754
		Zero-Shot CoT	0.677	0.783	0.438	0.562	0.640	0.892	0.745
		Few-Shot	0.674	0.814	0.401	0.537	0.631	0.918	0.748
		Few-Shot CoT	0.667	0.712	0.494	0.583	0.645	0.822	0.723
	GPT-3.5-turbo <sup>[9]</sup>	Zero-Shot	0.673	0.737	0.521	0.611	0.638	0.819	0.718
		Zero-Shot CoT	0.618	0.577	0.722	0.641	0.678	0.526	0.592
		Few-Shot	0.707	0.405	0.562	0.471	0.851	0.751	0.798
		Few-Shot CoT	0.700	0.877	0.424	0.571	0.648	0.947	0.769
	Llama3-8B <sup>[58]</sup>	Zero-Shot	0.627	0.887	0.240	0.377	0.589	<b>0.973</b>	0.734
		Zero-Shot CoT	0.635	0.841	0.281	0.421	0.596	0.952	0.734
		Few-Shot	0.668	0.851	0.360	0.506	0.623	0.944	0.750
		Few-Shot CoT	0.654	0.700	0.469	0.562	0.633	0.820	0.715
<b>Ours</b>	—	<b>0.883</b>	<b>0.895</b>	<b>0.853</b>	<b>0.872</b>	<b>0.873</b>	0.910	<b>0.890</b>	

表 5 在 Weibo21 数据集上同大语言模型方法的实验结果对比

数据集	LLM	Method	Accuracy	Fake News			Real News		
				Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo21	DeepSeek-V3 <sup>[37]</sup>	Zero-Shot	0.751	0.637	0.960	0.766	<b>0.953</b>	0.596	0.733
		Zero-Shot CoT	0.771	0.659	0.952	0.779	0.947	0.637	0.762
		Few-Shot	0.763	0.658	0.919	0.767	0.915	0.647	0.758
		Few-Shot CoT	0.773	0.666	0.935	0.778	0.932	0.654	0.768
	Qwen2.5-32B <sup>[57]</sup>	Zero-Shot	0.759	0.703	0.905	0.791	0.863	0.609	0.714
		Zero-Shot CoT	0.763	0.692	0.953	0.802	0.922	0.568	0.703
		Few-Shot	0.751	0.739	0.782	0.760	0.764	0.718	0.740
		Few-Shot CoT	0.779	0.734	0.882	0.801	0.849	0.675	0.752
	GLM-4-flashx <sup>[32]</sup>	Zero-Shot	0.653	0.597	<b>0.972</b>	0.739	0.919	0.328	0.483
		Zero-Shot CoT	0.714	0.646	0.960	0.772	0.918	0.464	0.616
		Few-Shot	0.729	0.668	0.923	0.775	0.871	0.532	0.660
		Few-Shot CoT	0.720	0.653	0.948	0.773	0.901	0.488	0.633
	GPT-3.5-turbo <sup>[9]</sup>	Zero-Shot	0.668	0.563	0.961	0.710	0.941	0.452	0.611
		Zero-Shot CoT	0.763	0.663	0.900	0.763	0.899	0.661	0.762
		Few-Shot	0.590	0.509	0.960	0.665	0.916	0.318	0.472
		Few-Shot CoT	0.718	0.617	0.886	0.727	0.876	0.595	0.709
	Llama3-8B <sup>[58]</sup>	Zero-Shot	0.662	0.780	0.461	0.579	0.612	0.868	0.718
		Zero-Shot CoT	0.692	0.763	0.566	0.650	0.649	0.821	0.725
		Few-Shot	0.666	0.782	0.468	0.586	0.615	0.867	0.720
		Few-Shot CoT	0.704	0.744	0.630	0.682	0.673	0.779	0.722
<b>Ours</b>	—	<b>0.931</b>	<b>0.902</b>	0.908	<b>0.933</b>	0.932	<b>0.929</b>	<b>0.929</b>	

表 6 在 Weibo 数据集上同大语言模型方法的实验结果对比

数据集	LLM	Method	Accuracy	Fake News			Real News		
				Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	DeepSeek-V3 <sup>[37]</sup>	Zero-Shot	0.780	0.701	0.946	0.805	0.925	0.626	0.746
		Zero-Shot CoT	0.819	0.754	0.928	0.832	0.915	0.718	0.805
		Few-Shot	0.786	0.728	0.886	0.799	0.867	0.692	0.770
		Few-Shot CoT	0.789	0.721	0.917	0.807	0.896	0.670	0.767
	Qwen2.5-32B <sup>[57]</sup>	Zero-Shot	0.773	0.708	0.904	0.794	0.878	0.648	0.746
		Zero-Shot CoT	0.757	0.679	0.947	0.791	0.920	0.579	0.710
		Few-Shot	0.763	0.741	0.786	0.763	0.786	0.741	0.763
		Few-Shot CoT	0.794	0.743	0.881	0.806	0.864	0.713	0.781
	GLM-4-flashx <sup>[32]</sup>	Zero-Shot	0.643	0.579	<b>0.971</b>	0.725	0.925	0.333	0.490
		Zero-Shot CoT	0.702	0.624	0.969	0.759	<b>0.939</b>	0.452	0.610
		Few-Shot	0.729	0.660	0.913	0.766	0.871	0.555	0.678
		Few-Shot CoT	0.708	0.634	0.941	0.758	0.898	0.488	0.632
	GPT-3.5-turbo <sup>[9]</sup>	Zero-Shot	0.640	0.562	0.936	0.702	0.881	0.395	0.545
		Zero-Shot CoT	0.732	0.680	0.909	0.778	0.848	0.543	0.662
		Few-Shot	0.635	0.572	0.958	0.716	0.896	0.335	0.487
		Few-Shot CoT	0.707	0.635	0.924	0.752	0.877	0.506	0.641
	Llama3-8B <sup>[57]</sup>	Zero-Shot	0.675	0.795	0.444	0.570	0.630	0.892	0.738
		Zero-Shot CoT	0.702	0.765	0.558	0.645	0.668	0.838	0.743
		Few-Shot	0.672	0.843	0.400	0.542	0.621	0.930	0.745
		Few-Shot CoT	0.731	0.812	0.581	0.678	0.688	0.874	0.770
<b>Ours</b>	--	<b>0.951</b>	<b>0.978</b>	0.921	<b>0.947</b>	0.924	<b>0.978</b>	<b>0.949</b>	

足会导致特征提取不完整;MCAN使用多模态协同注意力网络,存在对多模态交互捕捉不足和协同注意力机制粗糙等问题;FND-CLIP依赖CLIP的图文对比学习能力,但无法适应虚假新闻特有的跨模态矛盾模式;EANN依赖特定事件和领域知识,对于复杂任务缺少泛化能力;BMR的计算开销太大并且难以捕捉动态的模态关系,在GossipCop数据集上优于我们方法的原因可能是他们所使用的数据集样本多于我们,尤其是真实新闻的样本,由于大语言模型无法处理包含敏感信息的样本,导致我们的样本数较BMR方法更少,未来我们会针对这个问题做出改进;AKA-Fake模型通过关联新闻文本和图像子图来联系外部知识,缺少大语言模型的泛化能力;LogicDM的模型基于逻辑符号,存在对逻辑谓词的定义困难、复杂关系处理能力有限等问题;MRHFR拥有过度设计的层级结构与静态融合机制导致模态交互滞后和细粒度特征丢失等问题;FSRU首次尝试从频率的角度检测虚假新闻,然而这种方法无法集成评论等新闻数据;CroMe利用了图像编码器和大语言模型来捕获跨模态的图文表示,再通过Transformer对融合特征进行处理,但

是单纯的图文模态信息已经不能满足目前虚假新闻检测的要求,需要大语言模型来辅助提取新闻内容的深层信息。大语言模型在不同提示下尽管性能有所提升,并且在某些指标上要超过我们的方法。比如在GossipCop数据集上,Llama3-8B的真实新闻Recall指标优于我们的方法,说明Llama3模型在减少对虚假新闻的漏检方面具有一定优势;在Weibo21数据集上,GLM-4-flashx模型在对虚假新闻的漏检和DeepSeek-V3模型在对真实新闻的误判中优于我们的方法;在Weibo数据集中,GLM-4-flashx模型在对虚假新闻的漏检和真实新闻的误判中都优于我们的方法。但这些大语言基准模型整体的准确率和F1综合分数都不如我们,因此大语言模型暂时还不具备直接充当虚假新闻检测模型的条件。

我们的方法优于其他基线模型的原因如下:

(1)动态评论交互融合模块通过自适应评论聚合器剔除了一部分生成评论的幻觉信息,使用SA层和CA层聚焦了与新闻内容相关的真实评论信息,然后结合了二者所包含的新闻信息,既解决了利用大语言模型生成评论的检测方法缺少多元评

论而导致泛化能力弱的局限性，又显著提高了检测的准确率。

(2)三重视角通过大语言模型的自我审查和同行评估，利用了它的深层理解与上下文推理能力，可以解决一部分大语言模型在分析新闻信息时所产生的幻觉内容，而它在分析时所提供的背景知识则可以帮助两个检测模块在鉴定虚假新闻时提高其泛化性和准确性。三重视角的最后一重是综合评测，即通过动态评论交互融合模块和新闻信息分析融合模块对新闻文本、各类评论、图像及新闻信息分析数据进行特征提取和跨模态语义融合来解决前两重视角所遗漏的幻觉部分。实验结果显示，三重视角确实有助于解决现有大语言模型检测方法在提取新闻语义信息时容易受到幻觉影响的局限性。

(3)本文设计的新闻信息分析融合模块，首先利用 BERT 和 Swin-T 有效提取了新闻文本和图像

的深度语义信息和局部特征信息，增强了模态的特定特征。而 CLIP 作为图文对比学习的多模态预训练模型来对新闻分析数据和新闻图像进行编码，可以有效增强其模态对齐特征，其中新闻分析数据包含了大语言模型对文本和评论等信息的交叉验证与专业意见。将 BERT 编码的文本特征与 CLIP 编码的新闻分析数据特征拼接，强化了文本方向的特征信息，而 Swin-T 编码的图像特征与 CLIP 编码的图像特征拼接，强化了图像方向的特征信息。最后再通过双向的交叉注意力机制进行跨模态的特征融合，显著增强了跨模态的图文语义理解。这种增强跨模态语义理解的检测框架可以应对多种检测场景，有助于提高对复杂虚假新闻的检测能力。

#### 4.5 消融实验分析

为了对模型每个部分的性能做一个评估，我们进行了一些消融实验，具体结果如表 7 所示。

表 7 不同数据集上的消融结果对比

数据集	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
GossipCop	w/o S	0.808	0.615	0.623	0.603	0.874	0.871	0.869
	w/o T	0.810	0.570	0.625	0.582	<b>0.888</b>	0.863	0.873
	w/o Z	0.592	0.318	0.672	0.432	0.852	0.567	0.681
	w/o SA	0.852	<b>0.910</b>	0.771	0.829	0.809	0.929	0.863
	w/o cmts	0.854	0.730	0.510	0.600	0.876	0.949	0.911
	w/o BS	0.802	0.862	0.716	0.777	0.758	0.880	0.811
	w/o mv	0.862	0.801	0.482	0.602	0.871	<b>0.967</b>	<b>0.917</b>
	<b>Ours</b>	<b>0.883</b>	0.895	<b>0.853</b>	<b>0.872</b>	0.873	0.910	0.890
Weibo21	w/o S	0.870	0.816	0.912	0.857	0.920	0.845	0.876
	w/o T	0.915	0.900	0.919	0.906	0.932	0.911	0.918
	w/o Z	0.706	0.595	<b>0.955</b>	0.733	<b>0.941</b>	0.523	0.673
	w/o SA	0.831	0.756	0.884	0.813	0.899	0.790	0.839
	w/o cmts	0.876	0.888	0.773	0.823	0.853	<b>0.948</b>	0.896
	w/o BS	0.806	0.830	0.688	0.741	0.798	0.897	0.838
	w/o mv	0.831	0.733	0.908	0.841	0.907	0.692	0.809
	<b>Ours</b>	<b>0.931</b>	<b>0.902</b>	0.908	<b>0.933</b>	0.932	0.929	<b>0.929</b>
Weibo	w/o S	0.931	0.925	0.936	0.929	<b>0.942</b>	0.923	0.931
	w/o T	0.945	0.958	0.928	0.942	0.931	0.956	0.942
	w/o Z	0.698	0.619	<b>0.966</b>	0.755	0.935	0.450	0.608
	w/o SA	0.852	0.856	0.830	0.838	0.844	0.872	0.853
	w/o cmts	0.934	0.956	0.900	0.926	0.915	0.964	0.938
	w/o BS	0.832	0.861	0.769	0.808	0.808	0.886	0.841
	w/o mv	0.869	0.929	0.891	0.910	0.895	0.932	0.913
	<b>Ours</b>	<b>0.951</b>	<b>0.978</b>	0.921	<b>0.947</b>	0.924	<b>0.978</b>	<b>0.949</b>

w/o S: 缺少自我审查视角，没有经过第一个大语言模型 GLM-4-flash 的第一轮分析，直接将评论和文本放入第二个大语言模型 DeepSeek-R1-32B 进行分析；w/o T: 缺少同行评估视角，将 GLM-4-flash

的第一轮分析直接作为新闻分析数据输入到模型中；w/o Z: 缺少综合评测视角，让第二个大语言模型 DeepSeek-R1-32B 根据文本、评论和第一轮分析的结果直接来判断新闻的真假；w/o cmts: 不使用

动态评论交互融合模块; w/o SA: 在仅使用动态评论交互融合模块时不引入 SA 层; w/o mv: 不使用新闻信息分析融合模块; w/o BS: 在仅使用新闻信息分析融合模块时不引入 BERT 和 Swin-T。

不使用自我审查视角(w/o S), 即 GLM-4-flash 不对自己生成的评论和新闻文本进行分析, 将文本和生成评论直接输入给 DeepSeek-R1-32B 来进行分析, 得到的新闻分析数据输入至新闻信息分析融合模块。由于没有加入 GLM-4-flash 的自我分析, 导致了模型性能有一定的下降, 说明大语言模型对自己生成的数据还是具备了初步的判断能力, 并且可以消除一部分的幻觉, 在优化提示的情况下, 模型的性能可能会继续提升。

缺少同行评估视角(w/o T), 即 GLM-4-flash 对新闻文本和自己生成的评论做完一轮分析后得到的新闻分析数据直接输入至新闻信息分析融合模块, 不再通过 DeepSeek-R1-32B 的同行分析。相比于缺少自我审查(w/o S)的模型仍保持了一部分的性能, 说明了在一定程度上自我审查功能的不可替代。在 Weibo 数据集上的实验证明了同行评估的重要性, 同行评估主要是可以从其它大语言模型那里得到一些独特的视角, 这是自我审查和综合评测所不具备的。

缺少综合评测视角(w/o Z), 即 DeepSeek-R1-32B 根据新闻文本、GLM-4-flash 的生成评论和分析结果来直接判断新闻的真假。实验结果表明综合评测对模型性能的影响最大, 尽管通过大语言模型的自我审查和同行评估可以消除部分幻觉, 提高检测效率, 但在没有深度学习模型的加入下, 大语言模型仅依靠自己的判断还不能完全胜任对虚假新闻的检测工作。

为了能够更直观地检验 SA 层在评论模块中的作用, 我们在仅使用动态评论交互融合模块时剔除了 SA 层, 直接将文本与真实评论输入至 CA 层(w/o SA)。实验结果显示缺少了自注意力机制对文本和评论的核心内容聚焦, 导致了模块检测率有所下降。尽管 BERT 提供了丰富的上下文, 但也包含了大量与当前特定分类任务无关的信息, 因此使用自注意力机制对内容的关键部分进行信息聚焦是十分有必要的。

缺少动态评论交互融合模块(w/o cmts)导致了模型性能的下降, 说明评论确实包含了一些关于新闻内容的有效信息, 而提取生成评论与真实评论融

合后的深层语义有助于加强模型检测虚假新闻的能力。在实例分析中我们还做了补充实验验证两种评论的结合是否真的可以提高虚假新闻检测的准确率。

为了能够直观地验证 BERT 和 Swin-T 在新闻信息分析融合模块中所起到的效果, 我们在仅使用该模块的情况下不采用 BERT 和 Swin-T 编码新闻文本和图像, 只将新闻分析数据和图像数据使用 CLIP 进行编码后输入该检测模块(w/o BS)。由于缺少了最重要的新闻文本和图像特征, 尽管新闻分析数据拥有大语言模型对新闻内容的分析结果, 以及 CLIP 的图文编码, 但实验的准确率还是下降较多。因此, 使用 BERT 和 Swin-T 对文本和图像的编码是有效的, 可以显著增强模型的检测能力。

去除新闻信息分析融合模块(w/o mv)与完整模型的对比实验结果表明: 该模块在整个模型中尤为重要。该模块包含了图像和新闻内容分析两大重要信息, 在 CLIP 模型和交叉注意力机制的帮助下与新闻内容进行跨模态的特征融合和语义对齐, 有效提高了模型整体的检测性能。

## 5 实例分析

为了能够验证生成评论和真实评论的结合是否真的可以提高虚假新闻检测的准确性, 我们首先做了评论信息的搜集。如图 10 所示, 我们在社交平台上找到了几则与数据集中比较相似的新闻内容并截取了新闻的真实评论与数据集的生成评论。直接观察发现, 大语言模型所生成的评论与人类发表的真实评论已经达到高度相似的程度。因此生成评论可以弥补现实世界中真实评论数量和质量较低、部分新闻缺少评论的现状。

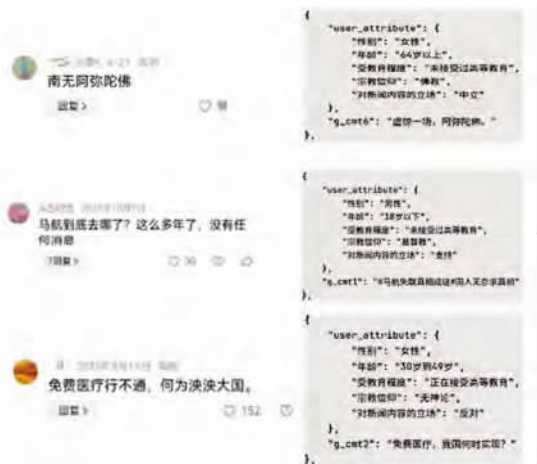


图 10 真实评论与生成评论部分示例

生成评论是否可以完全替代真实评论?为了研究这个问题,我们从 Weibo 数据集中随机抽取了 1000 条样本,其中包括 500 条真实新闻和 500 条虚假新闻。先使用 BERT 编码再计算每个样本下真实评论与生成评论的平均余弦相似度和 Jaccard 相似度,然后再统计所有样本的数据之后取平均值,表 8 为整个数据集真实评论和生成评论的平均 BERT 相似度和 Jaccard 相似度(平均 BERT 相似度是一种利用 BERT 来计算两个文本序列语义层面相似度的方法,在 BERT 向量空间中,语义相似度通常落在 0 到 1 之间,越接近 1 表示语义越相似;Jaccard 相似度是一种用于衡量两个有限集合相似性的统计指标,具有计算速度快,资源消耗低等优势,取值范围在数学上被固定为 0 到 1 之间,越接近 1 相似度越高<sup>[59]</sup>)。统计结果显示平均 BERT 相似度达到 0.837 超过 0.800,在语义关系上处于强相关,表明生成评论和真实评论在语义层面比较接近,因此在语义相似度上也证明了生成评论可以弥补真实评论的一些不足。然而 Jaccard 相似度为 0.009 远低于弱重叠标准 0.05,过低的 Jaccard 相似度说明生成评论用词与真实评论几乎无重叠,暴露了生成评论刻意回避自然表达,不如真实评论表达多样且符合人类的语言习惯。生成评论存在太过刻板、样式单一的缺点,而真实评论在表达上会有颜文字、俚语、反讽语句等多样化形式,真实评论的独特性决定了生成评论无法完全替代真实评论。

表 8 评论相似度统计信息

平均 BERT 相似度	平均 Jaccard 相似度
0.837	0.009

为了能够进一步验证生成评论和真实评论在虚假新闻检测中的具体效果,我们在上述 1000 条 Weibo 样本的基础上选取了 5 种情况进行实验,结果如表 9 所示。我们的方法是使用 BERT 编码文本和评论,分别做自注意力操作后,再使用交叉注意力,降维后放入分类器来最终得到各项实验指标。cmts 是只使用真实评论, g\_cmts 是只使用生成评论, mix\_cmts 为混合两种评论与文本做交叉注意力操作, g\_cmts+cmts 则是做完交叉注意力操作后拼接两种不同的文本与评论融合表示再放入分类器中,最后一种是使用我们的动态评论交互融合模块。实验结果表明结合了两类评论的方法效果比单

一评论的效果更好,而我们的方法既优于单一评论的方法,也强于其他两种结合了两类评论的方法。

表 9 使用不同方法对两种评论的实验结果对比

数据集	Accuracy	Fake News			Real News		
		Precision	Recall	F1-score	Precision	Recall	F1-score
cmts	0.795	0.832	0.708	0.736	0.720	0.840	0.753
g_cmts	0.810	0.815	0.722	0.737	0.743	0.860	0.773
mix_cmts	0.835	0.728	0.752	0.725	0.783	0.758	0.743
g_cmts+cmts	0.822	0.632	0.702	0.650	<b>0.894</b>	0.866	<b>0.878</b>
<b>Ours</b>	<b>0.875</b>	<b>0.908</b>	<b>0.846</b>	<b>0.867</b>	0.836	<b>0.902</b>	0.863

## 6 可视化实验

为了进一步衡量我们所提出的方法鉴别虚假新闻的能力,我们引入了 AUC 指标<sup>[60]</sup>。这是一种可以在类别不平衡场景中衡量模型的整体区分能力的实验标准,值域为 0 到 1,越接近 1,模型的区分能力越强。在虚假新闻检测领域,AUC 指标可以在真假新闻样本分布不均的情况下有效评估模型将虚假新闻(正例)和真实新闻(负例)区分开来的能力。图 11、图 12 和图 13 分别显示了我们在三个数据集中利用我们的方法与 5 个大语言模型进行 AUC 指标的对比,其中大语言模型使用了四种提示方法。可以看到我们的方法在真假新闻分布不平衡的情况下仍旧保持了较好的分类性能,模型学习到了区分真假新闻的有效模式,拥有比较良好的鲁棒性和泛化潜力。其他信息如 DeepSeek-V3 作为新晋国产大语言模型,在中英文数据集上的表现结果都较为良好,大部分指标都接近甚至超过平均 AUC 指标;而紧随其后的是 Qwen2.5-32B,在两个中文数据集中也取得了不错的效果,但在英文数据集中效果不是很理想,这可能与它训练数据侧重于中文数据,缺少大规模的英文训练数据有关;GLM-4-flashx 在中文数据集中随着提示方法的改变,效果逐渐得到提升,而在英文数据集中,效果反而呈现下降趋势,提示工程依赖于大语言模型基础的语言能力,在非原生语言的基础上优化提示可能会放大模型的语言缺陷。GPT-3.5-turbo 作为老牌的大语言模型也取得了较为优异的效果,但是在中文数据集中依赖 CoT 的方式,在英文数据集中依赖 Fewshot 的方式;Llama3-8B 整体的效果不如其他大语言模型,在与其他模型的对比实验中超过半数处于靠后位置,这可能与它包含的参数较低有关。

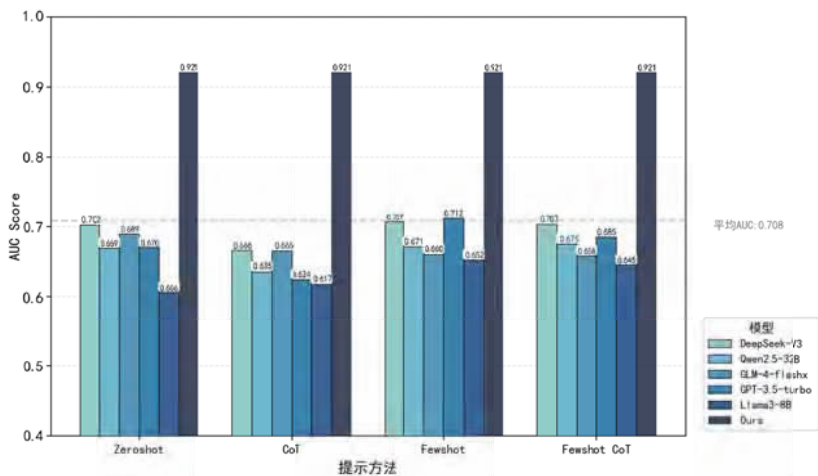


图 11 在 GossipCop 数据集上同大语言模型方法的性能比较(AUC)

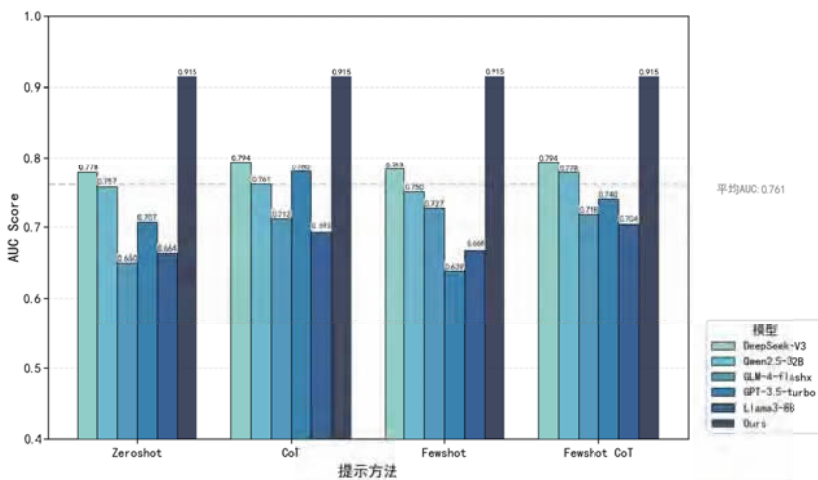


图 12 在 Weibo21 数据集上同大语言模型方法的性能比较(AUC)

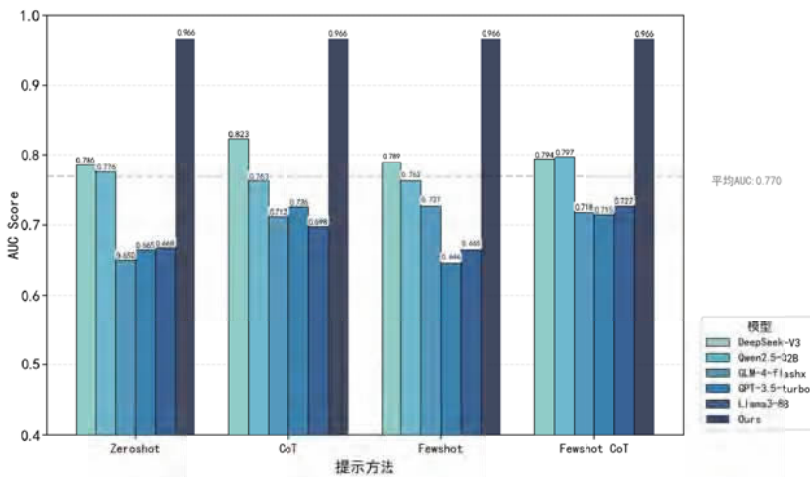


图 13 在 Weibo 数据集上同大语言模型方法的性能比较(AUC)

### 7 总结与展望

为了解决现有方法缺少多元评论和模型泛化性弱、无法处理复杂场景等局限性。提出了一种基于多元评论和三重视角的多模态虚假新闻检测方法。

该方法首先利用大语言模型为新闻内容生成评论，并通过大语言模型的自我审查和同行评估来分析新闻，得到新闻分析数据，同时减轻了部分使用大语言模型过程中产生的幻觉问题，从而有效保证了检测模型的准确性，再通过动态评论交互融合模块和

新闻信息分析融合模块来聚合两种评论信息并融合新闻分析数据用于虚假新闻检测。既充分发挥了深度学习模型的优势,又有效利用了大语言模型技术,为之后的虚假新闻检测提供了新思路。

本文的不足之处在于使用大语言模型生成评论和分析文本时,涉及政治、暴力等敏感信息会调用失败。为了保证数据集的稳定性,我们过滤了这一类数据,但这一类新闻才是虚假新闻的重灾区,未来我们将针对这类敏感问题展开专项研究。

## 参 考 文 献

- [1] Castillo C, Mendoza M, Poblete B. Information credibility on twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India. 2011: 675-684
- [2] Agarwal S, Farid H, Gu Yu-Ming, et al. Protecting world leaders against deep fakes//CVPR workshops. Long Beach, USA. 2019: 6-12
- [3] Barman D, Guo Zi-Yi, Conlan O. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. Machine Learning with Applications, 2024, 16: 100545
- [4] Chen Can-Yu, Shu Kai. Can llm-generated misinformation be detected? arXiv preprint arXiv:2309.13788, 2023
- [5] Nan Qiong, Sheng Qiang, Cao Juan, et al. Exploiting user comments for early detection of fake news prior to users' commenting. Frontiers of Computer Science, 2025, 19(10): 1910354
- [6] Cui Li-Meng, Wang Su-Hang, Lee D. Same: Sentiment-aware multi-modal embedding for detecting fake news//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Vancouver, Canada. 2019: 41-48
- [7] Qi Peng, Cao Juan, Sheng Qiang. Semantics-enhanced multi-modal fake news detection. Journal of Computer Research and Development, 2021, 58(7): 1456-1465(in Chinese)  
(元鹏, 曹娟, 盛强. 语义增强的多模态虚假新闻检测. 计算机研究与发展, 2021, 58(7): 1456-1465)
- [8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need// Advances in neural information processing systems 30 (NeurIPS 2017). Long Beach, USA, 2017: 5998-6008
- [9] OpenAI. "ChatGPT: Optimizing language models for dialogue," Blog, OpenAI, 2022. <https://openai.com/blog/chatgpt/>. Accessed: 2025-06-03
- [10] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023
- [11] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682, 2022
- [12] Hu Bei-Zhe, Sheng Qiang, Cao Juan, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 38(20): 22105-22113
- [13] Nan Qiong, Sheng Qiang, Cao Juan, et al. Let silence speak: Enhancing fake news detection with generated comments from large language models//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise, USA, 2024: 1732-1742
- [14] Devlin J, Chang Ming-Wei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186
- [15] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2021: 8748-8763
- [16] Liu Ze, Lin Yu-Tong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/ CVF International Conference on Computer Vision. Tokyo, Japan, 2021: 10012-10022
- [17] Ajaio O, Bhowmik D, Zargari S. Sentiment aware fake news detection on online social networks//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Glasgow, UK, 2019: 2507-2511
- [18] Jing Jing, Wu Hong-Chen, Sun Jie, et al. Multimodal fake news detection via progressive fusion networks. Information Processing & Management, 2023, 60(1): 103120
- [19] Liu Yin-Han, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019
- [20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014
- [21] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778
- [22] Zhang Shao-Qin, Du Sheng-Dong, Zhang Xiao-Bo, et al. Social network rumor detection method integrating multimodal information. Computer Science, 2021, 48(5): 117-223(in Chinese)  
(张少钦, 杜圣东, 张晓博等. 融合多模态信息的社交网络谣言检测方法. 计算机科学, 2021, 48(5): 117-123)
- [23] Liu Hao-Tian, Li Chun-Yuan, Wu Qing-Yang, et al. Visual instruction tuning. Advances in Neural Information Processing Systems, 2023, 36: 34892-34916
- [24] Xuan Ke-Yang, Yi Li, Yang Fan, et al. LEMMA: Towards LVLm-enhanced multimodal misinformation detection with external knowledge augmentation. arXiv preprint arXiv:2402.11943, 2024
- [25] Liu Xuan-Nan, Li Pei-Pei, Huang Huai-Bo, et al. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented llms//Proceedings of the 32nd ACM International Conference on Multimedia. Orlando, USA, 2024: 10154-10163
- [26] Wan He-Run, Feng Shang-Bin, Tan Zhao-Xuan, et al. Dell: Generating reactions and explanations for llm-based misinformation detection. arXiv preprint arXiv:2402.10426, 2024
- [27] Ke Jing, Xie Zhe-Yong, Xu Tong, et al. An implicit semantic enhanced fine-grained fake news detection method based on large

- language models. *Journal of Computer Research and Development*, 2024, 61(5): 1250-1260. doi: 10.7544/issn 1000-1239.202330967 (in Chinese)  
(柯婧, 谢哲勇, 徐童等. 基于大语言模型隐含语义增强的细粒度虚假新闻检测方法. *计算机研究与发展*, 2024, 61(5): 1250-1260)
- [28] Liu Yu-Han, Chen Xiu-Ying, Zhang Xiao-Qing, et al. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. arXiv preprint arXiv:2403.09498, 2024
- [29] Ignat O, Xu Xiao-Meng, Mihalcea R. Maide-up: Multilingual deception detection of gpt-generated hotel reviews. arXiv preprint arXiv:2404.12938, 2024
- [30] Wang L Z, Ma Yi-Ming, Gao Ren-Fei, et al. Megafake: A theory-driven dataset of fake news generated by large language models. arXiv preprint arXiv:2408.11871, 2024
- [31] Shu Kai, Sliva A, Wang Su-Hang, et al. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 2017, 19(1): 22-36
- [32] Bei jing Zhipu Huazhang Technology Co., Ltd. ZhipuAI, 2023. <https://chatglm.cn/>. Accessed:2025-06-03  
(北京智谱华章科技有限公司. 智谱清言. <https://chatglm.cn/>, 2025-06-03)
- [33] Shu Kai, Zhou Xin-Yi, Wang Su-Hang, et al. The role of user profiles for fake news detection//*Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Vancouver, Canada, 2019: 436-439
- [34] Boyer P. Why divination? Evolved psychology and strategic interaction in the production of truth. *Current Anthropology*, 2020, 61(1): 100-123
- [35] Greenberg N E, Wallick A, Brown L M. Impact of COVID-19 pandemic restrictions on community-dwelling caregivers and persons with dementia. *Psychological Trauma: Theory, Research, Practice, and Policy*, 2020, 12(S1): S220
- [36] Zhang Yue, Li Ya-Fu, Cui Le-Yang, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219, 2023
- [37] Hangzhou DeepSeek Artificial Intelligence Co., Ltd. DeepSeek, 2023. <https://www.deepseek.com/>. Accessed 2025-06-03  
(杭州深度求索人工智能基础技术研究有限公司. DeepSeek. <https://www.deepseek.com/>, 2025-06-03)
- [38] Bansal S, Singh N S, Dar S S, et al. MMCfND: Multimodal multilingual caption-aware fake news detection for low-resource indic languages. arXiv preprint arXiv:2410.10407, 2024
- [39] Su Xin-Qi, Cui Ya-Wen, Liu A-Jian, et al. DAAD: Dynamic analysis and adaptive discriminator for fake news detection. arXiv preprint arXiv:2408.10883, 2024
- [40] Nan Qiong, Cao Juan, Zhu Yong-Chun, et al. MDFEND: Multi-domain fake news detection//*Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Melbourne, Australia, 2021: 3343-3347
- [41] Jin Zhi-Wei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs//*Proceedings of the 25th ACM international conference on Multimedia*. Mountain View, USA, 2017: 795-816
- [42] Shu Kai, Mahudeswaran D, Wang Su-Hang, et al. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 2020, 8(3): 171-188
- [43] Zhou Xin-Yi, Wu Jin-Di, Zafarani R: Similarity-aware multi-modal fake news detection//*Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Singapore, 2020: 354-367
- [44] Singhal S, Shah R R, Chakraborty T, et al. Spotfake: A multi-modal framework for fake news detection// *Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. Beijing, China. 2019: 39-47
- [45] Singhal S, Kabra A, Sharma M, et al. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)//*Proceedings of the AAAI Conference on Artificial Intelligence*. New York, USA, 2020, 34(10): 13915-13916
- [46] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arxiv preprint arxiv:1910.01108, 2019
- [47] Chen Yi-Xuan, Li Dong-Sheng, Zhang Peng, et al. Cross-modal ambiguity learning for multimodal fake news detection//*Proceedings of the ACM Web Conference 2022*. Lyon, France, 2022: 2897-2905
- [48] Wu Yang, Zhan Peng-Wei, Zhang Yun-Jian, et al. Multimodal fusion with co-attention networks for fake news detection//*Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Osaka, Japan, 2021: 2560-2569
- [49] Zhou Yang-Ming, Yang Yu-Zhou, Ying Qi-Chao, et al. Multimodal fake news detection via clip-guided learning//*Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME)*. Xiamen, China, 2023: 2825-2830
- [50] Wang Ya-Qing, Ma Feng-Long, Jin Zhi-Wei, et al. Eann: Event adversarial neural networks for multi-modal fake news detection//*Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*. London, UK, 2018: 849-857
- [51] Ying Qi-Chao, Hu Xiao-Xiao, Zhou Yang-Ming, et al. Bootstrapping multi-view representations for fake news detection// *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2023, 37(4): 5384-5392
- [52] Liu Hui, Wang Wen-Ya, Li Hao-Liang. Interpretable Multimodal Misinformation Detection with Logic Reasoning. arXiv preprint arXiv:2305.05964, 2023
- [53] Wu Lian-Wei, Liu Pu-Sheng, Zhang Yan-Ming. See how you read? Multi-reading habits fusion reasoning for multi-modal fake news detection//*Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2023, 37(11): 13736-13744
- [54] Lao An, Zhang Qi, Shi Chong-Yang, et al. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector//*Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2024, 38(16): 18426-18434
- [55] Zhang Li-Tian, Zhang Xiao-Ming, Zhou Zi-Yi, et al. Reinforced adaptive knowledge learning for multimodal fake news detection// *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2024, 38(15): 16777-16785
- [56] Choi E, Ahn J, Piao Xin-Yu, et al. CroMe: Multimodal fake news detection using cross-modal tri-transformer and metric learning. arxiv preprint arXiv:2501.12422, 2025

- [57] Alibaba Cloud Computing Co., Ltd. Chat Qwen AI, 2024. <https://chat.qwen.ai/>. Accessed: 2025-06-03  
(阿里云计算有限公司. 通义千问. <https://chat.qwen.ai/>, 2025-06-03)
- [58] Meta.Llama3.1, 2024. <https://www.llama.com/>. Accessed: 2025-06-03
- [59] Zhang Xiao, Sun Su-Fen, Wei Qing-Feng, et al. Research on

tomato question answering model based on BERT multi-feature fusion. *Information and Computer (Theoretical Edition)*, 2021, 33(17): 69-73(in Chinese)

(张笑, 孙素芬, 魏清风等. 基于 BERT 多特征融合的番茄问答模型研究. *信息与电脑(理论版)*, 2021, 33(17): 69-73)

- [60] Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, 27(8): 861-874



**YU Yong-Yi**, M.S. candidate. His research interests include natural language processing, fake news detection.

**XIAO Cong**, M.S. His research interests include natural language processing, fake news detection.

**WANG Ming-Wen**, Ph.D., professor. His research

interests include natural language processing, information extraction, information retrieval, data mining.

**HUANG Qi**, Ph.D., lecturer. His research interests include social network analysis, fake news detection, sentiment analysis.

**LUO Wen-Bing**, Ph.D. His research interests include educational knowledge graph, natural language processing, personalized recommendation.

**ZHU Ying-Ting**, M.S., lecturer. Her research interests include natural language processing, computer education.

## Background

The development of the Internet and new media technologies has driven the dissemination and exchange of news. However, due to some unstable factors, news posts containing false information have spread on social platforms, causing certain social harm. Currently, fake news is developing in a multimodal form that combines text, images, and other elements.

With the advancement of large language model technology, these models, which possess vast knowledge reserves and strong semantic analysis capabilities, have been applied to the field of fake news detection. Currently, they are mainly used in text generation and text analysis. For instance, large language models are used to generate comments to simulate silent users or to analyze the overall logic and background of news. However, existing methods that solely rely on large language models to generate comments still have certain limitations, as they ignore genuine comments with human thinking and rich emotions, which can lead to weak generalization of the model. Additionally, the hallucination problem of large language models still needs to be further addressed, and the extraction of deep information from news

analysis by large language models also requires further improvement. To solve these problems, this paper proposes a multimodal false news detection method based on diverse comments and a threefold perspective, which effectively enhances the robustness and generalization ability of the fake news detection model by integrating the effective information of two types of comments and the three perspectives of self-review, peer evaluation, and comprehensive evaluation.

Experimental results on three datasets show that our proposed method outperforms the latest baseline model methods in terms of accuracy and other metrics, and surpasses the existing five large language benchmark models.

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62266023, 62466028), the Natural Science Foundation of Jiangxi Province (Grant No. 20242BAB20045), the 2023 Annual Project of Jiangxi Higher Education Society (Grant No. ZX1-B-001), and the Science and Technology Research Project of Jiangxi Provincial Department of Education (Grant No. GJJ2401903). These projects mainly focus on areas such as natural language processing and social sentiment analysis.