面向社交网络的多模态谣言检测方法:现状与未来

李雨泽[1],刘 颖[1,2*]
(1. 西安邮电大学图像与信息处理研究中心,陕西西安 710121;
2. 青海理工学院高原生态安全视觉计算联合实验室,青海西宁 810016)

摘 要: 社交网络凭借其即时、开放与传播广泛的特性,在促进信息交流的同时,也沦为谣言滋生的温床。多模态谣言融合文本及图像等多种信息形式,呈现出更强的欺骗性与煽动性,增加了网络治理的难度。因此,实现高效准确的多模态谣言检测,已成为维护清朗网络空间与公共安全的关键任务。近年来,学界针对多模态谣言的数据特性提出了多种检测方法,技术体系日趋多样。为系统梳理该领域当前的研究进展,本文从信息构成与来源视角,构建了一个涵盖内容、外部知识、社交上下文与外部环境四大维度的多模态谣言检测分类体系。区别于现有综述对内容维度的单一侧重,本文通过这一四维分析框架,揭示了谣言检测正从局部的静态内容核查向融合动态社交演化与外部环境交叉验证演进的深层范式转移。在此框架下,深入剖析了基于图文内容交互、外部知识增强、社交上下文信息以及外部环境感知四类技术方法的核心机制、演进脉络与代表性模型,并总结了其技术优势与局限性。此外,本文归纳了常用数据集与评估指标,并从纵向技术演进脉络、跨方法类别统计趋势以及特定检测场景下模型表现三个维度,对比了各类模型的效能差异与适用边界。最后,本文探讨了模态异构与特征对齐瓶颈、外部知识依赖与大语言模型(Large Language Model,LLM)幻觉、动态社会与环境建模复杂性、跨域领域偏差以及高逼真人工智能生成内容(Artificial Intelligence Generated Content,AIGC)谣言鉴伪等五大关键挑战,并对未来的研究方向进行展望。

关键词: 社交网络;多模态谣言检测;多模态融合;知识增强;社交上下文;外部环境感知
基金项目: 西安邮电大学研究生创新基金(No.CXJJBDL2024003)
中图分类号: TP391;TP393 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-28
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260187

Multi-modal Rumor Detection in Social Networks: Recent Advances and Future Trends

LI Yuze[1], LIU Ying[1,2*]
(1. *Center for Image and Information Processing*, *Xi'an University of Posts and Telecommunications*, *Xi'an*, *Shaanxi* 710121, *China*;
2. *Center for Visual Computing in Plateau Ecological Integrity*, *Qinghai Institute of Technology*, *Xining*, *Qinghai* 810016, *China*)

Abstract: With characteristics of instantaneity, openness, and widespread dissemination, social networks facilitate information exchange but have simultaneously become a hotbed for the proliferation of rumors. Integrating heterogeneous information forms such as text and images, multimodal rumors exhibit greater deceptiveness and provocativeness, thereby significantly complicating network governance. Consequently, achieving efficient and accurate multimodal rumor detection has emerged as a critical mission for maintaining a clean cyberspace and ensuring public safety. In recent years, the research community has proposed diverse detection methodologies tailored to the data characteristics of multimodal rumors, leading to an increasingly diversified technical landscape. To systematically review current research, this paper constructs a multimodal rumor detection taxonomy from the perspective of information composition and sources, encompassing four dimensions: content, external knowledge, social context, and external environment. Unlike existing reviews with a singular emphasis on the content dimension, this paper, through its four-dimensional analytical framework, reveals a profound paradigm shift in rumor detection from localized static content verification to an integrated approach that incorporates dynamic social evolution and external environmental cross-validation. Within this framework, it analyzes the core mechanisms, evolutionary trajectories, and representative models of four methodological categories: visual-textual interaction, external knowledge enhancement, social context information, and external environment perception, summarizing their technical advantages and limitations. Additionally, this paper summarizes commonly used datasets and evaluation metrics, and comparatively analyzes the performance disparities of various models from the three dimensions of longitudinal technical evolution trajectories, cross-category statistical trends, and model performance under specific detection scenarios. Finally, this paper discusses five

收稿日期:2026-04-21;录用日期:2026-05-05;责任编委:唐旭
*通信作者:刘颖

critical challenges: bottlenecks in modality heterogeneity and feature alignment, risks of external knowledge dependence and large language model (LLM) hallucinations, complexities of dynamic social and environmental modeling, cross-domain biases, and the authentication of highly realistic artificial intelligence generated content (AIGC) rumors. Furthermore, it outlines future research directions.

Keywords: social networks; multi-modal rumor detection; multi-modal fusion; knowledge enhancement; social context; external environment perception

Foundation Item(s): Graduate Innovation Fund of Xi'an University of Posts and Telecommunications (No.CXJJBDL2024003)

0 引言

近年来,以微博、Twitter等为代表的交互式社交网络平台已深度融入公众生活,成为信息获取、分享与传播的核心渠道。但其固有的随意性、隐匿性与多元化特点也使之成为谣言滋生的"温床"。谣言作为一种发布时真实性未经验证、具有社会传播风险的信息载体,在社交网络中呈现出"病毒式"的传播模式,其速度与广度往往远超真实信息[1]。根据中国互联网络信息中心发布的第57次《中国互联网络发展状况统计报告》,截至2025年12月,我国网民规模已达11.25亿,互联网普及率突破80%[2],庞大的网民基数也加剧了谣言扩散的风险。仅2025年,中国互联网联合辟谣平台共发布各类辟谣稿件1.7万余篇[3],网络谣言治理的现实需求已十分迫切。随着人工智能生成内容(Artificial Intelligence Generated Content,AIGC)技术快速迭代,高拟真的多模态谣言信息制作成本大幅降低,这也导致谣言检测与治理难度显著加大。

面对不断升级的谣言检测挑战,相关技术手段也在持续迭代不断发展。早期研究主要集中在单模态领域,从文本[4]、图像[5]或音视频[6-7]等单一内容模态中进行特征建模。这些方法多依赖手工特征或利用卷积神经网络(Convolutional Neural Network,CNN)、循环神经网络(Recurrent Neural Network,RNN)及图神经网络(Graph Neural Network,GNN)等深度学习模型自动提取特征[8]。然而,面对海量多模态谣言数据与AIGC伪造内容的双重冲击,单模态谣言检测方法逐渐显露出其固有局限。如图1所示,单模态谣言检测不仅受限于证据链条单一,难以通过多源信息互证来抵御干扰,更因语义理解缺失而无法捕捉图文不符等依赖跨模态一致性校验的谣言线索,并且,由于过度依赖特定模态的伪造痕迹,单模态模型在面对高拟真的AIGC伪造谣言时也会表现出泛化性不足。本质上,当前社交网络中的谣言特征已演变为多模态信息的关联断裂与语义冲突。因此,不再局限于对单模态模型的改进,而是,构建如图1所示的多模态联合架构,已成为应对当前复杂社交网络环境的技术必然。

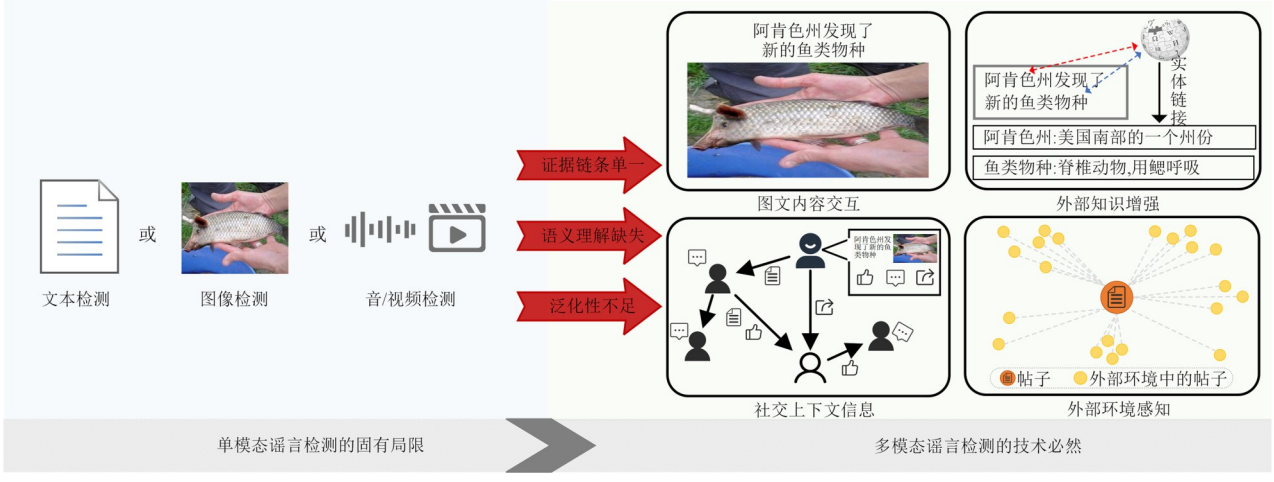


图1 谣言检测技术从单模态到多模态的演进脉络

Figure 1 Evolution of rumor detection techniques from unimodal to multimodal

梳理该领域的研究进展可以发现,尽管现有综述已对该领域进行了不同程度的总结,但在覆盖广度与时效性上仍有不足。早期研究受限于当时的技术背景,Cao等人[9]仅关注手工特征方法,高玉军等人[10]和Lakzaei等人[11]虽跟进了深度学习和GNN模型,但仍局限于单一或传统的技术框架。近期的综述开始关注多模态视角,刘华玲等人[12]从单模态到多模态进行了梳理,但其对多模态的定义主要局限于信息内

容与社交上下文,未能充分纳入外部知识与环境感知等新兴维度。Hu等人[13]虽补充了外部知识特征,却忽略了宏观社会环境对谣言传播的驱动作用。更关键的是,面对ChatGPT等大语言模型(Large Language Model,LLM)带来的技术冲击,AIGC技术已使谣言生态发生质变,现有综述未能及时审视LLM作为谣言检测引擎和谣言内容生成源的双重角色[14-16]。针对上述局限,本文从信息构成与来源视角出发,跳出了单一的谣言内容分析框架,构建了一个涵盖内容、知识、社交与环境四大维度的多模态检测分类体系。在此体系下,本文将现有的多模态谣言检测方法系统归纳为基于图文内容交互、外部知识增强、社交上下文信息以及外部环境感知四大类别,如图2所示。

具体而言,本文的贡献主要体现在以下三个方面:

(1)从信息构成与来源出发,构建了一个涵盖内容、外部知识、社交上下文与外部环境四个维度的多模态谣言检测分类体系,梳理了这四类技术方法的核心机制与演进脉络,并分析总结了其技术优势与存在的局限性。

(2)归纳了谣言检测领域内常用数据集与评估指标,并从纵向技术演进脉络、跨方法类别统计趋势以及特定检测场景下模型表现三个维度,对比分析了各类检测模型的实验性能与适用边界。

(3)总结了当前多模态谣言检测面临的五大关键挑战,并从深层语义理解与时序对齐机制、LLM赋能的统一推理框架、低资源与动态环境下的高效学习方法、跨域泛化机制与多语种综合数据集构建,以及AIGC防伪与谣言检测交叉融合五个层面,系统展望了未来的研究趋势。

本文的后续章节结构安排如下:第2节介绍相关概念及问题定义;第3节梳理四类多模态检测方法;第4节归纳数据集与实验比较与分析;第5节探讨关键挑战与未来方向;最后在第6节对全文进行总结。

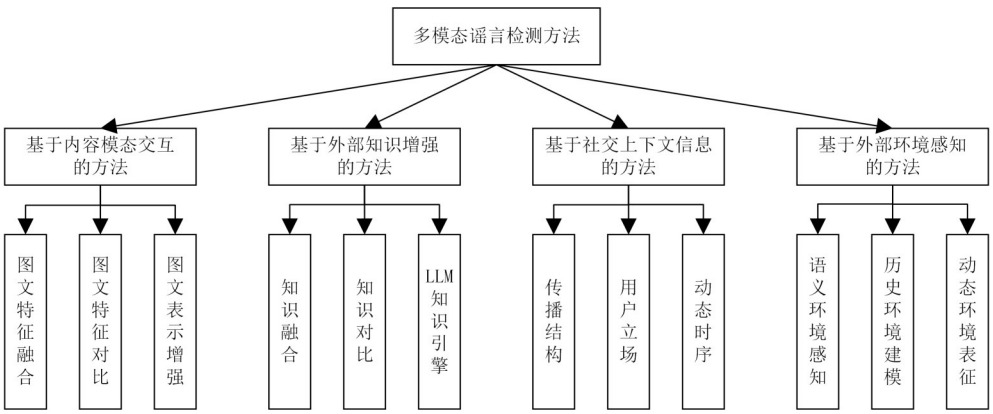


图2 多模态谣言检测方法的分类体系

Figure 2 Classification framework of multimodal rumor detection methods

1 相关概念及问题定义

1.1 谣言的定义

谣言一词在权威辞书的释义中本质特征十分明确。《现代汉语词典》第七版和《辞海》[17]中将其定义为没有事实根据的消息,《韦氏词典》[18]中明确谣言是“未经权威证实的闲话、传闻、舆论、陈述或报告”。尽管中外辞书的表述侧重略有差异,但强调了谣言的两大关键特征:未经权威验证和缺乏事实依据。

在学术研究中多沿用DiFonzo等人[19]在2007年提出的谣言经典定义“未经证实且与流通中的信息相关的陈述”。后续研究中进一步将谣言的真实性状态细分为真实、部分真实、虚假及未证实四种类型[20]。Arkaitz等人[21]通过实验证明,相较于已确证的信息,未经证实的谣言信息更易引发公众关注,更易被传播扩散。尽管部分学者[22-23]对谣言的定义侧重于虚假性,但立足于多模态检测的实际场景,本文将谣言界定为:发布时真实性未经验证,但已具备传播扩散特征的多模态信息集合,其最终真伪需通过检测模型判定。

社交网络中,谣言与虚假新闻常被交替使用。多数研究[24-25]将两者同归属于虚假信息范畴,具体分类构成如图3所示。但在概念边界上存在本质区别,谣言侧重于实时未验证性,多为突发场景下的非经证实信息;虚假新闻则强调故意捏造内容的证伪[26-27],即主观故意制造的假新闻。相较于虚假新闻的认知操纵属性,谣言的高度不确定性更容易在社交网络中引发爆炸式传播。因此,本文的定义与检测目标将紧扣谣言的未验证属性及其带来的传播风险。

1.2 多模态谣言检测的问题定义

传统单模态谣言检测通常将谣言检测建模为一

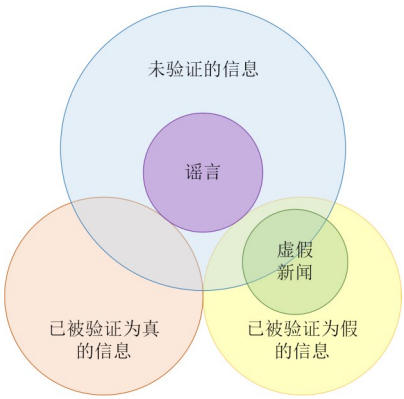


图3 虚假信息分类构成图

Figure 3 Composition diagram of misinformation classification

个经典的二分类问题。给定社交网络中存在的信息集合$P=\{p_1,p_2,\cdots,p_n\}$，集合中每条信息$p_i\in P$会被分配一个标签$y_i\in\{0,1\}$，其中$y_i=1$表示谣言，$y_i=0$表示非谣言。目标是通过检测模型F实现映射$F(p_i)\to y_i$，以最小化预测误差。

面对社交网络中内容的多模态演变，单模态视角已显局限。多模态谣言检测将输入空间进行了扩展：对任意信息p_i，其表征为模态集合$\mathcal{M}_i=\{m_i^{(1)},m_i^{(2)},\cdots,m_i^{(K)}\}(K\geqslant 2)$。其中$K$为模态数目，$m_i^{(j)}$表示$p_i$的第$j$种模态特征$\left(j\in\{1,2,\cdots,K\}\right)$。任务目标转化为学习多模态分类函数$\mathcal{F}_M$实现映射$\mathcal{F}_M\left(\Phi(\mathcal{M}_i)\right)\to y_i$。此处$\Phi$为跨模态融合算子。

本文构建的多模态谣言检测分类体系中，为了实现对复杂谣言的精准捕捉，信息的模态不再局限于文本、图像、音视频等内容特征，而是进一步补充了外部知识、社交上下文以及外部环境感知三类关键的扩展模态。这些扩展维度与内容模态互为补充，共同支撑起本文构建的多模态谣言检测架构。信息$\mathcal{M}_i$的构成被形式化为涵盖这四个正交维度的多模态集合$\mathcal{M}_i=\{\mathcal{C}_i,\mathcal{K}_i,\mathcal{S}_i,\mathcal{E}_t\}$。

（1）内容模态$\mathcal{C}$：指待检信息本身的内在属性，可形式化为$\mathcal{C}_i=\{T_i,V_i,A_i\}$，分别代表文本、视觉及音视频特征。该模态主要用于通过跨模态交互捕捉内生语义冲突。

（2）外部知识模态$\mathcal{K}$：侧重于静态、结构化的客观事实。可定义为事实三元组的集合$\mathcal{K}_i=\left\{(e_h,r,e_t)\right\}$，其中，$e_h$和$e_t$分别代表经过验证的头实体和尾实体，$r$为两者之间确定的逻辑关系。该模态为模型提供不受短期时间波动的绝对事实参考，主要用于辅助完成跨模态的逻辑核查。

（3）社交上下文模态$\mathcal{S}$：侧重于网络传播拓扑与群体反馈。可形式化为图结构$\mathcal{S}_i=(\mathcal{V},\mathcal{R})$，其中$\mathcal{V}$代表参与交互的用户节点集合，$\mathcal{R}$代表转发、评论等传播关系集合。该模态主要用于挖掘信息传播过程中的结构异常与群体立场偏差。

（4）外部环境感知模态$\mathcal{E}$：侧重于动态、非结构化的宏观时序语境。可定义为特定时间窗口$[t-\Delta t,t]$内的主流信息流与舆论分布$\mathcal{E}_t=\{d_1,d_2,\cdots,d_n\}$，其中$d_k\left(k\in\{1,2,\cdots,n\}\right)$代表该窗口内的具体外部环境信息条目。该模态通过将待检信息置于宏观演化的背景下，用于度量局部特征相较于全局背景的异常度与流行度。

1.3 多模态谣言检测的核心理论挑战

基于1.2节的问题定义，多模态谣言检测任务目标转化为学习多模态分类函数，实现映射$\mathcal{F}_M\left(\Phi(\mathcal{C}_i,\mathcal{K}_i,\mathcal{S}_i,\mathcal{E}_t)\right)\to y_i$。然而，此处的跨模态融合算子$\Phi$并非简单的特征拼接，它在理论上必须克服多模态扩展带来的三大核心挑战。

（1）信息冗余与互补性处理：多源特征的引入必然带来大量冗余信息与噪声干扰。不同模态间常包含重叠信息，直接融合会导致表征冗余。同时，某些模态也可能因生成质量、传输损耗或恶意伪造而引入噪声干扰。因此，融合算子需具备信息瓶颈能力，在压缩多模态输入$\mathcal{M}_i$与融合表征Z之间互信息以滤除冗余噪声的同时，最大化Z与标签y_i间的互信息以保留核心特征，其优化目标形式化为$\max_\theta I(Z;y_i)-\beta I(Z;\mathcal{M}_i)$。其中$\theta$为模型参数，$\beta$控制压缩强度，平衡信息保留与噪声过滤。

（2）跨模态冲突与一致性量化：模型需在多维空间中精确量化模态间的语义对齐程度。对于高伪装谣言，系统需识别深层的逻辑冲突。以内容模态$\mathcal{C}_i$与知识模态$\mathcal{K}_i$的对齐为例，模型在最小化两者表征间距离$D(\mathcal{C}_i,\mathcal{K}_i)$的同时，必须能够捕捉违反常识的异常扰动。其中$D(\cdot,\cdot)$为特征空间中的距离度量函数。

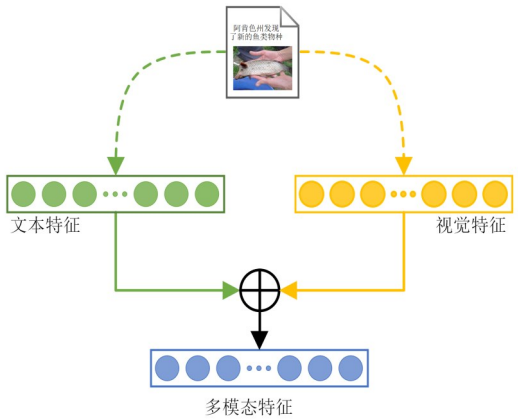
（3）跨维度的因果推理机制：为了抵御AIGC深度伪造带来的特征空间偏移，模型需超越表层的统计相关性，建立起包含事实约束与环境反事实的因果图模型。通过引入因果干预$P\left(y_i|\mathrm{do}(\mathcal{C}_i),\mathcal{K}_i,\mathcal{S}_i,\mathcal{E}_t\right)$，切断环境偏见或伪造表象带来的虚假混淆，实现深度的逻辑溯源。其中P为预测概率，$\mathrm{do}(\cdot)$为因果干预算子。

多模态谣言检测并非各扩展维度的简单特征堆砌，而是要求模型在复杂的跨模态交互中，系统性地攻克一致性量化、冗余信息过滤以及深度因果推理三大理论难题。

2 多模态谣言检测方法：模型与演进

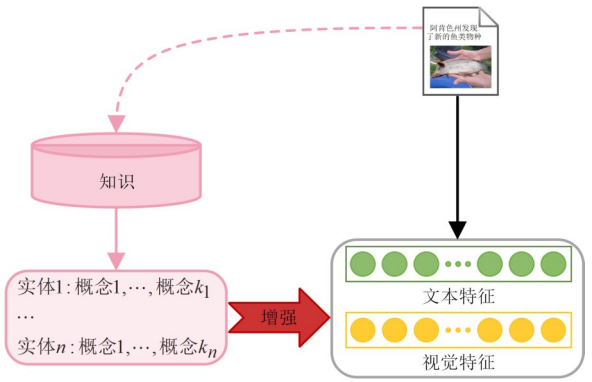
多模态谣言检测聚焦于多源信息的利用方式与建模差异，现有研究可归纳为四类主流技术路线：基于图文内容交互、外部知识增强、社交上下文信息以

及外部环境感知。如图4所示,本节将依此脉络,展开分析这四类方法的建模机理与技术演进路径。



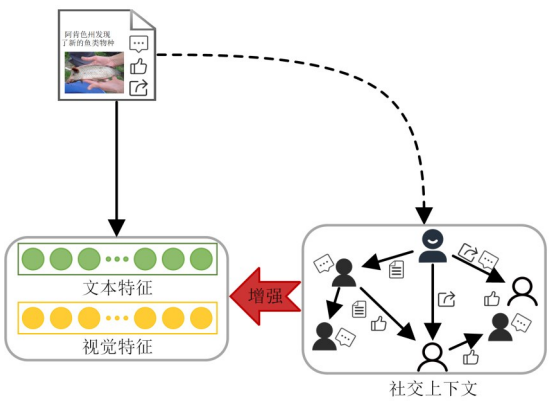
(a) 基于图文内容交互的多模态谣言检测方法

(a) Multimodal rumor detection method based on image-text content interaction



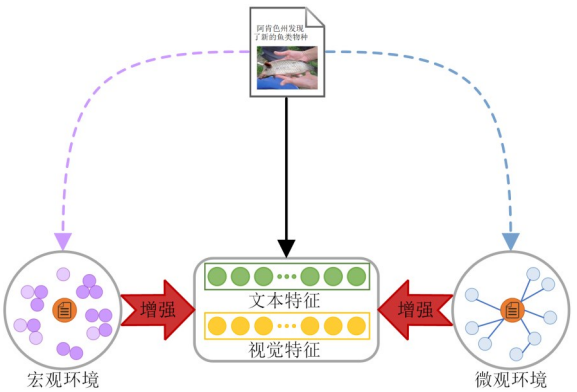
(b) 基于外部知识增强的多模态谣言检测方法

(b) Multimodal rumor detection method based on external knowledge enhancement



(c) 基于社交上下文信息的多模态谣言检测方法

(c) Multimodal rumor detection method based on social context information



(d) 基于外部环境感知的多模态谣言检测方法

(d) Multimodal rumor detection method based on external environment perception

图4 多模态谣言检测核心技术流程对比

Figure 4 Comparison of core technical processes of multimodal rumor detection

2.1 基于图文内容交互的多模态谣言检测方法

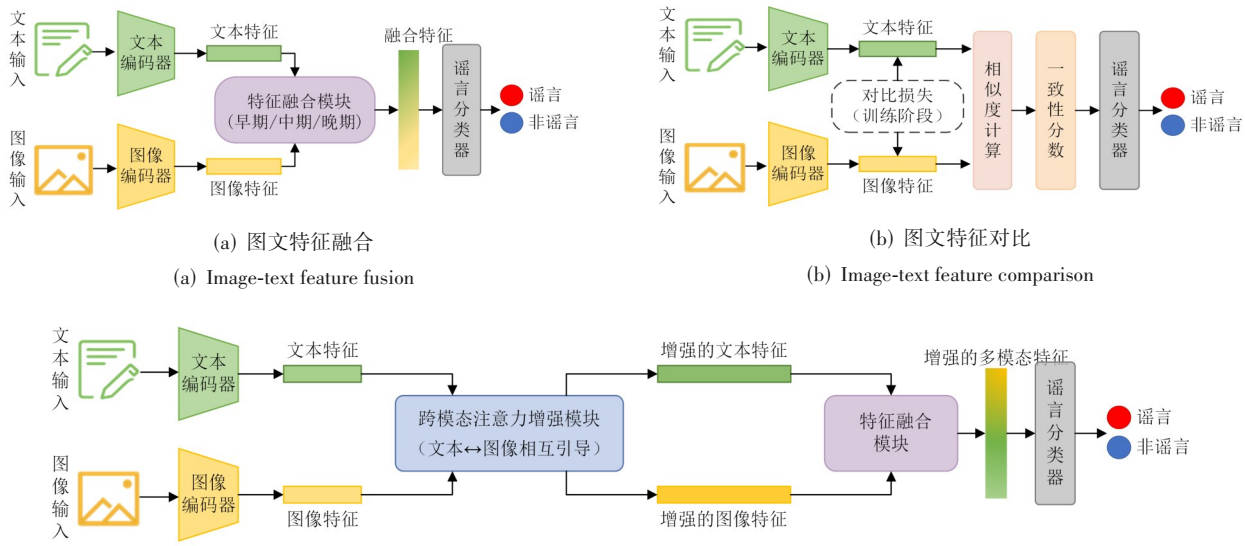
基于图文内容交互的检测方法是多模态谣言检测的基石,直接面向图文信息本身,侧重于挖掘文本与图像的内在关联,通过多源信息的联合表征来提升检测性能[13]。根据模态间交互方式不同,相关研究可分为三类:特征融合、特征对比与表示增强,其主流范式如图5所示。

2.1.1 图文特征融合

图文特征融合致力于将文本与图像整合为统一的联合表征。依据融合操作的介入阶段,现有方法可划分为早期、中期和晚期融合三类。

早期融合(即特征级融合)通常在输入层或浅层网络直接对文本与视觉特征进行拼接,模态间交互较弱[13]。Singhal等人[28]提出的SpotFake模型是其中的典型代表,该模型分别使用VGG19和BERT(Bidirectional Encoder Representations from Transformers)提取图像与文本特征,将得到的特征向量进行简单拼接后输入分类器,构成了一个基础但完整的多模态检测流程。早期融合方法虽实现简单且计算高效,但模态交互仅停留在浅层数值维度,难以挖掘深度的非线性语义关联。

中期融合(即模型级融合)通过网络中间层的交互机制,实现文本与图像间细粒度的跨模态对齐与深层语义理解。2017年,Jin等人[29]提出att-RNN率先引入注意力机制捕捉图文隐含关联,奠定了该方向基础。为进一步解决模型在特定事件上的过拟合问题,Wang等人[30]设计了一个基于生成对抗网络(Genera-



(a) 图文特征融合

(a) Image-text feature fusion

(b) 图文特征对比

(b) Image-text feature comparison

(c) 图文表示增强

(c) Image-text representation enhancement

图5 基于图文内容交互的多模态谣言检测主流范式

Figure 5 Mainstream paradigms of multimodal rumor detection based on image-text content interaction

tive Adversarial Networks, GAN)的端对端的EANN(Event Adversarial Neural Networks)模型,利用对抗学习思想剥离文本和图像特征融合后事件特有的独有特征,从而提取出更具泛化性的跨模态不变特征;Khattar等人[31]则构建了多模态变分自动编码器(Multimodal Variational Autoencoder, MVAE),通过学习图文共享的隐层分布来增强表征的鲁棒性。针对语义交互的粒度,后续研究转向了更精细的注意力设计。Hu等人[32]和Song等人[33]分别通过图像-文本匹配感知及双向协同注意力机制,强化了模态间的相互增强与语义校准;Qi等人[34]更进一步引入视觉实体来理解高层语义并对不一致性进行建模;Jing等人[35]则设计了混合器架构,实现了多层级特征的渐进式融合。近期,Wang等人[36]利用Transformer的全局建模能力实现了深度跨模态推理,并引入多准则评估机制,有效缓解了以往模型对单一不一致性指标的过度依赖。中期融合方法能有效捕捉图文不符的语义不一致性,是检测误导性配图谣言的关键。

晚期融合(即决策级融合)侧重于先独立提取模态特征或生成初步判别,再于输出层进行集成。Singhal等人[37]在SpotFake[28]的基础上提出了SpotFake+模型是该策略的典型代表,通过引入更强的预训练模型XLNet(eXtreme Language Understanding NETwork)提取文本特征,并将深度特征向量连接后输入全连接层进行决策,证明了基于高质量深度表征的晚期融合性能的决定性作用。晚期融合的最新进展聚焦于提升决策的可靠性和可解释性。Wang等人[38]设计的UEFN(Uncertainty Estimation Fusion Network)模型创新性地结合了D-S证据理论与狄利克雷分布,通过量化各模态的不确定性来指导融合权重,从而在提升精度的同时,为模型决策补充了关键的可解释性维度。

图文融合策略已从简单的特征组合,发展为复杂的语义交互与逻辑推理。目前,中期融合是捕捉深层语义线索的核心手段,而晚期融合则在提升决策置信度方面展现出巨大潜力。

2.1.2 图文特征对比

图文特征对比有别于特征融合的检测方法,侧重于保持各模态的独立性,通过引入特定的相似度度量机制,直接计算文本与视觉在语义空间中的匹配得分。其立论基础在于,真实信息通常具有高度的图文一致性,而谣言往往伴随着显著的语义冲突。因此,这种跨模态的不匹配度便成为了判定谣言的核心线索。

早期研究多依赖于显式相似度计算。经典的SAFE[39](Similarity-Aware Fake news detection)模型通过图像描述生成技术将视觉信息转化为文本描述,在统一的语义空间中计算原始文本与生成文的相似度。为解决转化过程中的误差累积,Xue等人[40]提出了基于隐式表征对齐的改进方案,采用权值共享编码器将双模态映射到统一空间进行对比学习,避免了SAFE[39]的双重依赖问题,实现了端到端的不一致性检测。

当前,特征对比技术大多转向以CLIP[41](Contrastive Language-Image Pretraining)等为代表的大规模预训练视觉-语言模型。这类模型凭借海量数据的对比训练,天然具备了强大的跨模态匹配能力。在此基础上,Zhou等人[42]利用CLIP相似度分对特征进行适应性加权,以强化关键线索;Ma等人[43]则通过引入多尺度卷积,实现了文本与图像局部区域的精细化映射。为了进一步解决决策可信度问题,Huang等人[44]引入了双证据理论框架,将语义一致性转化为可量化的证据强度,并融合不确定性估计,为检测结果提供了数理层面的可靠性保障。

特征对比方法经历了从浅层相似度计算到深层隐式对齐的转变,为识别跨模态语义冲突提供了有效的判别依据。未来的研究或将不再满足于捕捉表层的图文不匹配,而是致力于挖掘更隐蔽的深层逻辑矛盾,以提升模型在复杂语境下的鉴别精度。

2.1.3 图文表示增强

图文表示增强不同于前两者侧重跨模态交互,其直接面向数据质量,通过特定的学习策略优化文本和图像的特征分布,以确保输入信息在进入分类器前具备足够的纯净度与区分度。

早期研究侧重于通过改进网络架构来强化特征表达。Qian等人[45]提出的HMCAN(Hierarchical Multi-modal Contextual Attention Network)利用分层多模态Transformer,实现了对文本上下文与深层语义的联合建模;Guo等人[46]设计的跨模态信息增强融合网络(Cross-modal Information-enhanced Fusion Network,CIFN)则引入选择性注意力机制,通过动态加权来过滤冗余信息并强化关键特征,从而捕捉模态间的依赖关系。进阶研究进一步深入到特征内部结构的优化,针对视觉特征中的噪声干扰,Han等人[47]设计了语义解耦模块,通过将图像分离为前景与背景,直接剔除了无关的视觉信息。而为了解决模态间的语义偏差,Xing等人[48]则从空间映射角度入手,利用双向网络将图文投影至统一语义空间,并通过严格的对齐约束强迫模型学习更稳健的特征表达。图文表示增强通过优化特征表达来提升模型的判别力。未来的趋势将侧重于自监督预训练,利用海量无标注数据学习更具鲁棒性的通用表征,降低对高质量标注数据的过度依赖。

特别需要注意的是,短视频已成为当前社交网络中重要的谣言载体。上述图文匹配方法主要依赖静态语义对齐,难以直接处理短视频中融合了动态画面、音频与字幕的多维时序信息,也难以捕捉音画失步或帧间逻辑断裂等造假线索。为应对短视频带来的检测挑战,现有研究开始向视频多模态特征拓展。Wu等人[49]从模态篡改的视角出发,通过设计交叉模态匹配预训练任务,实现了具备强解释性的短视频谣言检测。在跨模态交互深化方面,Wang等人[50]设计了视觉语言模型驱动的混合专家适配器,专用于捕捉音视频间的动态语义冲突。此外,Pang等人[51]尝试借鉴视频伪造检测技术,将传统的谣言话题演化空间转化为动态的时序视频特征进行特征识别与检测。从静态图文到动态视频的延伸,正成为内容模态交互技术演进的重要方向。

综上所述,基于内容交互的方法主要以特征融合为核心,辅以特征对比与表示增强,旨在挖掘图文及音视频的内部关联,并向短视频等复杂时序模态拓展。尽管此类方法在提取模态内部关联方面已趋于成熟,但面对高伪造谣言,若其在图文或音视频层面做到逻辑自洽却完全违背客观事实,模型仅凭内部特征往往难以有效鉴别。这种事实核查机制的缺失,驱使检测方法突破单一的内容边界,向外部知识增强方向演进。

2.2 基于外部知识增强的多模态谣言检测方法

基于图文内容交互的方法主要关注内部语义的一致性,难以甄别图文匹配但违背客观事实的谣言。引入外部知识模态正是为了提供不受短期波动的绝对事实参考,弥补了单纯内容交互的短板,将检测重心从单纯的图文一致性拓展至事实真实性。现有研究主要分为知识融合、知识对比与LLM知识引擎三类。图6展示了这三类方法的典型范式与核心流程。

2.2.1 知识融合

知识融合通过将外部知识作为深层语义补充注入内容表征,以缓解社交网络中短文本的语义稀疏问题。在技术实现上,主要分为注意力机制与图拓扑结构建模。

基于注意力机制的方法将外部知识视为与文本、图像平等的一类特征源,通过设计特定的注意力网络动态融合知识信息。Dun等人[52]提出了知识感知注意力网络(Knowledge-aware attention network,Kan),通过识别信息中的实体并链接至知识图谱,进而设计双重注意力机制来融合外部知识。类似地,Zhang等人[53]构建多通道知识增强模型利用注意力机制实现文本对视觉特征的单向引导,并融入外部知识概念作为附加证据。

基于图拓扑的方法侧重于将内容与知识构建为异质信息图,将检测任务转化为图分类问题。Wang等人[54]提出的知识驱动多模态图卷积网络(Graph Convolution Network,GCN)通过构建一个包含文本实体、外部知识和视觉目标的异质图,利用GCN实现了

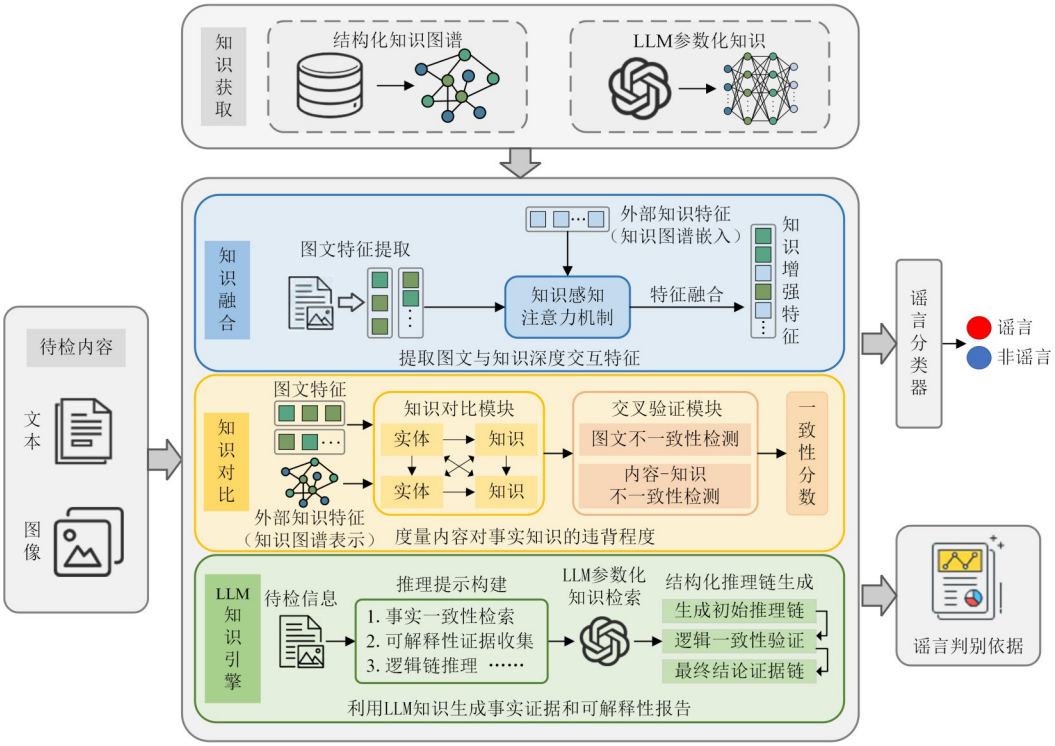


图6 基于外部知识增强的多模态谣言检测方法范式图

Figure 6 Paradigm diagram of multimodal rumor detection method based on external knowledge enhancement

多源信息的联合表征；Sun等人[55]在此基础上引入时间维度，设计双动态GCN同时捕捉传播动态与知识演化；Cui等人[56]则结合了注意力机制与GAN，在通过注意力实现知识局部注入的同时，利用对抗网络挖掘更深层的全局语义关联，显著提升谣言检测性能。

知识融合方法虽然拓宽了语义空间，但依然仅将外部知识信息作为谣言内容的补充，而非判断谣言的核心依据，导致模型对谣言直接违背常识这一关键特性的利用相对不足。

2.2.2 知识对比

知识对比将外部知识作为客观参考，不追求特征的深度整合，而是直接度量待检内容与背景知识之间的一致性来识别谣言。

Hu等人[57]提出的CompareNet模型通过实体链接将新闻描述与知识图谱中的规范化表示进行相似性比较，有效量化了内容可信度。Sun等人[58]提出的知识引导双一致性网络（Knowledge-guided Dual-Consistency Network，KDCN）进一步扩展了这一思路，能够同时检测内容跨模态间和内容-知识间的不一致性，并设计了处理视觉模态缺失的鲁棒机制。

该方法的核心优势在于能够直观量化待检内容对客观事实的偏离程度，具有较强的可解释性。然而，其检测性能高度依赖所接入知识库的准确性与覆盖范围，在面对知识库更新滞后或覆盖不全的新发突发事件时，判别效果将受到明显制约。

2.2.3 LMM作为知识引擎与推理机

LLM的应用推动了外部知识增强技术的升级，可以直接调用模型参数中内嵌的通用知识与逻辑推演能力[16,59]。图7展示了以LLM为核心的检测架构，整合了检索增强、指令微调与思维链推理等关键组件。

面对复杂的谣言检测场景，仅依赖模型内部参数知识的常规提示方法由于缺乏外部事实约束，极易诱发LLM的知识幻觉问题。为提升事实核查的严谨性，当前LLM在谣言检测中的应用已演化出检索增强生成（Retrieval Augmented Generation，RAG）、多智能体协同验证与多轮自反思机制三种结构化推理模式。

RAG机制的核心是通过集成外部检索模块实时获取客观事实，将检索到的最新背景知识与待检测信息拼接后输入LLM。这种方法能有效缓解LLM固有的知识滞后与幻觉问题，赋予模型处理突发谣言的时效性。Li等人[60]引入了多轮迭代检索机制，通过反复从互联网搜集并校验事实构建了初步的证据链闭环。为提升检索系统在复杂网络环境下的鲁棒性，He等人[61]提出了多阶段RAG优化框架，利用混合检索机制显著增强了模型面对结构化语义扰动时的抗干扰能力。Xiao等人[62]则拓展了RAG的深度，利用LLM对检索到的事实进行评估与可解释性推理，并交

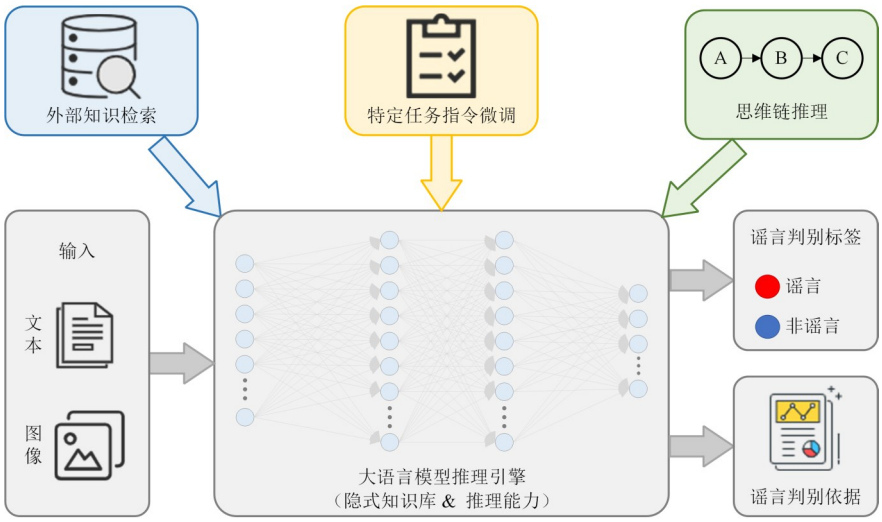


图7 LLM作为知识推理引擎与推理机的谣言检测框架

Figure 7 Rumor detection framework with LLM as a knowledge reasoning engine and inference engine

由专家模型分类,实现了从基础检索到复杂推理的深化。此外,随着多模态大模型(Multimodal Large Language Model, MLLM)的发展,Qi等人[63]提出的SNIFFER模型在实现语义对齐后,接入外部工具辅助核查,强化了脱离社交上下文场景下的谣言判别能力。然而,RAG机制的检测性能高度依赖外部检索源的质量与客观性。若检索到的外部证据本身充斥着同质化噪音、蓄意伪造的错假信息或搜索引擎固有的算法偏见,易对LLM形成误导,进而导致最终的分类决策边界发生偏移。

多智能体协同验证模式则是通过赋予LLM不同的专家角色,模拟同行评审来交叉验证所获取的外部证据与内部参数知识。这种角色扮演机制能强化对信息事实的剖析,正如Liu等人[64]构建的正反方智能体,通过结构化辩论对信息事实进行深度的证据链剖析。在此基础上,为打破不同事件间的知识壁垒,Li等人[65]设计了多专家智能体框架,实现了领域规则的自动优化,成功将协同验证的泛化能力拓展至跨领域场景。更为前沿的探索则将这种多智能体交互进一步延伸至模态之间,近期已有研究专门构建多角色系统,通过深度协同来严格评估图文多模态特征与外部检索证据的一致性[66]。这种多智能体协同机制能有效降低单一模型对事实判定的偏见,但面临计算开销大与智能体调度复杂的瓶颈。

多轮自反思与思维链旨在引导模型在调用内部参数化知识时,按特定逻辑路径推导,并对提取的中间事实依据进行自我纠错与修正。为规范模型的推导路径,柯婧等人[67]引导LLM按照主干事件、细粒度次要事件和隐含信息的顺序进行层级式推演,不依赖外部工具实现隐含语义增强。这种链式推理逻辑也被拓展至多模态场景,Xu等人[68]通过设计检查、推理及决定的多阶段提示,显著强化了图文证据核查的连贯性。除了正向的逻辑推导,模型自身的纠错能力同样是该机制的核心。Yang等人[69]引入了自反思知识构建机制,使模型自主分析历史判断中的推理错误,并建立动态内部校验准则。而Wang等人[70]则让LLM同时扮演生成器与检测器,通过内部对抗形成生成和检验的知识核查闭环。这种策略极大提升了事实逻辑链条的深度,但其局限在于,若模型在初始阶段陷入强烈的确认偏误,自反思机制极易退化为对错误知识的自我确证。

上述三种主流LLM推理范式的机制差异与优劣对比总结于表1中,LLM将知识获取方式从静态图谱检索转为参数化隐式推理。这种机制突破了传统结构化知识库的覆盖限制,使模型能够直接调用常识对谣言进行逻辑查证,但仍需配合上述结构化机制以确保事实的准确性,最终建立可靠的多模态交叉验证体系。

基于外部知识增强的方法有效弥补了单纯图文内容交互在事实核查上的短板。从知识融合与对比,到以LLM为核心的结构化推理,多模态谣言检测正逐步跨越浅层的特征拼接,具备了深度的逻辑查证能力。然而,静态知识库的更新滞后与LLM固有幻觉,使得模型难以始终获得绝对可靠的事实基准。并且,谣言在社交网络中的传播趋势是动态变化的,往往伴随复杂的群体交互与情绪演变。动态传播特性是静态的事实核查无法完全涵盖的,这就要求检测技术必须突破单一的内容查证,向融合社交上下文信息的方向继续延伸。

2.3 基于社交上下文信息的多模态谣言检测方法

谣言不仅仅是图文的组合,更是社会交互的产物。不同于前述方法侧重于静态内容的检测,基于社

表1 LLM在谣言检测中的主流推理模式对比

Table 1 Comparison of mainstream reasoning paradigms of LLM in rumor detection

方法类别	核心机制	主要技术优势	主要局限与理论瓶颈
检索增强生成	集成外部检索模块,实时获取外部事实并与待检内容拼接	缓解知识幻觉,提供高时效性的客观事实依据	依赖检索源质量,易被同质化噪音或搜索偏见误导
多智能体协同验证	赋予LLM多角色,通过多智能体交互辩论得出结论	模拟同行评审,降低单一模型偏见,提供多维解释证据	计算与时间开销巨大,面临复杂的智能体协同调度难题
多轮自反思与思维链	引导模型对中间推理步骤进行自我评估与修正,形成闭环	无需外部工具,提升逻辑推理深度与证据链一致性	初始方向错误时易陷入确认偏误的自我循环,难起纠偏作用

交上下文的方法把研究视角拓展至谣言动态传播过程,将谣言的传播结构与用户立场等信息建模为一种新的模态,通过与图文特征的交叉验证,识别那些内容逻辑正确但传播途径异常的谣言。该类方法的总体研究范式如图8所示,本节将从传播结构、用户行为及动态时序建模三个方面展开讨论。

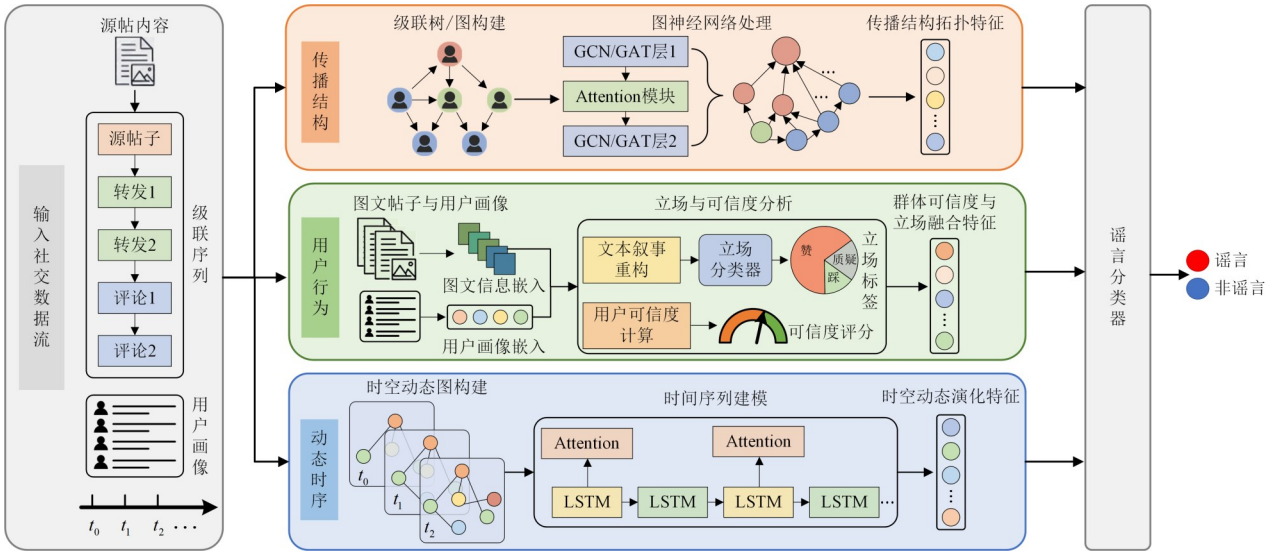


图8 基于社交上下文信息的多模态谣言检测方法范式图

Figure 8 Paradigm diagram of multimodal rumor detection method based on social context information

2.3.1 基于传播结构模态的检测方法

此类方法将传播拓扑结构视为独立模态,利用谣言在传播途径上的统计差异进行判别。早期研究依赖于手工设计的统计特征。研究者通过构建传播树,提取其深度、宽度、节点数、节点平均度等结构属性作为分类特征[71]。Ma等人[72]在谣言检测任务中首先提出了传播树核心的概念,Castillo等人[73]则利用传播树的平均深度或大小来捕捉基本传播模式。

GNN的引入推动了从手工特征向深度图表征的演进,研究方法随之分化为同构与异构网络[74-75]。同构网络侧重于单一实体(用户或帖子)的统一建模[76],以捕捉社交关联对信息扩散的宏观结构影响。Bian等人[77]提出的双向图卷积网络(Bi-directional Graph Convolutional Networks, Bi-GCN)模型是此方向的经典工作,它将用户间的回复、评论关系建模为同构的传播树,并通过Bi-GCN分别从自上而下的传播与自下而上的扩散中学习谣言深度传播的拓扑模式。在此基础上,张鑫昕等人[78]补充了用户可信度视角,利用改进的注意力模块挖掘传播树的层内依赖关系,有效细化了Bi-GCN的结构特征提取。Zheng等人[79]则聚焦于用户固有的社交拓扑,构建了基于用户双向关注关系的同构网络,通过量化用户间的兴趣分布、通信频率等同质性指标,精炼出高同质性社交圈子以进行检测。

为克服同构网络在表征复杂社交上下文方面的固有局限,研究者转向了表达能力更强的异构可信度网络[80]。通过融合用户、帖子、评论、创作者等多种类型的节点与关系,构建了更贴近真实社交生态的计算模型[81]。欧阳祺等人[82]构建了包含推文、词与用户的异质图,将异质图分解为推文-词与推文-用户子图分别进行特征学习,有效融合了文本的全局-局部语义关系与传播结构特征,实现了对语义内容与传播结构的深度耦合。Zhang等人[83]则利用帖子、创作者和主题信息构建异构网络,学习深度扩散表征。

Yang等人[84]构建了包含用户、源帖和评论的异质图以深度解析对话上下文,而Rahimi等人[85]提出的多视图框架则构建了事件与用户的交互网络以获取宏观传播视角。

近期研究开始突破单纯依靠结构建模的局限,尝试引入LLM以弥补GNN在语义推理上的短板。不同于GNN侧重于数据驱动的统计模式,LLM能够从内容层面解析传播逻辑。Li等人[86]利用LLM的生成能力为少数类节点合成文本,实现了语义级的数据增强,有效解决了类别不平衡问题;Yuan等人[87]提出的多尺度语义协同推理(Multi-Scale Semantic Collaborative Reasoning,MSSCR)模型则调用LLM解析传播路径上的节点内容,精准刻画了信息扩散时的语义演化轨迹。这些LLM探索推动了传播分析从拓扑感知逐步升级为语义理解。

2.3.2 基于用户立场与行为模态的检测方法

该类方法关注参与谣言信息传播的用户本身及其产生的反馈,将用户信誉与群体立场引入谣言检测框架,利用个体可信度与群体智慧进行交叉验证。

现有研究首先通过构建用户画像以评估个体可信度。通过提取账户属性(如注册时长、认证状态)与行为统计特征(如粉丝/关注比)[88-89],此类方法主要用于识别社交机器人等恶意传播节点[90]。另一种路径侧重于挖掘谣言评论区内的群体立场,通过分析评论和转发内容中表达的观点、立场和情感来推测原帖真实性[91]。该方向的研究由早期浅层情感特征提取逐渐转变为对深层反馈结构建模,Shu等人[92]通过句子-评论协同注意力机制对信息进行交互式编码,能同时定位源帖中的可疑句子和评论中的关键观点,而Xie等人[93]提出的立场提取和推理网络模型进一步采用图推理网络对评论间的复杂立场关系进行隐式建模,避免了繁重的人工标注。

LLM的引入推动用户立场分析从表层观测迈向深度语义推理,主要是语义重构与生成增强两类方法。Li等人[94]提出了叙事整合的元路径图自编码器(Narrative-Integrated Metapath Graph Auto-Encoder,NIMGA),该模型采用双智能体架构,利用LLM对评论进行叙事重构与演化追踪,弥补了传统立场分类在语义连贯性上的不足。Qiong等人[95]则利用LLM模拟多样化用户画像以生成代表性评论,通过合成数据策略缓解了因沉默用户导致的数据稀疏问题。这两项研究分别从语义深度与数据广度层面,强化了模型对社交舆论场的感知能力。

2.3.3 融合动态时序信息的检测方法

社交网络中的谣言传播其实也是一个动态变化的过程,仅对社交上下文的静态信息进行分析,会遗漏关键的内容。融合动态时序信息的方法引入了时间维度,可以追踪整个谣言传播过程中传播结构与群体立场的实时变化,侧重解决谣言早期检测问题。

该方向可以分为序列建模和动态图学习两类,具体如图9所示。序列建模是将传播过程视为时间序列。早期研究利用RNN等捕捉宏观的谣言传播指标波动进行建模,以捕捉谣言的生命周期模式。Chen等人[96]结合注意力机制,从帖子序列中筛选关键时序特征;Liu等人[97]则将传播路径建模为多元时间序列,融合RNN-CNN的时序分类器,实现了不依赖内容语义的早期检测。动态图模型能精细刻画谣言传播变化过程中的复杂结构关系。Huang等人[98]设计了双动态GCN,分别处理信息传播图与用户交互图,并利用交叉注意力机制融合多视角时空(Multiview Spatio-Temporal,MST)特征,实现了对动态事件的更精细表征。Xu等人[99]提出的D^2框架采用动态异构GNN联合建模社交与传播结构,基于有限观测数据预测潜在传播路径,从而强化模型对早期谣言检测的性能。

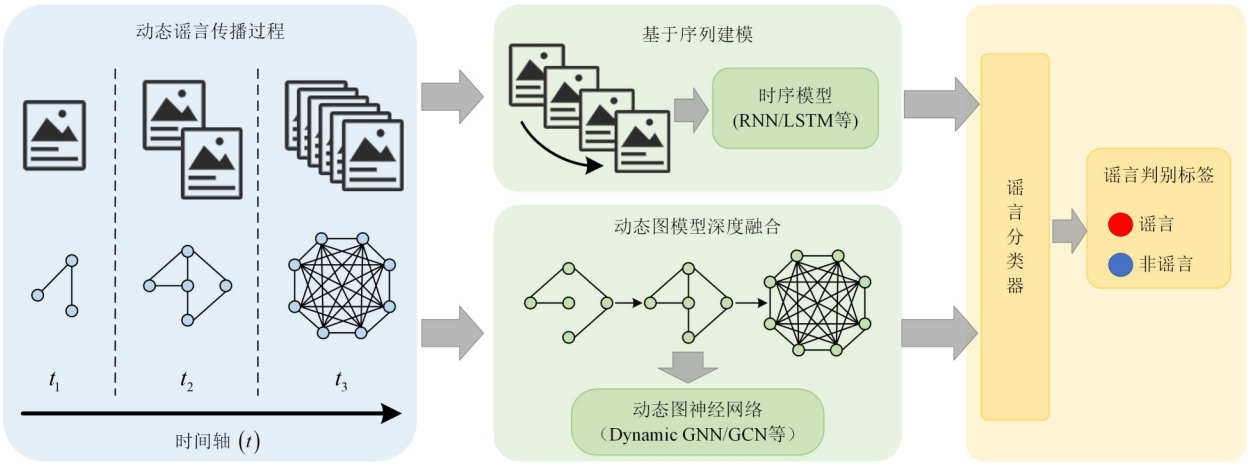


图9 基于序列建模与动态图学习的谣言检测通用框架

Figure 9 General framework of rumor detection based on sequence modeling and dynamic graph learning

近期研究开始关注模型的可解释性，尝试引入LLM解释引擎的关键角色。Zhong等人[100]设计了一种检测-解释协同架构，先利用动态图模型捕捉时空特征完成分类，再调用Llama-3-8B针对检测结果生成判别依据进行解释。这一尝试将研究重心从谣言的二分类判别拓展至考虑这条信息为何是谣言，并给出解释证据，直接提升了模型决策的透明度。

基于社交上下文的方法整合了结构、立场与时序特征。这种策略完成了从静态内容分析向动态交互分析的跨越。它擅长识别那些内容逻辑正确，但传播轨迹异常的谣言。但此类方法仍局限于信息内部的传播，忽视了谣言所存在的外部环境。这种视角的封闭性迫使研究重心从内部特性挖掘向外部环境分析延伸。

2.4 基于外部环境感知的多模态谣言检测方法

谣言的产生与传播并非孤立事件，往往紧密关联当下的热点事件或公众情绪[101]。因此，检测视角不能只局限于放大观察信息本身的内容、传播结构与知识关联，而应缩小俯瞰信息，将其置于更广阔的外部环境背景中进行分析。外部环境感知方法主要在于将待检测信息与动态演化主流媒体或公众焦点进行对照，通过度量其相较于宏观背景的差异与异常度来识别谣言。如图10所示，现有研究主要分为语义环境感知、历史环境建模与动态环境表征三类。

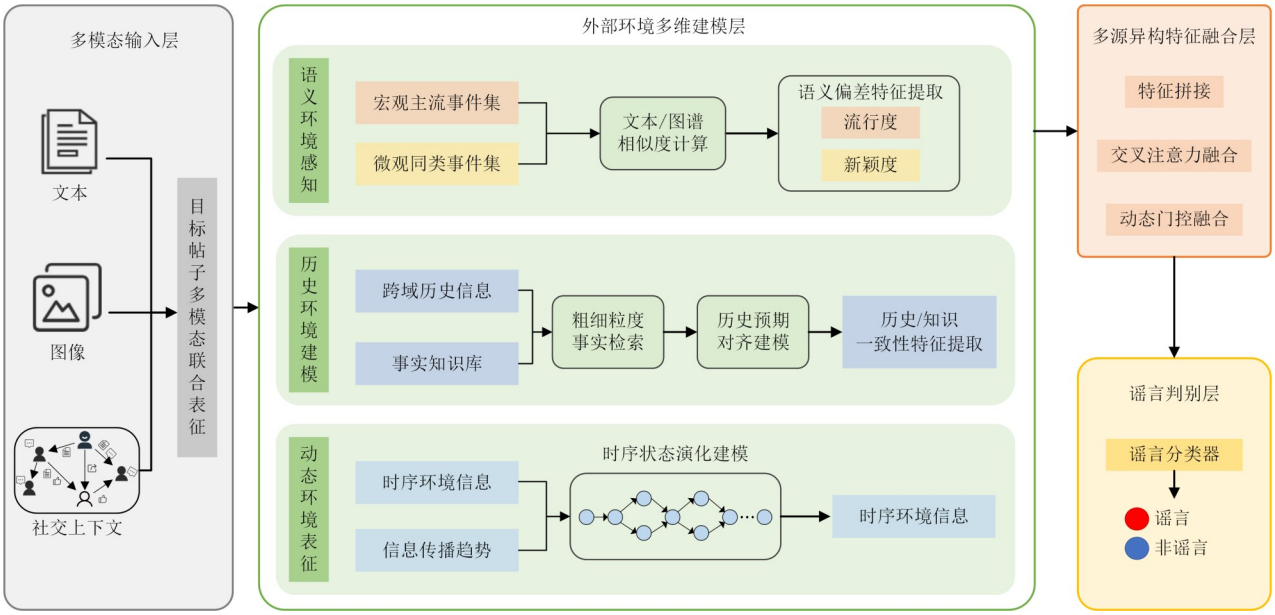


图10 基于外部环境感知的多模态谣言检测方法框架图

Figure 10 Framework diagram of multimodal rumor detection method based on external environment perception

2.4.1 基于语义环境感知的检测方法

该类方法旨在为待检测信息构建一个即时的、宏观的语义背景板，通过量化其与背景板之间的流行度与新颖度等关系来识别谣言。其基本假设是：真实信息通常与主流媒体议程保持一致，而谣言则倾向于通过制造与主流叙事相悖或过于新奇的内容来吸引眼球[101-102]。

该类方法的开创性工作是Sheng等人[101]提出的新闻环境感知框架（News Environment Perception，NEP）。该框架首次明确提出了缩放观察的研究视角，为待检测信息构建一个包含宏观环境（近期主流信息）和微观环境（紧密相关的信息）的语义环境，并分别设计了流行度导向和新颖度导向的感知模块，从中提取环境信号以增强基础检测器。这一工作奠定了语义环境感知的基本思路。NEP框架的成功启发了后续一系列改进研究。Fang等人[102]提出的新闻语义环境感知（News Semantic Environment Perception，NSEP）框架引入了时间约束区间以更精细地划分语义环境，并利用GCN来感知信息内容与宏观环境中帖子之间的语义不一致性，同时利用稀疏注意力在微观环境中捕捉语义矛盾，为早期谣言检测提供了显式证据。Liu等人[103]则将外部环境感知应用于谣言传播检测，通过图结构整合新闻的流行度、新颖度与文本信息，捕获其间的互补性高阶特征。

为了更精准地捕捉环境中的隐藏信号，Wang等人[104]采用了谱聚类算法，通过计算质心向量和熵来挖掘信息关联，利用多任务学习框架，将不同环境视为独立任务进行联合训练，提升模型的泛化能力。总体来看，即时语义环境的感知有助于捕捉谣言在宏观层面的语义差异。

2.4.2 基于历史环境建模的检测方法

语义环境感知侧重于当下,而历史环境建模则进一步将时间轴拉长,致力于从相关的历史信息中构建一个更具延续性的参考框架。其核心思想是:当前事件的发展脉络和公众对其的预期,深受其相关历史事件所形成的认知环境的影响,谣言往往会在这种预期方向上刻意制造偏差或完全违背历史规律。

该类方法分为跨领域和同领域建模。Yu等人[105]针对多领域检测问题,通过分析领域信息与其历史信息环境样本的关系构建了历史与主流信息环境,利用领域门控机制融合特征,提升了多领域泛化能力。Liu等人[106]则关注公众预期,通过构建了跨领域粗细粒度环境(Cross-domain Coarse-Fine Grained Environment,CFGE),并分别提取特征进行门控融合,利用历史语境反映的公众预期来辅助进行谣言检测。

部分研究选择从用户历史行为的角度构建环境。Xiao等人[107]通过检索用户过往发布的高相似度帖子来构建历史上下文,无需维护庞大知识库适合在线检测,为应对突发谣言信息提供了新思路。Yu等人[108]则结合了历史环境与外部知识库,形成双证据感知(Dual Evidence Perception,DEP)框架,实现了不同证据源的有效融合。

2.4.3 基于动态环境表征的检测方法

无论是语义环境还是历史环境,多数方法仍将其视为静态的背景。然而,真实的信息环境是持续演变的,谣言的真伪也可能随着新证据的出现而动态变化[109]。基于动态环境表征的方法正是为了捕捉这种时序演化特性,将环境感知提升到一个连续的、动态的新层次。

Wang等人[109]明确挑战了主流方法所遵循的静态学习方法,尝试为每个时间段学习一个社会环境表征,并利用LSTM、常微分方程或预训练动力学系统等时序模型来预测未来时间段的表征,从而使检测模型能够适应动态演化的社会环境。Li等人[110]则从宏观-微观舆论建模的角度切入动态性,通过分析微观层面传播对谣言子帖的影响,并结合宏观层面公众意见的时间变化,获得更清晰的新闻传播表征,平滑用户评论中的噪声。

在公共卫生等高风险领域,动态环境中的不确定性成为关键信号。Lu等人[111]提出的环境不确定性感知架构首次将信息环境的不确定性作为核心特征,从物理环境、宏观媒体环境、微观交流环境和信息框架四个维度进行度量,提升了疫情谣言的检测精度。此外,动态环境感知正逐步与社交传播结构走向深度融合。Wu等人[112]提出了增强传播图卷积网络(Enhanced Propagation Graph Convolutional Network,EPGCN),该模型开创性地将外部新闻环境中的相关节点直接注入谣言的微观传播图中,通过联合学习传播拓扑与宏观语境的全局结构特征,实现了外部动态环境与信息传播链条的无缝衔接。

综上所述,基于外部环境感知的检测方法通过缩放观察,拓展了检测视野,不再局限于信息的局部内容,而是关注信息宏观的外部动态环境,也经历了从静态参照,到挖掘历史关系,再到建模动态演化的过程。当前研究倾向于将环境与社交上下文及外部知识结合,同时利用谣言内外部信号协同工作[108,112]。然而,如何高效、实时地构建与更新环境,如何量化与融合环境中的复杂信号并降低对高质量环境数据的依赖,仍是未来研究面临的关键挑战。

2.5 总结与分析

本节梳理了四类主流的多模态谣言检测方法,分别是基于图文内容交互、外部知识增强、社交上下文以及外部环境感知,表2进一步从核心机制、主要优势与局限及深层理论瓶颈等方面,对上述方法进行了综合对比。

表2 多模态谣言检测方法综合对比

Table 2 Comprehensive comparison of multimodal rumor detection methods

方法类别	核心机制	关键技术	代表性模型	主要优势	主要局限与理论瓶颈
图文内容交互(2.1)	聚焦图文模态间语义对齐与逻辑冲突检测	特征融合 特征对比 表示增强	EANN[30] SAFE[39] HMCAN[45]	实现相对简单,能有效检测图文不符等表层不一致性	严重依赖内容质量,面对高拟真AIGC内容时易因缺乏事实约束而导致跨模态对齐失效
外部知识增强(2.2)	引入外部知识作为客观事实基准进行验证	知识融合 知识对比 LLM推理	KAN[52] KDCN[58] SNIFFER[63]	为检测提供可验证的事实依据,增强模型推理能力,可解释性强	效果受限于知识库完备性,且存在知识幻觉传导风险,LLM产生的伪证据易污染融合表征并导致决策边界偏移
社交上下文感知(2.3)	利用信息传播的动态模式与群体反馈作为辅助证据	传播结构 用户立场 时序动态	Bi-GCN[77] NIMGA[94] MST[98]	不依赖内容语义,能识别内容伪装精巧但传播行为异常的谣言,利于早期检测	传播初期数据稀疏时效果受限,且GNN聚合机制极易过度放大微观的短期极端情绪,掩盖早期的微弱拓扑异常
外部环境感知(2.4)	构建宏观舆论参照系,度量信息偏离度	语义环境 历史环境 动态环境	NEP[101] DEP[108] Misder[109]	提供全局性评估,能识别与主流信息流格格不入的异常信息,抗伪造性强	环境构建成本高,且宏观无关背景与微小语义偏差易被过度聚合,导致动态环境噪声放大并淹没核心事实线索

如表2所示，多模态谣言检测的研究重心已不再局限于单一的图文内容分析，而是逐步向融合知识、社交与环境的多维信息空间拓展，这也反映了研究人员对谣言本质认知的深化，谣言不是孤立的静态信息片段，而是一个具有动态演化特性的复杂实体。然而，深入剖析上述方法的底层机制可知，现有模型在应对复杂社交网络环境时仍存在特定的理论瓶颈。

（1）基于图文内容交互的方法中，跨模态对齐失败机制主要源于特征空间映射的浅层性与事实推理的缺失。现有的对齐机制主要依靠在特征空间中进行距离度量来判断是否对齐。这种方法本质上是在学习图文之间的统计相关性，但在面对高拟真的AIGC伪造内容时，由于缺乏外部事实约束，检测模型极易被其表面的语义一致性所欺骗，导致基于距离度量的对齐惩罚机制失效。

（2）基于外部知识增强的方法中，尽管LLM作为知识推理引擎时会弥补静态知识库的覆盖短板，但极易引发知识幻觉的传导风险。若缺乏严格的检索增强等事实校验约束，LLM产生的幻觉内容会转化为看似合理的伪证据。在多模态特征融合阶段，这种具有高语言连贯性的伪证往往会被交叉注意力机制赋予异常高的权重，进而沿着网络层级污染原始的图文表征，导致分类器的决策边界发生严重偏移。

（3）基于社交上下文与外部环境感知的方法中，共同面临动态环境噪声放大问题。真实网络环境中充斥着大量微观的情绪化宣泄与宏观的无关背景讨论。当利用GNN等模型对这些动态拓扑建模时，基于消息传递的聚合机制很容易将处于传播中心的短期极端情绪与微小的宏观语义偏差过度聚合。这种机制会导致动态环境噪声放大，掩盖谣言早期的微弱拓扑异常与语义新颖度信号。

不同的多模态谣言检测方法并非相互替代，而是形成了多视角的协同互补。单一的内容分析难以应对真图假文类别的谣言，结合社交或环境信息能有效识别此类谣言。基于知识的方法虽然精确，但受限于静态知识库的覆盖范围，引入LLM丰富的参数化知识与推理能力则有望突破这一瓶颈。当前研究致力于融合不同方法的优势，将知识增强与环境感知结合[108]，或者关联传播结构与外部环境[112]，最终的目标都是构建更鲁棒的检测系统。

3 实验比较与分析

本节汇总了常用公开数据集与评价指标，并对代表性模型的实验结果进行系统对比，重点剖析了不同方法的性能差异与技术优劣。

3.1 数据集

构建高质量的多模态谣言数据集耗时耗力，导致该领域的公开资源相对有限[113]。多数研究通常依赖现有公开数据集，且数据源主要来自Twitter和微博等平台。表3梳理了相关数据集。

表3直观展示了公开数据集的整体演变趋势。早期数据集仅含纯文本信息，后期逐渐融合了图像与社交上下文内容。近年来以FakeSV[114]和FakeTT[115]为代表的新数据集进一步引入了音视频模态，并且还有研究开始引入外部环境信息[101]。这也反映了谣言检测研究重心正从单一内容向多维度的环境延伸。同时，数据集的规模与复杂性也在逐步提升，总数据量持续增长，涉及领域从通用的社会问题细分出针对新冠疫情、特定方言等垂直领域，标签体系也由二分类发展为精细化多分类。

尽管新数据集不断涌现，但PHEME[21]、Weibo[116]和Twitter15/16[72]等经典数据集凭借其丰富的多模态信息和规范的构建流程，目前仍是评估基于图文内容交互与基于社交上下文信息的检测算法最常用且最具可比性的基准[117]。而近年发布的Chinese[101]、English[101]及MR^2[118]等数据集则推动了研究前沿，Chinese[101]和English[101]专为验证基于外部环境感知的检测算法而设计，MR^2[118]则通过提供检索到的外部证据，为基于外部知识增强的检测方法设立了新的评估标准。

表3 谣言检测公开数据集概览

Table 3 Overview of public datasets for rumor detection

数据集	年份	平台	具体模态包含					语言	总数据量	标签	标签数据量	领域
			文本	图像	音视频	社交上下文	外部环境					
KWON[119]	2013	Twitter	√	—	—	—	—	英文	102	谣言 非谣言	47 55	社会问题
MediaEval[120]	2015	Twitter	√	√	—	—	—	英文	13 924	谣言 非谣言	7 898 6 026	社会问题
PHEME[21]	2016	Twitter	√	√	—	√	—	英文	7 507	谣言 非谣言	1 972 3 830	社会问题

续表

数据集	年份	平台	具体模态包含					语言	总数据量	标签	标签数据量	领域
			文本	图像	音视频	社交上下文	外部环境					
Weibo[116]	2016	Weibo	√	√	—	√	—	中文	4 664	谣言 非谣言	2 313 2 351	社会问题
Twitter15[72]	2016	Twitter	√	√	—	√	—	英文	1 490	非谣言 错误谣言 真实谣言 未证实谣言	374 370 372 374	社会问题
Twitter16[72]	2016	Twitter	√	√	—	√	—	英文	818	非谣言 错误谣言 真实谣言 未证实谣言	205 205 205 203	社会问题
RumorEval[121]	2017	Twitter	√	—	—	—	—	英文	325	错误谣言 真实谣言 未证实谣言	145 74 106	社会问题
Jin-Weibo[29]	2017	Weibo	√	√	—	—	—	英文	9 528	谣言 非谣言	4 749 4 779	社会问题
GossipCop[122]	2020	Twitter	√	—	—	—	—	英文	1 043	谣言 非谣言	619 424	社会问题
ArCov19[123]	2021	Twitter	√	—	—	—	—	阿拉伯语	3 157	谣言 非谣言	1 480 1 677	Covid-19
Covid-19 Rumor[124]	2021	News Twitter	√	—	—	—	—	英文	6 934	错误谣言 真实谣言 未证实谣言	3 664 1 752 1 518	Covid-19
Weibo-all[125]	2021	Weibo	√	—	—	—	—	中文	8 050	谣言 非谣言	3 851 4 199	社会问题
Chinese[101]	2022	Twitter Weibo	√	—	—	—	√	中文	39 066	真实 错误	19 543 19 523	社会问题
English[101]	2022	Twitter Weibo	√	—	—	—	√	英文	6 483	真实 错误	3 293 3 190	社会问题
Facebook-C[126]	2023	Facebook	√	—	—	√	—	中文粤语	1 925	谣言 非谣言	565 1 360	社会问题
MR^2[118]	2023	Twitter Weibo	√	√	—	√	—	英文 中文	14 700	谣言 非谣言 未证实谣言	3 172 4 927 6 601	社会问题政治 公共卫生 自然科学
FakeSV[114]	2023	抖音 快手	√	√	√	√	—	中文	5 538	真实 虚假 辟谣	1 827 1 827 1 884	社会问题
FakeTT[115]	2024	TikTok	√	√	√	√	—	英文	1 991	真实 虚假	819 1 172	社会问题

此外,现有数据集大多局限于单一平台和语种,导致模型在实际应用中面临明显的领域偏差问题。由于不同平台用户的表达习惯和信息传播机制存在差异,在单一平台训练的模型往往难以直接迁移到其他特定的网络环境,因此跨平台的谣言检测正成为急需突破的瓶颈[127]。同时,当前主流的LLM多依赖英文语料进行预训练,在处理中文等其他语言的谣言时,容易因为缺乏相应语言环境的文化背景和语义常识而导致识别准确率下降[14,24]。而且,当前跨平台传播的谣言正向短视频等音视频形式演变[114-115],如何克服跨平台与跨语言的领域偏差,提升模型的跨域泛化能力,是未来多模态谣言检测走向实际应用急需

解决的关键问题。这也说明未来仍需构建涵盖音视频模态、具备高质量标注且能反映跨文化与跨平台传播特性的大规模数据集。

3.2 评价指标

为了客观评估多模态谣言检测模型的性能，现有研究通常采用基于混淆矩阵的全局分类指标，以及针对早期检测的时间敏感性指标。在具体计算中，真正例（True Positive，TP）代表正确识别的谣言，真负例（True Negative，TN）代表正确识别的非谣言，假正例（False Positive，FP）代表将非谣言误判为谣言，假负例（False Negative，FN）代表将谣言误判为非谣言。社交网络里的谣言传播速度极快，如果只看最终的分类结果显然不够，还需要评估模型在谣言刚发酵时的响应速度。因此，部分研究在基础分类之外，补充了早期检测准确率和检测延迟等时间敏感性指标。表4汇总了这些常用的评价标准。

表4 多模态谣言检测常用评价指标

Table 4 Overview of common evaluation metrics for multimodal rumor detection

指标类型	指标名称	计算公式与说明	核心定义与侧重点
全局分类	准确率 (Accuracy)	$\frac{TP+TN}{TP+TN+FP+FN}$	模型总体预测正确的比例，直观但易受类别不平衡影响
	精确率 (Precision)	$\frac{TP}{TP+FP}$	衡量模型预测结果的可靠性，即预测为谣言的样本中真正是谣言的比例
	召回率 (Recall)	$\frac{TP}{TP+FN}$	衡量模型的查全能力，即所有真实谣言中被模型成功找出的比例
	F1分数 (F1 Score)	$2\times\frac{Precision\times Recall}{Precision+Recall}$	精确率与召回率的调和平均数，是综合评价模型性能的核心指标
	宏平均F1 (Macro-F1)	$\frac{1}{N}\sum_{i=1}^{N}F1_i$	对所有类别的F1分数求算术平均，平等对待每一类，衡量模型在多分类任务上的整体性能
早期检测	早期检测准确率 (Accuracy @T)	限制模型仅获取发布后T小时内，或前T个转发评论的稀疏上下文进行预测的准确	侧重评估模型在谣言爆发初期多模态证据不完备条件下的提前预警能力
	检测延迟	$T_{delay}=t_p-t_0$ t_0为信息发布时刻 t_p为模型给出正确预测时刻	侧重量化模型从异常信息出现到成功触发发现机制的响应速度

在基础的全局分类中，由于Weibo[116]和Twitter15/16[72]主流数据集的正负样本比例相对均衡，准确率常被直接拿来当作直观的性能参考。现有研究也倾向于将F1分数作为核心评判依据，因为它有效权衡了精确率与召回率，能更深入地反映模型识别的鲁棒性。对于多分类任务，宏平均F1通过平等考量各个类别的表现，能客观衡量模型的整体性能，避免结果被单一数据较多的类别带偏。而实际场景下谣言的危害会随时间成倍放大，引入时间敏感性指标去计算模型在极少证据条件下的早期检测准确率和发现延迟，可以直观反映出模型的提前预警能力，这对于真正落地搭建实时的谣言拦截系统非常关键。

3.3 实验分析

本节将从纵向技术演进脉络、跨方法类别统计趋势以及特定检测场景下模型表现三个维度，对各类谣言检测方法的代表性模型进行对比分析。表5详细汇总了相关模型的方法类别、模态构成、性能增益及主要创新点。

从谣言检测的纵向技术演进脉络来看，模型性能呈现出阶梯式上升趋势。早期基于特征工程的单模态方法受限于表达能力，准确率普遍低于80%。随后，深度学习与跨模态融合技术的引入推动了谣言检测领域的首轮性能增长，以att-RNN[29]、EANN[30]为代表的基于图文内容模态的中期融合模型通过注意力机制、对抗训练等方式挖掘模态间深层关联，将模型准确率推升至80%以上。近年来，得益于BERT和CLIP等预训练模型及LLM的强大表征能力，该领域迎来了新一轮突破。Bi-GCN[77]在Weibo[116]数据集上准确率高达96.1%，MSSCR[87]、CIFN[46]等模型在PHEME[21]数据集上的准确率也突破90%，这些实验结果表明多模态检测技术正日趋成熟。

从跨方法类别的统计趋势来看，多模态信息的引入均带来了显著的性能增益，且展现出鲜明的优势侧重。基于图文内容交互的方法相较于早期基线平均取得了15%左右的准确率提升，其在识别图文不一致型谣言时表现优异，尤其是SAFE[39]在GossipCop[122]数据集上取得83.8%的准确率。然而，仅关注信息内部逻辑难以应对违背客观事实的高伪装谣言，实证数据显示，在PHEME[21]数据集上，仅基于内容融合模型的性能普遍不及引入外部知识的模型。基于外部知识

增强的方法进一步显著提升了检测效能的上限,以KAGN[56]、KDCN[58]等为代表的常规知识融合与对比模型,平均性能增益达到了11%至18%;而以SNIFFER[63]为代表的LLM推理引擎前沿模型更是取得了突破性进展,其在面对NewsCLIPpings[128]数据集时准确率高达88.4%,相较于基线模型SAFE[39]实现了35.6%的显著性能提升,充分凸显了隐式知识推理的强大潜力。此外,基于社交上下文信息的方法其平均性能增益约为11%至18%,在特定数据下表现极佳。在传播结构清晰、数据完备的Weibo[116]数据集上,Bi-GCN[77]取得了96.1%的极高准确率,证实了宏观传播拓扑在谣言鉴伪中的决定性能力。另一方面,作为宏观补充视角的外部环境感知方法,拓展了谣言检测的研究边界。自NEP[101]首次提出缩放观察的研究视角以来,该路线确立了新闻环境作为独立模态的有效性,并在Chinese[101]数据集上奠定了76.4%的基准性能。面对宏观舆论噪声的干扰,该方向正全面转向与其他模态的深度融合,并展现出显著的协同增强效应。DEP[108]通过融合历史环境与外部知识,将准确率推升至83.6%,EPGCN[112]则将新闻环境节点直接注入微观传播图以实现与社交上下文的联合建模,在Weibo[116]和Chinese[101]的联合数据集上取得93.6%的极高精度,较单一结构基线Bi-GCN[77]提升12.7%。这两项工作充分证实了将外部环境与异构模态优势互补的巨大潜力。

表5 多模态谣言检测代表性方法及性能对比

Table 5 Representative methods and performance comparison for multimodal rumor detection

研究工作	年份	方法类别	模态构成					数据集	评价指标		性能增益(vs. Baseline)	主要创新点
			内容		外部知识	社交上下文	外部环境		准确率	F1分数		
			文本	视觉								
Yang等人[4]	2012	单模态文本SVM分类器	√	—	—	—	—	自建微博数据集[4]	0.780	—	—	首次系统性地将SVM与多维度特征工程应用于新浪微博的谣言检测
Galitsky等人[129]	2015	单模态文本网页挖掘	√	—	—	—	—	自建产品评论数据集[129]	0.672~0.806	—	—	基于网页挖掘与句法分析,通过实体替换识别谣言
att-RNN[29]	2017	图文内容交互中期融合	√	√	—	—	—	MediaEval[120]	0.664	0.689	NeuralTalk[130] ↑7.2%	开创性地使用注意力机制融合图文特征,首次揭示了图文间的隐含关联
								Jin-Weibo[29]	0.772	0.764	NeuralTalk[130] ↑6.2%	
EANN[30]	2018	图文内容交互中期融合	√	√	—	—	—	MediaEval[120]	0.715	0.719	att-RNN[29] ↑5.1%	提出事件GAN,通过对抗训练学习事件不变特征,提升模型跨事件泛化能力
								Jin-Weibo[29]	0.827	0.829	att-RNN[29] ↑4.8%	
MVAE[31]	2019	图文内容交互中期融合	√	√	—	—	—	MediaEval[120]	0.745	0.758	att-RNN[29] ↑8.1%	采用变分自编码器学习图文共享隐层表示
								Jin-Weibo[29]	0.824	0.758	att-RNN[29] ↑5.2%	
Spotfake+[37]	2020	图文内容交互晚期融合	√	√	—	—	—	MediaEval[120]	0.777	0.82	MVAE[31] ↑3.2%	提出不依赖子任务的纯净多模态框架,通过BERT与VGG-19的特征拼接实现对多模态谣言的高效检测
								Jin-Weibo[29]	0.892	0.932	MVAE[31] ↑6.8%	
SAFE[39]	2020	图文内容交互图文特征对比	√	√	—	—	—	GossipCop[122]	0.838	0.895	att-RNN[29] ↑9.5%	通过图像描述生成与相似度计算,实现可解释一致性检测

续表

研究工作	年份	方法类别	模态构成					数据集	评价指标		性能增益 (vs. Baseline)	主要创新点
			内容		外部知识	社交上下文	外部环境		准确率	F1分数		
			文本	视觉								
HMCAN[45]	2021	图文内容交互图文表示增强	√	√	—	—	—	PHEME[21]	0.881	0.834	MVAE[31] ↑2.9%	提出分层多模态上下文注意力网络,通过联合建模文本上下文与分层语义,提取信息量更丰富的高阶特征
CIFN[46]	2024	图文内容交互图文表示增强	√	√	—	—	—	PHEME[21]	0.901	0.882	SAFE[39] ↑7.8%	采用选择性注意力机制增强跨模态特征表示,动态强化关键信息
KAN[52]	2021	外部知识增强知识融合	√	—	√	—	—	PHEME[21]	0.783	0.746	DKN[131] ↑5.7%	设计知识感知注意力网络,将知识图谱信息动态融入多模态表征
KAGN[56]	2023	外部知识增强知识融合	√	—	√	—	—	PHEME[21]	0.865	0.834	DKN[131] ↑13.8%	提出知识感知注意力与GNN模型,通过局部注意力和全局图结构双重机制融合外部知识
								Twitter15[72]	0.892	0.895	Bi-GCN[77] ↑0.6%	
								Twitter16[72]	0.901	0.897	Bi-GCN[77] ↑2.1%	
KDCN[58]	2023	外部知识增强知识对比	√	√	√	—	—	PHEME[21]	0.862	0.832	SAFE[39] ↑8.0%	提出知识引导的双一致性网络,同步检测跨模态与知识的一致性,并对视觉模态缺失具有鲁棒性
								Weibo[116]	0.943	0.942	SAFE[39] ↑1.8%	
SNIFFER[63]	2024	外部知识增强LLM推理	√	√	√	—	—	NewsCLIPpings[128]	0.884	-	SAFE[39] ↑35.6%	通过两阶段指令微调提升跨模态差异感知与推理能力,并能生成解释
Bi-GCN[77]	2020	社交上下文感知同构可信度网络	√	—	—	√	—	Weibo[116]	0.961	0.961	Castillo等[73] ↑13.0%	提出Bi-GCN,通过构建自上而下和自下而上的传播图,同时捕捉谣言深度传播与广度扩散结构
MSSCR[87]	2025	社交上下文感知多尺度语义推理	√	√	—	√	—	PHEME[21]	0.903	0.876	SAFE[39] ↑8.1%	提出多尺度语义协同推理框架,集成跨模态对齐、传播结构分析与LLM细粒度语义解析,提升检测可解释性
								Weibo-all[125]	0.910	0.896	SAFE[39] ↑5.3%	
NIMGA[94]	2025	社交上下文感知用户立场与行为	√	—	—	√	—	Weibo[116]	0.934	0.933	Hg-sl[80] ↑3.0%	采用双智能体架构,通过LLM实现评论的叙事重构与语义演化追踪,突破了传统立场分类的局限
MST[98]	2025	社交上下文感知动态时序	√	—	—	√	—	PHEME[21]	0.864	0.853	Bi-GCN[77] ↑1.7%	提出基于双动态GCN网络的多视角时空学习模型,联合建模信息传播与用户交互的时空演化,实现早期检测
								Weibo[116]	0.958	0.959	Bi-GCN[77] ↑3.9%	

续表

研究工作	年份	方法类别	模态构成					数据集	评价指标		性能增益(vs. Baseline)	主要创新点
			内容		外部知识	社交上下文	外部环境		准确率	F1分数		
			文本	视觉								
NEP[101]	2022	外部环境感知语义环境	√	—	—	—	√	Chinese[101]	0.764	0.789	—	开创缩放观察视角,为待检信息构建宏观与微观信息环境,通过流行度、新颖性感知提升检测性能
								English[101]	0.716	0.716	—	
CFGE[106]	2023	外部环境感知历史环境	√	—	—	—	√	Chinese[101]	0.791	0.800	NEP[101] ↑2.7%	构建跨领域粗细粒度历史环境,捕捉公众预期以指示谣言方向;通过门控机制融合粗细环境嵌入
								English[101]	0.727	0.705	NEP[101] ↑1.1%	
DEP[108]	2024	外部环境感知历史环境双证据感知	√	—	√	—	√	Chinese[101]	0.836	0.854	NEP[101] ↑1.7%	提出双证据感知框架,首次将历史信息环境与外部知识库证据融合,通过域门机制协同增强检测
EPGCN[112]	2024	外部环境感知社交上下文感知	√	—	—	√	√	Weibo[116]+Chinese[101]	0.936	0.936	Bi-GCN[77] ↑12.7%	开创性地将外部新闻环境节点直接注入微观传播图中,利用GCN联合学习多维度结构特征
Misder[109]	2025	外部环境感知动态环境表征	√	—	—	—	√	Chinese[101]	0.856	0.768	EANN[30] ↑1.2%	提出动态环境表征框架,通过LSTM/常微分方程/预训练动力学系统建模社会环境演化,适应信息真实性的动态变化趋势
								GossipCop[122]	0.854	0.799	EANN[30] ↑0.8%	

从特定检测场景下的模型表现来看,各类方法的适用边界得到了进一步凸显。在早期检测场景中,由于谣言发布初期的社交反馈稀疏,依赖长程拓扑的传统图模型面临特征提取失效的困境。早期工作尝试利用循环与卷积网络联合提取初始传播路径的序列特征以实现早期预警[97],基于动态时序的MST[98]进一步突破静态限制,联合建模信息与用户的时空演化,无需等待传播树成型即可敏锐捕捉初始异常,展现出极佳的时效性。在跨领域与跨事件泛化场景下,由于不同领域谣言的表层特征差异巨大,传统融合方法极易产生过拟合。为此,引入对抗学习剥离事件独有属性以提取跨事件的通用不变特征的EANN[30],以及构建跨领域粗细粒度历史环境以捕捉公众预期的CFGE[106],均有效挖掘了深层不变特征,大幅提升了模型面对未知突发谣言的泛化鉴伪能力。

综合上述三个维度的实验对比分析可以发现,当前尚未出现能够适配所有复杂场景的单一最优模型。各类技术路线的优势侧重点均不同,内容融合聚焦信息内部模态冲突,知识增强侧重客观事实校验,社交上下文与环境感知则分别捕捉异常传播路径与宏观语境偏离。特别是面对早期预警的高时效要求与跨领域泛化的强鲁棒性挑战,单一视角的局限性尤为凸显。因此,对于多模态谣言检测领域,未来研究的核心突破将不再局限于单一模态的优化,而是需针对具体的数据特性对异构证据源实施自适应融合,最终目标是协同互补信息,构建兼具高精度、强泛化力与决策透明度的综合检测系统。

4 关键挑战与未来方向

尽管多模态谣言检测研究已取得明显进展,但现有方法在实际应用中仍面临一系列挑战。本节将梳理该领域现存的关键挑战,并指出潜在的未来研究方向。

4.1 面临的关键挑战

当前,多模态谣言检测面临的关键挑战集中体现在模态异构与短视频特征对齐瓶颈、外部知识依赖与LLM幻觉风险、动态社会感知与低资源环境建模的复杂性、跨平台与跨语言的领域偏差,以及高逼真AIGC谣言的鉴伪挑战五个方面。

(1)模态异构性与短视频特征对齐瓶颈。多模态

谣言检测的效果取决于异构信息的融合质量。不同模态的数据结构与语义表征存在显著差异,为深层对齐带来困难。现有研究已尝试了从早期的浅层特征拼接[28],到深层细粒度交互的中期融合[29],再到基于高质量表征的晚期融合[37]。尽管融合机制不断深化,但面对模态缺失或跨模态权重自适应分配时,常规方法依然受限[35]。这种对齐能力的短板导致模型难以捕捉隐晦的模态逻辑冲突。

随着TikTok、抖音、YouTube Shorts等短视频平台的兴起,这一挑战变得更加突出。短视频谣言融合了动态视觉帧、语音旁白、背景音乐以及时序弹幕等多种信息流,直接提升了模态异构性的维度。现有基于图文匹配的检测方法难以直接迁移至短视频场景,因为它们无法有效处理音画异步、时序剪辑等动态伪造线索[49-50]。短视频的普及对多模态融合技术提出了更高要求,针对短视频的高维时空特征对齐,仍是当前亟待突破的技术瓶颈。

(2)外部知识依赖的局限性与LLM幻觉风险。外部知识的引入虽能增强模型的逻辑推理深度,但其实际效果会受知识源本身的可靠性影响。传统结构化的知识库存在覆盖率不足与更新滞后的问题,无法及时应对突发的谣言事件,从而限制模型的泛化能力[52,56]。而LLM虽提供了丰富的参数化隐性知识,但受限于固有的幻觉问题,容易生成违背事实的内容,极易引入新的错误信息,反而可能成为误导检测的噪声源[14,63,67]。特别是当这些具备高语言连贯性的错误知识参与跨模态交互时,将严重干扰整体的语义表征,导致最终的分类决策边界发生偏移。如何在发挥LLM强大推理能力的同时,有效规避其生成风险,是当前亟待攻克的技术难题。

(3)动态社会感知与低资源环境建模的复杂性。谣言并非孤立存在的信息片段,其产生与演化始终与动态复杂的社会舆论环境紧密交织。传播结构、用户立场等社交上下文信息具有高度的演化性和稀疏性[77,87],特别是对于传播初期的谣言,由于社交数据极为有限,难以构建有效的图结构或立场模型进行建模与推理[94,99]。此外,实时、大规模信息环境的构建与更新需要巨大的计算与数据获取成本,且流行度、新颖度等环境信号的提取与量化难度较大,如何精确度量待检信息与动态环境在语义层面的细微异常,并避免被极端或嘈杂的舆论环境所干扰,是实现精准早期检测的主要障碍[106,109]。

(4)跨平台与跨语言的领域偏差。多模态谣言检测在实际应用中面临严重的泛化瓶颈。现有模型大多在特定平台或单一语种上进行训练,导致模型易拟合于特定平台的数据偏置或文化背景下的表达模式。一旦网络环境发生迁移,原本依赖特定领域知识的人工设计规则就会失效,特征空间分布的偏移会导致模型性能显著下降[127]。此外,尽管LLM展现出了强大的常识推理潜力,但其预训练语料的语言不平衡性,使其在处理中文等非英语谣言时,难以进行深层语义推理,受限于知识盲区与文化偏差[14,24]。由于缺乏能反映多平台动态演化的数据支撑,现有算法难以在异构网络中提取通用的领域不变特征,跨域泛化能力不足已成为制约模型落地应用的关键瓶颈。

(5)高逼真AIGC谣言的鉴伪挑战。现有的多模态谣言检测多依赖于挖掘图文之间的不一致性[39-40]。然而随着AIGC技术的发展,恶意传播者能够以极低的成本批量生成图文高度匹配且逻辑完全自洽的高逼真多模态谣言[132]。由于机器生成的图像与文本在语义空间上具备高度的对齐特性,基于传统跨模态特征对比的方法面临失效风险[44]。此外,当前的深度伪造技术已能生成连贯的视频序列与匹配音频,进一步加剧了传统虚假特征提取的难度[7,49]。如何揭示表层逻辑自洽下的底层伪造痕迹,并应对海量AIGC内容的自动化攻击,是当前检测系统面临的重要挑战。

4.2 未来研究方向

针对上述五大核心挑战,未来研究应聚焦于以下五个紧密对应的重点方向,以推动该领域向更实用、可靠的方向发展。

(1)深层语义理解与时序对齐机制。单纯依赖表层特征的拼接已难以满足复杂场景下的检测需求,研究必须向深层语义对齐探索。这一过程不仅需要构建跨模态的语义关联网络以增强推理的逻辑性[56],更需借助LLM的上下文解析能力,精准捕捉传统模型难以察觉的隐晦矛盾[63,67]。

在此基础上,未来的研究还需将视野从静态图文拓展至动态短视频,并重点发展细粒度的时序对齐机制,使模型能够同时捕捉视觉帧序列与音频轨道之间的同步关系,识别音画失配、时序篡改等伪造痕迹[133]。跨模态因果推理框架也是未来值得关注的方向[134],通过因果干预与逻辑溯源,可帮助模型在复杂的背景音效和视觉特效干扰中,剥离噪声并直击篡改本质,实现对高维时序特征的鲁棒融合,提升对高伪装短视频谣言的鉴别能力[135]。

(2)LLM赋能的统一推理框架。针对外部知识依赖与LLM幻觉的挑战,未来研究的关键在于重新定义LLM的角色,从被动的知识检索器升级为可控、可验证的推理中枢。这一方向不再局限于简单的信息查询,而是致力于构建透明的逻辑架构。Yan等人[136]通过引入链式验证机制,利用生成-核查的迭代

循环确保了多模态证据的逻辑闭环;Zhao等人[137]则另辟蹊径,利用动态动作空间优化策略,在维持高精度的同时有效解决了LLM推理的效率瓶颈;而Zhoubian等人[138]提出的强化学习框架,通过自训练策略与解码机制的深度耦合,增强了模型在复杂逻辑推演任务中的表现。通过引入这些受控的验证与自纠错机制,检测框架能够在推理早期主动拦截并过滤掉LLM产生的幻觉噪声,有效防止伪证对跨模态特征融合的污染。这些研究共同推动了多模态谣言检测方法从依赖静态特征向动态、可控的推理过程演进,为实现高鲁棒性的多模态谣言交叉验证提供了新的方法论支撑。

(3)低资源与动态环境下的高效学习方法。面对传播初期社交数据稀疏与环境建模成本高的双重约束,传统的大数据驱动模式已不足够,轻量化、高效率的学习框架成为研究关键。Ye等人[139]将元学习机制引入多模态领域,通过快速多步更新策略捕捉早期谣言的瞬态特征,提升了模型在早期谣言检测中的适应性与稳定性。Choi等人[140]则聚焦于特征空间的拓扑优化,利用跨模态三重Transformer与度量学习重塑样本分布,在极少样本条件下依然保持了高判别力。这种轻量化机制能以低计算代价破解数据稀疏困局,并过滤极端环境噪声。未来研究应进一步将这种快速适应能力与样本高效策略深度整合,摆脱对大规模标注数据的路径依赖,实现数据稀缺场景下的高鲁棒检测。

(4)跨域泛化机制与多语种综合数据集构建。针对模型在跨平台与跨语言场景下的泛化局限,未来研究应避免对单一数据源的过度拟合,转向探索跨域不变特征的学习机制。可引入多智能体协同等方法,实现模型对目标数据域特征的动态自适应[65],并结合历史语境感知等策略[106],消除特定平台的浅层数据偏置,提取更具泛化性的本质特征。同时,亟需克服现有数据集的局限性,构建涵盖多语种、跨平台传播路径的综合数据集。重点追踪同一谣言在不同语言文化背景和异构社交网络间的演化轨迹,以刻画谣言的跨域传播演化规律,为LLM的跨语种微调和跨域多模态检测提供数据支撑。

(5)AIGC防伪与多模态谣言检测的交叉融合。针对AIGC生成的逻辑自洽型谣言,未来的多模态谣言检测框架应突破单一的语义一致性假设,实现语义理解与底层伪造特征提取的交叉融合[141]。具体来说,不仅需要将传统的多媒体鉴伪技术与谣言检测深度结合,引入频域异常分析、像素级噪声规律提取以及生成器指纹识别等底层鉴伪手段[51]。还需要引入动态对抗训练机制,让检测模型能够随着AIGC生成工具的升级而不断迭代[70]。此外,可依托MLLM挖掘视觉内容中违背物理规律的细微异常,比如光影失真、局部结构扭曲等[50]。通过构建覆盖物理信号层、跨模态语义层与外部事实层的多维交叉验证体系,弥补单纯依赖高层语义对齐的局限,从而提升模型应对高逼真AIGC谣言的鲁棒性

5 结束语

随着AIGC技术的爆发与社交媒体生态的演变,谣言已从早期的单一文本形态,呈现为具备高伪装性、多模态耦合及动态传播特征的复杂信息实体,这显著增加了谣言检测任务的难度。通过对现有研究的系统综述与分析,本文得出以下结论:首先,多模态谣言检测已超越传统的图文一致性校验,逐步演进为融合外部知识、社交传播拓扑与外部环境参照的协同推理模式。LLM的引入在提升推理能力的同时也带来了知识幻觉与证据可靠性方面的新问题。其次,现有各类方法各有明确的适用边界。内容交互方法对表层语义冲突较为有效,但在高逼真AIGC内容上泛化能力有限;知识增强与社交上下文方法能够提供深层验证依据,却分别受制于知识库滞后和传播初期数据稀疏;外部环境感知方法则面临环境构建成本较高和背景噪声干扰的问题。最后,当前研究正从孤立的特征提取转向异构证据的交叉验证,依靠单一视角应对所有复杂的网络生态已难以实现。面对AIGC技术引发的攻防升级与社交网络的跨域复杂化,未来的研究不必执着于寻找单一通用的万能模型,而应致力于探索深层时序对齐机制、LLM的可控推理框架以及低资源环境下的动态自适应策略。通过突破跨域泛化瓶颈并深度融合多媒体鉴伪技术,最终构建起一个兼具高精度、强鲁棒性与决策透明度的多维交叉验证体系与可信检测架构。

参考文献

[1] Vosoughi S, Roy D, Aral S. The spread of true and false news online[J]. Science, 2018, 359(6380): 1146-1151.

[2] 中国互联网络信息中心. 第57次«中国互联网络发展状况统计报告»[EB/OL]. (2026-02-05)[2026-03-25]. https://www.cnnic.net.cn/n4/2026/0304/c88-11549.html.
China internet network information center. The 57th Statistical Report on China's Internet Development[EB/OL]. (2026-02-05)[2026-03-25]. https://www.cnnic.net.cn/n4/2026/0304/c88-11549.html. (in Chinese).

[3] 闫丹丹. 盘点2025年度网络热点谣言, 新的一年可别再传了![EB/OL]. (2026-01-23)[2026-03-25]. https://www.piyao.org.cn/20260123/1ac78e02aad148398ddc4fa018b7abd

3/c.html.
Yan Dandan. A review of top online rumors in 2025: Don't spread them in the new year! [EB/OL]. (2026-01-23)[2026-03-25]. https://www.piyao.org.cn/20260123/1ac78e02aad148398ddc4fa018b7abd3/c.html. (in Chinese).

[4] Yang Fan, Liu Yang, Yu Xiaohui, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. New York: ACM, 2012: 13.

[5] Jin Zhiwei, Cao Juan, Zhang Yongdong, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608.

[6] Li Xiaojun, Xiao Xvhao, Li Jia, et al. A CNN-based misleading video detection model[J]. Scientific Reports, 2022, 12(1): 6092.

[7] Mcuba M, Singh A, Ikuesan R A, et al. The effect of deep learning methods on deepfake audio detection for digital investigation[J]. Procedia Computer Science, 2023, 219: 211-219.

[8] 黄学坚, 马廷淮, 荣欢, 等. 融合外部知识与证据的场景图注意力网络多模态谣言检测[J]. 计算机学报, 2025, 48(9): 2159-2180.
Huang Xuejian, Ma Tinghuai, Rong Huan, et al. Multimodal rumor detection with scene graph attention networks integrating external knowledge and evidence[J]. Chinese Journal of Computers, 2025, 48(9): 2159-2180. (in Chinese)

[9] Cao Juan, Guo Junbo, Li Xirong, et al. Automatic rumor detection on microblogs: A survey[PP/OL]. V1. arXiv (2018-07-10)[2026-01-30]. https://arxiv.org/abs/1807.03505.

[10] 高玉君, 梁刚, 蒋方婷, 等. 社会网络谣言检测综述[J]. 电子学报, 2020, 48(7): 1421-1435.
Gao Yujun, Liang Gang, Jiang Fangting, et al. Social network rumor detection: A survey[J]. Acta Electronica Sinica, 2020, 48(7): 1421-1435. (in Chinese)

[11] Lakzaei B, Haghir Chehreghani M, Bagheri A. Disinformation detection using graph neural networks: A survey[J]. Artificial Intelligence Review, 2024, 57(3): 52.

[12] 刘华玲, 陈尚辉, 曹世杰, 等. 基于多模态学习的虚假新闻检测研究[J]. 计算机科学与探索, 2023, 17(9): 2015-2029.
Liu Hualing, Chen Shanghui, Cao Shijie, et al. Survey of fake news detection with multi-model learning[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(9): 2015-2029. (in Chinese)

[13] Hu Linmei, Wei Siqi, Zhao Ziwang, et al. Deep learning for fake news detection: A comprehensive survey[J]. AI Open, 2022, 3: 133-155.

[14] Hu Beizhe, Sheng Qiang, Cao Juan, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(20): 22105-22113.

[15] Zhang Mingqing, Liu Qiang, Tao Xiang, et al. SINCon: Mitigate LLM-generated malicious message injection attack for rumor detection[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna: ACL, 2025: 12570-12581.

[16] Hu Ziang, Wei Lingwei, Li Kun, et al. LLM-driven external knowledge integration network for rumor detection[C]//Proceedings of the 20th International Conference on Advanced Intelligent Computing Technology and Applications. Heidelberg: Springer, 2024: 3-13.

[17] 翟月莹. 网络谣言的传播与治理[J]. 东岳论丛, 2023, 44(8): 150-156.
Zhai Yueying. Communication and governance of network rumors[J]. DongYue Tribune, 2023, 44(8): 150-156. (in Chinese)

[18] 张文祥, 杨林. 网络谣言的行政规制与协同治理: 兼论公共权力及其边界[J]. 新闻界, 2022(5): 71-80.
Zhang Wenxiang, Yang Lin. Administrative regulation and collaborative governance of online rumors: A study of public power and its boundaries[J]. Press Circles, 2022(5): 71-80. (in Chinese)

[19] DiFonzo N, Bordia P. Rumor, gossip and urban legends[J]. Diogenes, 2007, 54(1): 19-35.

[20] Zubiaga A, Aker A, Bontcheva K, et al. Detection and resolution of rumours in social media: A survey[J]. ACM Computing Surveys, 2018, 51(2): 32.

[21] Zubiaga A, Liakata M, Procter R, et al. Analysing how people orient to and spread rumours in social media by looking at conversational threads[J]. PLoS One, 2016, 11(3): e0150989.

[22] Cai Guoyong, Wu Hao, Lv Rui. Rumors detection in Chinese via crowd responses[C]//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). Piscataway: IEEE, 2014: 912-917.

[23] 刘雅辉, 靳小龙, 沈华伟, 等. 社交媒体中的谣言识别研究综述[J]. 计算机学报, 2018, 41(7): 1536-1558.
Liu Yahui, Jin Xiaolong, Shen Huawei, et al. A survey on

rumor identification over social media[J]. Chinese Journal of Computers, 2018, 41(7): 1536-1558. (in Chinese)

[24] Kumar A, Esposito C, Karras D A. Introduction to special issue on misinformation, fake news and rumor detection in low-resource languages[J]. Transactions on Asian and Low-Resource Language Information Processing, 2022, 21(1): 1.

[25] Wu Liang, Morstatter F, Carley K M, et al. Misinformation in social media: Definition, manipulation, and detection[J]. SIGKDD Explorations, 2019, 21(2): 80-90.

[26] Grinberg N, Joseph K, Friedland L, et al. Fake news on Twitter during the 2016 U.S. presidential election[J]. Science, 2019, 363(6425): 374-378.

[27] Allcott H, Gentzkow M. Social media and fake news in the 2016 election[J]. Journal of Economic Perspectives, 2017, 31(2): 211-236.

[28] Singhal S, Shah R R, Chakraborty T, et al. SpotFake: A multi-modal framework for fake news detection[C]//Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). Piscataway: IEEE, 2019: 39-47.

[29] Jin Zhiwei, Cao Juan, Guo Han, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 795-816.

[30] Wang Yaqing, Ma Fenglong, Jin Zhiwei, et al. EANN: Event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 849-857.

[31] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal variational autoencoder for fake news detection[C]//Proceedings of the World Wide Web Conference. New York: ACM, 2019: 2915-2921.

[32] Hu Linmei, Zhao Ziwang, Qi Weijian, et al. Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection[J]. Information Sciences, 2024, 664: 120310.

[33] Song Chenguang, Ning Nianwen, Zhang Yunlei, et al. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks[J]. Information Processing & Management, 2021, 58(1): 102437.

[34] Qi Peng, Cao Juan, Li Xirong, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1212-1220.

[35] Jing Jing, Wu Hongchen, Sun Jie, et al. Multimodal fake news detection via progressive fusion networks[J]. Information Processing & Management, 2023, 60(1): 103120.

[36] Wang Zihao, Zhao Tongzhou, Wang Jiageng, et al. Fake news detection based on attention mechanism[C]//Proceedings of the 2024 6th International Conference on Robotics and Computer Vision (ICRCV). Piscataway: IEEE, 2024: 359-363.

[37] Singhal S, Kabra A, Sharma M, et al. SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract) [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(10): 13915-13916.

[38] Wang Shuai, Ratnavelu K, Bin Shibghatullah A S. UEFN: Efficient uncertainty estimation fusion network for reliable multimodal sentiment analysis[J]. Applied Intelligence, 2025, 55(3): 171.

[39] Zhou Xinyi, Wu Jindi, Zafarani R. SAFE: Similarity-aware multi-modal fake news detection[C]//Proceedings of the 24th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD). Heidelberg: Springer, 2020: 354-367.

[40] Xue Junxiao, Wang Yabo, Tian Yichen, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5): 102610.

[41] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning (ICML). [s.l.]: PMLR, 2021: 8748-8763.

[42] Zhou Yangming, Yang Yuzhou, Ying Qichao, et al. Multimodal fake news detection via CLIP-guided learning[C]// Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2023: 2825-2830.

[43] Ma Bin, Zhang Yifei, Xian Yongjin, et al. A cross-modal rumor detection scheme via contrastive learning by exploring text and image internal correlations[PP/OL]. V1. arxiv (2025-08-15)[2026-01-31]. https://arxiv.org/abs/2508.11141.

[44] Huang Xuejian, Ma Tinghuai, Rong Huan, et al. Dual evidence enhancement and text-image similarity awareness for multimodal rumor detection[J]. Engineering Applications of Artificial Intelligence, 2025, 153: 110845.

[45] Qian Shengsheng, Wang Jinguang, Hu Jun, et al. Hierarchical multi-modal contextual attention network for fake news detection[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 153-162.

[46] Guo Zhiwei, Yang Zhenguo, Liu Dahuang. Rumor detection based on cross-modal information-enhanced fusion network[C]//2024 16th International Conference on Advanced Computational Intelligence (ICACI). Piscataway: IEEE, 2024: 158-164.

[47] Han Mingxing, Li Jiaxuan, Chen Yu, et al. EC-Fake: A fake news detection model based on external knowledge and contrast-driven feature augmentation[J]. Neurocomputing, 2025, 653: 131214.

[48] Xing Ying, Zhai Changhe, Che Zhanbin, et al. A multimodal fake news detection model based on bidirectional semantic enhancement and adversarial network under Web3.0[J]. Electronics, 2025, 14(18): 3652.

[49] Wu Kaixuan, Lin Yanghao, Cao Donglin, et al. Interpretable short video rumor detection based on modality tampering[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. Torino: ELRA, ICCL, 2024: 9180-9189.

[50] Wang Junxi, Wang Yaxiong, Cheng Lechao, et al. FakeSV-VLM: Taming VLM for detecting fake short-video news via progressive mixture-of-experts adapter[C]// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025. Stroudsburg: ACL, 2025: 4782-4798.

[51] Pang Yucai, Yang Zhou, Li Qian, et al. Topic videolization: A rumor detection method inspired by video forgery detection technology[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(6): 3753-3765.

[52] Dun Yaqian, Tu Kefei, Chen Chen, et al. KAN: Knowledge-aware attention network for fake news detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 81-89.

[53] Zhang Huaiwen, Fang Quan, Qian Shengsheng, et al. Multi-modal knowledge-aware event memory network for social media rumor detection[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1942-1951.

[54] Wang Youze, Qian Shengsheng, Hu Jun, et al. Fake news detection via knowledge-driven multimodal graph convolutional networks[C]//Proceedings of the 2020 International Conference on Multimedia Retrieval. New York: ACM, 2020: 540-547.

[55] Sun Mengzhu, Zhang Xi, Zheng Jiaqi, et al. DDGCN: Dual dynamic graph convolutional networks for rumor detection on social media[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(4): 4611-4619.

[56] Cui Wei, Shang Mingsheng. KAGN: Knowledge-powered attention and graph convolutional networks for social media rumor detection[J]. Journal of Big Data, 2023, 10(1): 45.

[57] Hu Linmei, Yang Tianchi, Zhang Luhao, et al. Compare to the knowledge: Graph neural fake news detection with external knowledge[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: ACL, 2021: 754-763.

[58] Sun Mengzhu, Zhang Xi, Ma Jianqiang, et al. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12736-12749.

[59] Hangloo S, Arora B. Combating multimodal fake news on social media: Methods, datasets, and future perspective[J]. Multimedia Systems, 2022, 28(6): 2391-2422.

[60] Li Guanghua, Lu Wensheng, Zhang Wei, et al. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors[PP/OL]. V1. arxiv (2024-03-14)[2026-01-31]. https://arxiv.org/abs/2403.09747.

[61] He Zimeng, Yang Yunhao, Yu Wei, et al. A multi-stage rumor detection framework based on retrieval-augmented generation optimization[C]//Proceedings of the 27th International Conference on Asian Digital Libraries on Intelligence and Equity: Shaping the Future of Knowledge. Heidelberg: Springer, 2026: 22-38.

[62] Xiao Liang, Shi Chongyang, Hao Shufeng, et al. Fact-augmented reasoning model for fake news detection[C]// Proceedings of the 32nd International Conference on Neural Information Processing (ICONIP). Heidelberg: Springer, 2026: 18-32.

[63] Qi Peng, Yan Zehong, Hsu W, et al. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 13052-13062.

[64] Liu Yuhan, Liu Yuxuan, Zhang Xiaoqing, et al. The truth becomes clearer through debate! Multi-agent systems with large language models unmask fake news[C]//Pro-

ceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2025: 504-514.

[65] Li Hui, Wang Ante, Li Kunquan, et al. A multi-agent framework with automated decision rule optimization for cross-domain misinformation detection[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2025: 5709-5725.

[66] Bukke R, Pandey S, Kumar S, et al. Agentic multi-persona framework for evidence-aware fake news detection[PP/OL]. V2. arxiv (2026-03-05)[2026-03-18]. https://arxiv.org/abs/2512.21039.

[67] 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[J]. 计算机研究与发展, 2024, 61(5): 1250-1260.
Ke Jing, Xie Zheyong, Xu Tong, et al. An implicit semantic enhanced fine-grained fake news detection method based on large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1250-1260. (in Chinese)

[68] Xu Yingrui, Ge Jingguo, Guangxu Lyu, et al. Multimodal fake news detection based on chain-of-thought prompting large language models[C]//Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway: IEEE, 2024: 559-566.

[69] Yang Yuzhou, Zhou Yangming, Zhu Zhiying, et al. RoE-FND: A case-based reasoning approach with dual verification for fake news detection via LLMs[PP/OL]. V1. arXiv (2025-06-04) [2026-03-18]. https://arxiv.org/abs/2506.11078.

[70] Wang Yifeng, Gu Zhouhong, Zhang Siwei, et al. LLM-GAN: Constructing generative adversarial network through large language models for explainable fake news detection[C]//Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2025: 1-5.

[71] Meel P, Vishwakarma D K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities[J]. Expert Systems with Applications, 2020, 153: 112986.

[72] Ma Jing, Gao Wei, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2017: 708-717.

[73] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 675-684.

[74] Bai Na, Meng Fanrong, Rui Xiaobin, et al. Rumour detection based on graph convolutional neural net[J]. IEEE Access, 2021, 9: 21686-21693.

[75] Bondielli A, Marcelloni F. A survey on fake news and rumour detection techniques[J]. Information Sciences, 2019, 497: 38-55.

[76] Zhou Xinyi, Zafarani R. Network-based fake news detection: A pattern-driven approach[J]. ACM SIGKDD Explorations Newsletter, 2019, 21(2): 48-60.

[77] Bian Tian, Xiao Xi, Xu Tingyang, et al. Rumor detection on social media with bi-directional graph convolutional networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 549-556.

[78] 张鑫昕, 潘善亮, 茅琴娇. 基于传播树的多特征谣言检测方法[J]. 电子学报, 2024, 52(5): 1609-1618.
Zhang Xinxin, Pan Shanliang, Mao Qinjiao. A rumor detection approach based on multi-feature propagation tree[J]. Acta Electronica Sinica, 2024, 52(5): 1609-1618. (in Chinese)

[79] Zheng Peng, Huang Zhen, Dou Yong, et al. Rumor detection on social media through mining the social circles with high homogeneity[J]. Information Sciences, 2023, 642: 119083.

[80] Sun Ling, Rao Yuan, Lan Yuqian, et al. HG-SL: Jointly learning of global and local user spreading behavior for fake news early detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(4): 5248-5256.

[81] Jin Zhiwei, Cao Juan, Jiang Yugang, et al. News credibility evaluation on microblog with a hierarchical propagation model[C]//Proceedings of the 2014 IEEE International Conference on Data Mining. Piscataway: IEEE, 2014: 230-239.

[82] 欧阳祺, 陈鸿昶, 刘树新, 等. 基于Bert-GNNs异质图注意力网络的早期谣言检测[J]. 电子学报, 2024, 52(1): 311-323.
Ouyang Qi, Chen Hongchang, Liu Shuxin, et al. Early rumor detection based on Bert-GNNs heterogeneous graph attention network[J]. Acta Electronica Sinica, 2024, 52(1): 311-323. (in Chinese)

[83] Zhang Jiawei, Dong Bowen, Yu P S. FakeDetector: Effective fake news detection with deep diffusive neural network[C]//Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2020: 1826-1829.

[84] Yang Xiaoyu, Yuefei Lyu, Tian Tian, et al. Rumor detec-

tion on social media with graph structured adversarial learning[C]//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama: ijcai.org, 2021: 197.

[85] Rahimi M, Roayaei M. A multi-view rumor detection framework using dynamic propagation structure, interaction network, and content[J]. IEEE Transactions on Signal and Information Processing over Networks, 2024, 10: 48-58.

[86] Li Yuanqing, Dai Mengyao, Zhang Sanfeng. FHGraph: A novel framework for fake news detection using graph contrastive learning and LLM[J]. Computers, Materials & Continua, 2025, 83(1): 309-333.

[87] Yuan Xue, Shu Anhao, Wu Yue, et al. MSSCR: Multiscale semantic collaborative reasoning model for explainable multimodal rumor detection[J]. Neurocomputing, 2025, 653: 131042.

[88] Wu Ke, Yang Song, Zhu K Q. False rumors detection on sina weibo by propagation structures[C]//Proceedings of the 2015 IEEE 31st International Conference on Data Engineering. Piscataway: IEEE, 2015: 651-662.

[89] Liu Xiaomo, Nourbakhsh A, Li Quanzhi, et al. Real-time rumor debunking on Twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Manageme. New York: ACM, 2015: 1867-1870.

[90] Hu Xia, Tang Jiliang, Gao Huiji, et al. Social spammer detection with sentiment information[C]//Proceedings of the 2014 IEEE International Conference on Data Mining. Piscataway: IEEE, 2014: 180-189.

[91] Liu Yang, Wu Y F B. FNED: A deep network for fake news early detection on social media[J]. ACM Transactions on Information Systems, 2020, 38(3): 25.

[92] Shu Kai, Cui Limeng, Wang Suhang, et al. dEFEND: Explainable fake news detection[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 395-405.

[93] Xie Jianhui, Liu Song, Liu Ruixin, et al. SERN: Stance extraction and reasoning network for fake news detection[C]// Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 2520-2524.

[94] Li Guoyi, Hu Die, Liu Zongzhen, et al. Semantic reshuffling with LLM and heterogeneous graph auto-encoder for enhanced rumor detection[C]//Proceedings of the 31st International Conference on Computational Linguistics. Stroudsburg: ACL, 2025: 8557-8572.

[95] Qiong Nan, Sheng Qiong, Cao Juan, et al. Let silence speak: Enhancing fake news detection with generated comments from large language models[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. New York: ACM, 2024: 1732-1742.

[96] Chen Tong, Li Xue, Yin Hongzhi, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]//Proceedings of the PAKDD 2018 Workshops on Trends and Applications in Knowledge Discovery and Data Mining. Heidelberg: Springer, 2018: 40-52.

[97] Liu Yang, Wu Yifang. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 354-361.

[98] Huang Xuejian, Ma Tinghuai, Jin Wenwen, et al. Multiview spatio-temporal learning with dual dynamic graph convolutional networks for rumor detection[J]. IEEE Transactions on Computational Social Systems, 2025, 12(5): 2469-2479.

[99] Xu Haowei, Gao Chao, Li Xianghua, et al. D2: Customizing two-stage graph neural networks for early rumor detection through cascade diffusion prediction[C]//Proceedings of the 18th ACM International Conference on Web Search and Data Mining. New York: ACM, 2025: 568-576.

[100] Zhong Nanjiang, Jiang Xinchen, Yao Yuan. From detection to explanation: Integrating temporal and spatial features for rumor detection and explaining results using LLMs[J]. Computers, Materials & Continua, 2025, 82(3): 4741-4757.

[101] Sheng Qiang, Cao Juan, Zhang Xueyao, et al. Zoom out and observe: News environment perception for fake news detection[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 4543-4556.

[102] Fang Xiaochang, Wu Hongchen, Jing Jing, et al. NSEP: Early fake news detection via news semantic environment perception[J]. Information Processing & Management, 2024, 61(2): 103594.

[103] Liu Xinyu, Ma Kun, Ji Ke, et al. Graph-based multi-information integration network with external news environment perception for propaganda detection[J]. International Journal of Web Information Systems, 2024, 20(2): 195-212.

[104] Wang Kun, Yang Yuzhen, Wang Xiaoyang. Spectral

clustering-guided news environments perception for fake news detection[J]. IEEE Access, 2024, 12: 197529-197539.

[105] Yu Wencheng, Ge Jike, Yang Zhaoxu, et al. Multi-domain fake news detection for history news environment perception[C]//Proceedings of the IEEE 17th Conference on Industrial Electronics and Applications. Piscataway: IEEE, 2022: 428-433.

[106] Liu Guojun, Ma Yuefeng, Liang Xun. Cross-domain fake news detection based on coarse-fine grained environments reflecting public expectation[C]//Proceedings of the 2023 IEEE International Conference on Big Data. Piscataway: IEEE, 2023: 403-412.

[107] Xiao Tianshu, Guo Sichang, Huang Jingcheng, et al. HiPo: Detecting fake news via historical and multi-modal analyses of social media posts[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York: ACM, 2023: 2805-2815.

[108] Yu Wencheng, Ge Jike, Chen Zuqin, et al. Research on fake news detection based on dual evidence perception[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108271.

[109] Wang Bing, Li Ximing, Wang Yiming, et al. Variety is the spice of life: Detecting misinformation with dynamic environmental representations[C]//Proceedings of the 34th ACM International Conference on Information and Knowledge Management. New York: ACM, 2025: 2945-2955.

[110] Li Li, Hu Jingyang, Fu Shihao, et al. Rumor detection based on macro-micro public opinion modeling and sentiment scores[C]//Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE, 2024: 3950-3955.

[111] Lu Jiahui, Zhang Huibin, Xiao Yi, et al. An environmental uncertainty perception framework for misinformation detection and spread prediction in the COVID-19 pandemic: Artificial intelligence approach[J]. JMIR AI, 2024, 3: e47240.

[112] Wu Yang, Zhang Yanqiang, Xu Zhen, et al. Rumor detection with news environment enhanced propagation structure[C]//Proceedings of the 20th International Conference on Advanced Intelligent Computing Technology and Applications. Heidelberg: Springer, 2024: 187-198.

[113] De Oliveira N R, Pisa P S, Lopez M A, et al. Identifying fake news on social networks based on natural language processing: Trends and challenges[J]. Information, 2021, 12(1): 38.

[114] Qi Peng, Bu Yuyan, Cao Juan, et al. FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(12): 14444-14452.

[115] Bu Yuyan, Sheng Qiang, Cao Juan, et al. FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 1351-1360.

[116] Ma Jing, Gao Wei, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: AAAI Press, 2016: 3818-3824.

[117] Tan Li, Wang Ge, Jia Feiyang, et al. Research status of deep learning methods for rumor detection[J]. Multimedia Tools and Applications, 2023, 82(2): 2941-2982.

[118] Hu Xuming, Guo Zhijiang, Chen Junzhe, et al. MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2023: 2901-2912.

[119] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media[C]//Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2013: 1103-1108.

[120] Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at mediaeval 2015[M]. Aachen: CEUR Workshop Proceedings, 2015.

[121] Derczynski L, Bontcheva K, Liakata M, et al. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours[C]//Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg: ACL, 2017: 69-76.

[122] Shu Kai, Mahudeswaran D, Wang Suhang, et al. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171-188.

[123] AL-Sarem M, Alsaeedi A, Saeed F, et al. A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs[J]. Applied Sciences, 2021, 11(17): 7940.

[124] Cheng Mingxi, Wang Songli, Yan Xiaofeng, et al. A COVID-19 rumor dataset[J]. Frontiers in Psychology, 2021, 12: 644801.

[125] Song Changhe, Yang Cheng, Chen Huimin, et al. CED:

Credible early detection of social media rumors[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(8): 3035-3047.

[126] Wang Xiaoda, Luo Chenxiang, Guo Tengda, et al. BGEK: External knowledge-enhanced graph convolutional networks for rumor detection in online social networks[C]//Proceedings of the 32nd International Conference on Artificial Neural Networks on Artificial Neural Networks and Machine Learning. Heidelberg: Springer, 2023: 291-303.

[127] Chen Xuelong, Pan Jinchao. A cross-platform rumor detection framework considering data privacy protection and different detection capabilities of online social platforms[J]. Decision Support Systems, 2025, 198: 114524.

[128] Luo G, Darrell T, Rohrbach A. NewsCLIPpings: Automatic generation of out-of-context multimodal media[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 6801-6817.

[129] Galitsky B. Detecting rumor and disinformation by web mining[C]//Proceedings of the 2015 AAAI Spring Symposium. Palo Alto: AAAI Press, 2015.

[130] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3156-3164.

[131] Wang Hongwei, Zhang Fuzheng, Xie Xing, et al. DKN: Deep knowledge-aware network for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference. New York: ACM, 2018: 1835-1844.

[132] Yu Xiaomin, Wang Yezhaohui, Chen Yanfang, et al. Fake artificial intelligence generated contents (FAIGC): A survey of theories, detection methods, and opportunities[PP/OL]. V2. arXiv (2024-05-03)[2026-03-20]. https://arxiv.org/abs/2405.00711.

[133] Zhu Xiaodong, Wang Suting, Yang Junqi, et al. Query-based audio-visual temporal forgery localization with register-enhanced representation learning[C]//Proceedings of the 33rd ACM International Conference on Multimedia. New York: ACM, 2025: 8547-8556.

[134] Liu Moyang, Yan Kaiying, Liu Yukun, et al. Deconfounded reasoning for multimodal fake news detection via causal intervention[PP/OL]. V2. arXiv (2025-04-12) [2026-03-20]. https://arxiv.org/abs/2504.09163.

[135] Chen Lizhi, Qian Zhong, Li Peifeng, et al. Disconfounding fake news video explanation with causal inference[C]//Proceedings of the 34th International Joint Conference on Artificial Intelligence. California: ijcai.org, 2025: 539.

[136] Yan Kaiying, Liu Moyang, Liu Yukun, et al. Debunk and infer: Multimodal fake news detection via diffusion-generated evidence and LLM reasoning[PP/OL]. V1. arXiv (2025-06-11) [2026-01-31]. https://arxiv. org/abs/2506.21557.

[137] Zhao Xueliang, Wu Wei, Guan Jian, et al. DynaAct: Large language model reasoning with dynamic action spaces[C/OL]//Proceedings of the 39th Annual Conference on Neural Information Processing Systems, 2025: 1-29. https://openreview.net/forum id=R24ZqNwoDz.

[138] Zhoubian S, Zhang Dan, Tang Jie. ReST-RL: Achieving accurate code reasoning of LLMs with optimized self-training and decoding[PP/OL]. V2. arXiv (2025-09-08)[2026-01-31]. https://arxiv.org/abs/2508.19576.

[139] Ye Yanyan, Chen Hongzhe, Cai Qianqian, et al. Multimodal meta-learning for early rumor detection based on few-shot learning[J]. IEEE Transactions on Computational Social Systems, 2026, 13(1): 475-484.

[140] Choi E, Ahn J, Xinyu Piao, et al. CroMe: Multimodal fake news detection using cross-modal tri-transformer and metric learning[J]. IEEE Access, 2025, 13: 197124-197132 .

[141] Yang Chang, Zhang Peng, Zhang Jing, et al. Rethink rumor detection in the era of LLMs: A review[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. Stroudsburg: ACL, 2025: 8730-8749.

作者简介

李雨泽 女，1998年6月出生于黑龙江省双鸭山市。现为西安邮电大学通信与信息工程学院博士研究生。主要研究方向为多模态信息处理。
E-mail: selene_zeze@stu.xupt.edu.cn

刘 颖 女，1972年1月出生于陕西省西安市。现为青海理工学院教授/西安邮电大学特聘教授、博士生导师。主要研究方向为图像理解与模式识别、多模态信息处理。中国电子学会会员编号：E190091672M。
E-mail: liuying_ciip@163.com