计算机科学
COMPUTER SCIENCE

社交媒体虚假信息检测研究综述

陈静, 周刚, 李顺航, 郑嘉丽, 卢记仓, 郝耀辉

引用本文

陈静, 周刚, 李顺航, 郑嘉丽, 卢记仓, 郝耀辉. 社交媒体虚假信息检测研究综述[J]. 计算机科学, 2024, 51(11): 1-14.

CHEN Jing, ZHOU Gang, LI Shunhang, ZHENG Jiali, LU Jicang, HAO Yaohui. Review of Fake News Detection on Social Media [J]. Computer Science, 2024, 51(11): 1-14.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700101

社交媒体虚假信息检测研究综述

陈 静 周 刚 李顺航 郑嘉丽 卢记仓 郝耀辉
信息工程大学数据与目标工程学院 郑州 450001

摘 要 社交媒体上的虚假信息不仅危害网络空间安全,还在重大事件中扮演着关键角色,严重误导公众,对政治和社会秩序造成负面影响。为此,围绕面向社交媒体的虚假信息检测技术研究展开综述,为构建高效检测技术和遏制社交媒体虚假信息泛滥奠定理论基础。首先,深入剖析虚假信息的内涵本质,探讨其在社交平台上的产生机理、具体表现形式,并界定检测任务的基础框架与目标;其次,从语义一致性视角出发,专注内容语义、社交上下文感知和知识驱动三大层面,对比梳理典型检测方法;在此基础之上,深入探究增强检测算法可解释性最新研究成果;进一步,从对抗博弈视角,深度剖析当前社交媒体虚假信息检测任务面临的挑战以及大型语言模型为虚假信息检测技术研究带来的机遇;最后,对社交媒体虚假信息检测技术未来的发展进行了展望。

关键词:虚假信息检测;跨模态关联;社交上下文感知;知识驱动;大语言模型
中图分类号 TP391

Review of Fake News Detection on Social Media

CHEN Jing,ZHOU Gang,LI Shunhang,ZHENG Jiali,LU Jicang and HAO Yaohui
School of Data and Target Engineering,Information Engineering University,Zhengzhou 450001,China

Abstract Fake news on social media not only jeopardizes cyberspace security,but also plays a pivotal role in major events,severely misleads the public and has a negative affect on political and social order. Therefore,this paper outlines social media fake news detection techniques,establishing a theoretical foundation for building efficient detection technology and curbing the proliferation of fake news on social media. Firstly,it deeply analyzes the connotation and essence of fake news,explores its generation mechanism and specific manifestations on social platforms,and defines the basic framework and objectives of the detection task. Next,from the perspective of semantic consistency,it focuses on three major levels:content semantics,social context awareness,and knowledge-driven,and compares and combs typical detection methods. On this basis,it deeply explores the latest research advancements in enhancing the explainability of detection algorithms. Furthermore,from the adversarial perspective,it deeply analyzes the challenges faced by current social media fake news detection tasks and the opportunities brought to research detection technology by large-scale language models. Finally,the future development of social media fake news detection technology is prospected.

Keywords Fake news detection,Cross-modal correlation,Social context awareness,Knowledge-driven,Large language model

1 引言

社交媒体网络空间的庞大用户量、高渗透率和快速传播性,使其成为公众获取信息和发表言论的重要途径。在信息传播的赋能、渲染、靶向和流变机理作用下,社交媒体网络空间中渗透与反渗透、攻击与反攻击、控制与反控制日益激烈。根据最新统计,2023 年微博共有效处理不实信息 8 万多条,抖音平台封禁发布不实信息账号约 1.9 万多个。中国社会科学院 2023 年发布的《中国新媒体发展报告 No.14》[1]指出,随着文本自动生成等新兴人工智能技术的问世,由人工智能生成的虚假信息呈爆发式增长,内容涵盖政治、社会等多个领域。

社交媒体上的虚假信息不仅威胁到网络空间安全,在重大事件中尤其扮演着重要角色。不实信息会对民众认知产生严重的干扰,致使群众做出错误的决策,从而对现实世界的政治和社会秩序造成严重的负面影响。例如,2020 年全世界范围内的 COVID-19 大流行期间出现的“信息瘟疫”,许多带有误导性内容的新闻报道通过社交媒体传播,导致社会经济

到稿日期:2024-07-16 返修日期:2024-09-03
基金项目:国家自然科学基金面上项目(62172433);河南省科技攻关项目(222102210081);河南省软科学研究项目(202400410084)
This work was supported by the National Natural Science Foundation of China(62172433),Henan Province Science and Technology Research Project(222102210081) and Henan Province Soft Science Research Project(202400410084).
通信作者:陈静(cathysilense@126.com)

紊乱，全球流行病预防效果减弱；2022 年，“俄乌冲突”爆发期间虚假信息在网络空间肆意传播，导致仇恨情绪无限蔓延，极大地干扰了公众对冲突态势的认知与判断。虚假信息已经成为引发全球关注的社会和科学问题。因此，综合应用自然语言处理、社交挖掘、跨模态分析等智能处理手段，挖掘并利用信息的内在特征、产生机理与传播规律，构建虚假信息智能检测模型，对于遏制目前社交媒体虚假信息持续泛滥的态势至关重要。

目前，针对社交媒体的虚假信息检测技术已成为研究热点，吸引了众多学者的广泛关注，并促成了大量创新性的技术探索与理论研究。然而，这些努力却面临着智能技术进步带来的新挑战。为了全面深入洞察社交媒体虚假信息检测领域的学术进展，众多学者开展了相关的综合评述工作，系统梳理该领域的理论发展脉络、关键技术突破、活跃的研究焦点、新兴的前沿议题，以及亟待克服的挑战。例如，Guo 等[2]从基于内容和结合社交上下文等多个方面系统性地概述了假信息检测的核心技术手段，并进一步探讨了假信息的早期检测、可解释性以及多模态数据等关键挑战。Zhang 等[3]综合分析了虚假信息最新的研究动态，尤其针对假信息的核心特征与检测技术方法进行了深入的探讨与综合评述。Bhattcharjee 等[4]则围绕网络信息生态系统中假信息的检测、缓解以及挑战 3 个方面展开论述。Alam 等[5]从多模态虚假信息检测角度展开论述，涵盖了各种模态组合：文本、图像、语音、视频以及社交媒体网络结构和时间信息。Hardalov 等[6]认为立场检测在对抗虚假信息中起到重要作用，因此重点从两个角度对立场检测与假信息检测之间的关系展开论述：将立场检测作为假信息检测的一个组成部分，以及将立场检测作为单独的任务与假信息检测任务联合训练。这些综合性研究从多个角度为构建假信息检测理论框架奠定了基础。然而，它们在面对智能生成技术产生的新型假信息，以及大型语言模型[1)]（Large-Scale Pre-Trained Language Models，LLMs）在假信息检测任务中既有的挑战与潜在的机遇方面，尚留有探讨的空间。

为此，本文以面向社交媒体的虚假信息检测任务为切入点，第 2 章深入剖析虚假信息的内涵本质，探讨智能生成时代背景下其在社交平台上的产生机理、表现形态，并详细界定检测任务的基础框架与目标；第 3 章聚焦基于深度学习的检测方法，着重从内容语义、社交上下文感知以及知识驱动三大层面，系统梳理假信息检测的技术手段与方法；在此基础之上，第 4 章深入探究增强检测算法可解释性的最新研究进展；第 5 章从对抗博弈视角，深度剖析智能生成时代社交媒体虚假信息任务面临的挑战以及大型语言模型为虚假信息检测技术未来研究创造的机遇。

2 问题描述

2.1 面向社交媒体的虚假信息的定义

目前，针对虚假信息并没有统一明确的定义，表 1 列出了不同权威机构对虚假信息的描述。

表 1 不同机构对虚假信息的描述

Table 1 Descriptions of fake news by different institutions

来源	定义
柯林斯词典	以新闻报道为幌子传播的虚假的、往往耸人听闻的信息[2)]
康桥词典	在互联网或其他媒体上传播，通常是为了影响政治观点，看似新闻的虚假故事[3)]
维基百科	经常被用来指代捏造的新闻。这类新闻出现在传统新闻、社交媒体或假新闻网站上，没有事实依据；撰写和发布假新闻通常是为了误导机构、实体或个人，和/或在经济或政治上获利[4)]
信息科学	制造者故意误导读者，并能够通过一些其他来源证实其结果为假的信息[7]

基于上述关于虚假信息的描述，可以提炼出其两个核心属性：1）蓄意误导性；2）可验证错误性。因此，面向社交媒体的虚假信息可被定义为以刻意欺骗为目的，在社交媒体平台上生成并迅速扩散，最终能够通过其他可靠渠道得到验证并确认为假的信息。

2.2 社交媒体虚假信息的产生机理

智能生成技术的迅猛发展使得 AI 创造的内容与人类作品之间的界限日益模糊，几乎达到难以辨识的程度。但与此同时，这一飞跃性的进展也无意中为虚假信息的制造及传播提供了温床。不良企图者正利用这类高级人工智能技术，系统性地炮制出大量仿人类的虚假信息，旨在操控公众舆论，误导大众对各类事件的认知与立场倾向，以服务于其特定的战略目标。此类假信息被学界与业界形象地称为“神经假信息”。

神经假信息的一个显著特征在于其内容往往具有明确的“针对性”，即信息生成过程中会依据特定条件进行定制化伪造。这一现象被学术界定义为神经假信息的“条件生成”特性[8]。根据生成器的条件与方法的不同，可以将虚假信息生成方法划分为 3 类：以知识元素为条件，以事实背景为条件，以及以风格和潜在意图为条件。

2.2.1 以知识元素为条件的虚假信息生成方法

这类方法认为假信息往往是基于操纵真实信息的知识元素（包括实体、关系和事件）而产生的。为此，它们以从真实信息中提取的知识元素为条件，通过篡改、扭曲等手段，创造出看似合理但实际失实的内容。具体过程如图 1 所示。

在条件生成器的训练阶段，以从原始数据集中提取的知识元素为条件，学习生成贴合条件的文本内容的能力。在假生成阶段，通过向生成器注入经过修改（如实体交换、添加新关系或事件、子图替换等）的知识元素，生成偏离事实、常识等类型的虚假信息。该方法被应用到多模态假信息生成时，容易生成明显的跨模态不一致的假信息。为了防止生成明显不一致的信息，文献[9]在信息生成过程中施加了跨媒体知识图操作约束，能够为训练检测器生成更真实/更具挑战性的数据。

1) 为了表述方便，后续我们将大语言模型简称为大模型

2) https://www.collinsdictionary.com/dictionary/english/fake-news

3) https://dictionary.cambridge.org/dictionary/english/fake-news

4) https://en.wikipedia.org/wiki/Fake_news

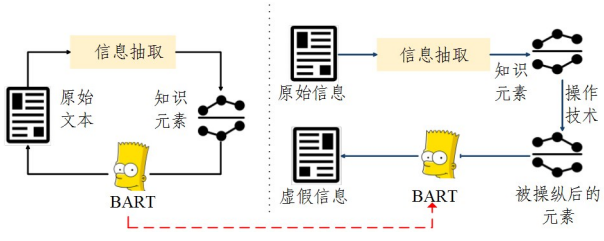


图 1　基于知识元素的假新闻生成器的训练与推理过程[9]

Fig. 1　Training and inference process of fake news generator based on knowledge elements[9]

2.2.2　以事实背景为条件的虚假信息生成方法

为了弥合人类创作真实信息与机器生成内容之间因本质属性而产生的事实性偏差，相关研究提出了以事实背景为条件的假信息生成方法。这种策略旨在通过深度学习和理解真实世界的背景知识，生成更为逼真且上下文连贯的内容，从而有效混淆真实与伪造虚假信息的边界。例如，文献[10]通过整合从权威数据源检索到的具体事实，增强了自动创建的虚假内容的真实感，具体如图 2 所示。

在训练阶段，为了确保生成内容既真实又准确，假设训练数据本身包含事实准确且语义清晰的陈述，通过从权威外部事实库中检索到的相关事实重塑这些关键的陈述，达到生成原始信息的目的。在生成器的推断阶段，给定一条虚假声明，模型能够检索相关的事实，帮助生成更逼近真实的假信息。这种方法巧妙地将外部权威事实融入虚假信息，显著增强了内容的可信度和权威性，同时使得虚假部分巧妙地潜藏于真实的事实背景中。这种混淆视听的方法极大地提升了识别不实内容的难度。

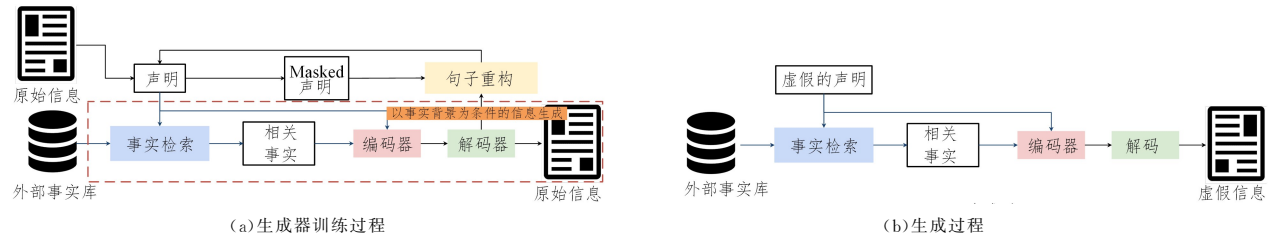


图 2　基于事实背景的假信息生成器训练以及推断过程[10]

Fig. 2　Training and inference process of fake news generator based on factual background[10]

2.2.3　以风格和潜在意图为条件的假信息生成方法

人工撰写的假信息一般存在两个潜在特征[11]：其一，大部分假信息仅包含 1～2 句错误内容，以有限的误导性陈述巧妙地嵌入大量看似真实的文本之中；其二，约有 1/3 的人工假信息运用了宣传技巧，旨在通过情感触发词汇或逻辑谬误来增强对读者的影响力，从而达到煽动情绪、左右观点的目的。基于此，相关学者提出了一种以风格和潜在意图为条件的假信息生成方法[12]。这类方法主要通过结合宣传技巧并保留大部分的正确信息来产生假信息。具体而言，将一个关键的句子替换成看似合理但虚假的信息。为了确保生成的句子是假的，可以使用自批评序列模型[13]进行训练，使模型能够自我校验生成文本的虚假程度，具体如图 3 所示。

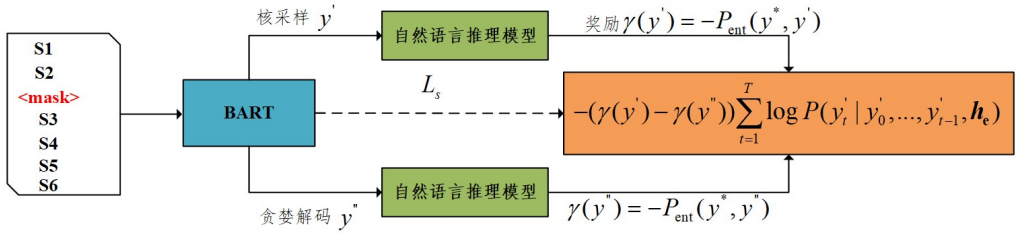


图 3　基于关键句子替换的假信息生成器的训练过程[12]

Fig. 3　Training process of fake news generator based on key sentence replacement[12]

在模仿人类制造假信息的过程中，着重强调了两种宣传技巧的应用：一是使用强烈的情感词汇来撩拨读者的情绪，即所谓的“负载语言”，通过激发读者的喜怒哀乐，促使他们对假信息产生共鸣；二是“诉诸权威”，即巧妙引用专家意见或权威机构的声明以支撑假信息中的论点，利用公众对专业权威的信任感来增强假信息的可信度。这两方面的策略共同作用，旨在构建出既具有内在欺骗性又能够有效触动读者心理防线的假信息。

2.3　社交媒体虚假信息的特点分析

相较于其他传播媒介中的虚假信息，社交媒体上的虚假信息突出以下 3 个特征。

1）短文本特性：社交媒体上的信息往往借助易于快速消费和传播的短文本格式，将信息进行高度浓缩，导致语义表达相对稀疏，增加了其核查与甄别的难度。图 4 展示了微博数据集上不同领域帖文的平均长度，各个领域的平均长度不超过 135。

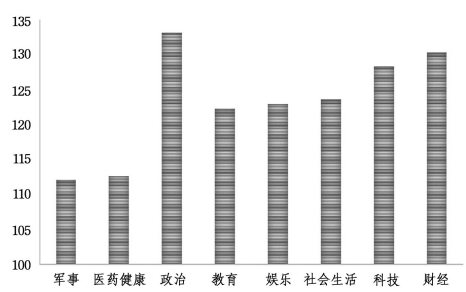


图 4　不同领域微博帖文的平均长度

Fig. 4　Average length of Weibo posts in different fields

2)多模态融合:图 5 给出了新浪平台辟谣的假信息中包含图片数量的占比,可以发现约有 60%左右的假信息中包含图像类数据。这反映出社交媒体信息普遍呈现出多模态数据融合的特点,即文本、图片等多种模态相互交织,使得虚假信息能通过更生动直观的方式包装和散布,增强了其欺骗性和感染力。

3)复杂语境依赖:图 6 给出了一条内容为假的推文发布后,在社交媒体传播的过程。可以观察出,社交媒体信息通常嵌入在用户之间的互动网络之中,包含丰富的上下文背景,这使得虚假信息可能借助情境关联增强其误导效果,同时也增加了甄别难度。

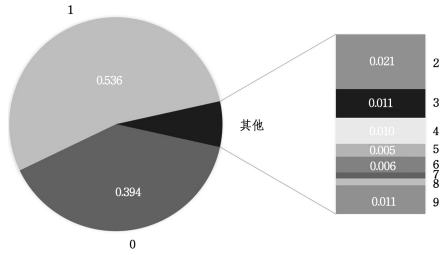


图 5 假信息中不同图片数量的占比

Fig. 5 Proportion of different numbers of images in fake news

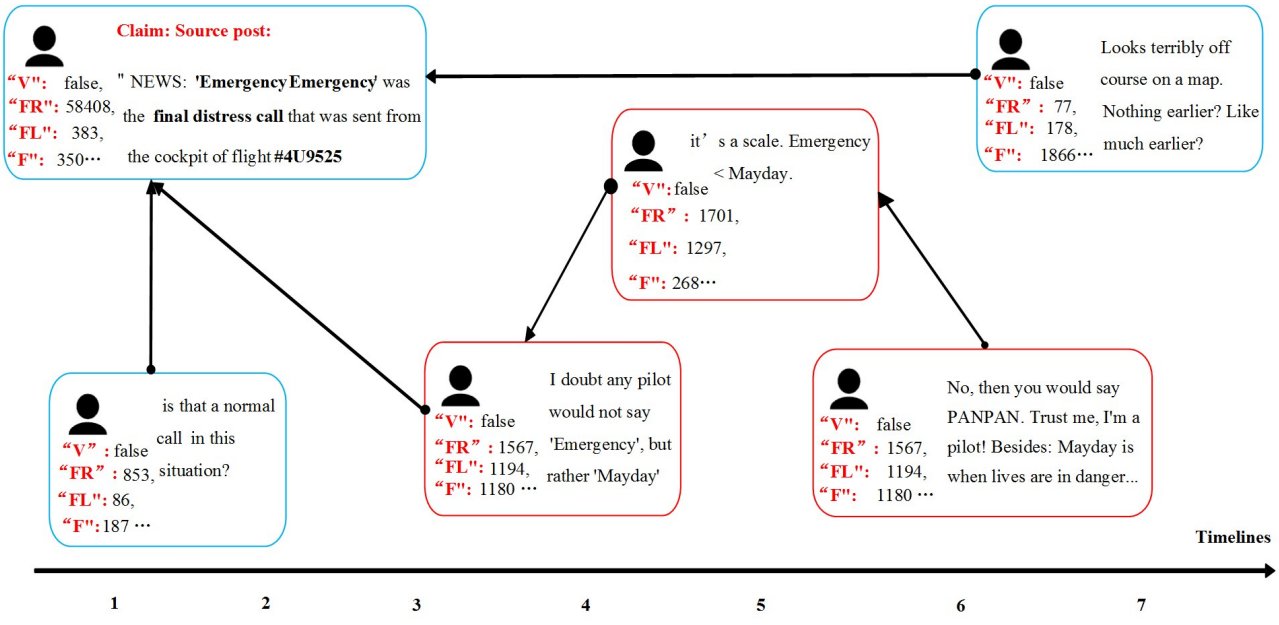


图 6 假信息传播过程示意图[1)]

Fig. 6 Schematic diagram of the process of fake news propagation[1)]

2.4 虚假信息检测任务描述

虚假信息检测属于内容可信度检测研究范围,旨在在信息传播的早期阶段有效地识别出虚假信息。该任务本质上可以被看作是一个分类问题,即在给定输入的情况下,输出真或假标签。具体形式化描述如下:

$$f: f(p)=\hat{y} \tag{1}$$

其中,p 表示待检测帖子信息;$\hat{y}\in\{0,1\}$,0 表示帖子内容为真,1 表示帖子内容为假;f 表示在训练数据集上学习到的决策函数。

早期主要侧重于人工特征工程,比如提取帖文的内容特征、用户特征以及传播特征(见图 7),利用传统的机器学习方法对信息的真假进行判别。

然而,随着智能技术的发展,虚假信息的生成方式以及表现形式也在不断地变化,人工特征难以及时应对新出现的虚假信息形式。为了能自动化地捕获假信息判别的高价值线索,一些基于深度学习的方法相继被提出。在这些检测模型训练过程中通常采用交叉熵作为损失函数(见式(2)),以衡量模型预测的概率分布和真实概率分布之间的相似性,其值越小说明分布越相似。

$$loss=-\frac{1}{N}\sum_{i=1}^{N}[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)] \tag{2}$$

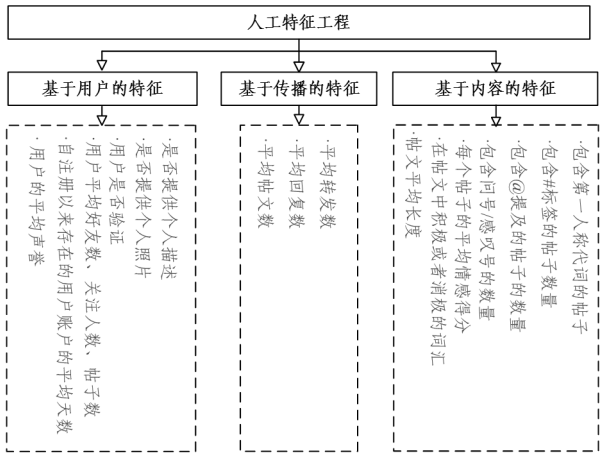


图 7 常见的人工总结的虚假信息特征

Fig. 7 Common features of artificially summarized fake news

3 面向社交媒体的虚假信息检测方法

根据分析视角的不同,进一步将基于深度学习的检测

1) 该条数据来自 PHEME 数据集,原推文中的英文在本文中被翻译成了中文。

方法划分为基于内容语义的检测方法、基于社交上下文感知的检测方法以及基于知识驱动的检测方法。

3.1　基于内容语义的虚假信息检测方法

基于内容语义的方法将检测任务视为特征挖掘过程，利用深度学习模型，如卷积神经网络、循环神经网或注意力机制等，挖掘文本、图像等不同模态数据中潜在的表现模式和语义线索来验证信息的真实性。根据数据类型的不同，将其划分为基于文本内容的检测方法和基于多模态语义关联的检测方法。

3.1.1　基于文本内容的检测方法

该类方法聚焦文本内容自身，利用语言模型实现文本内容语义关联的自动化捕捉。根据算法对文本建模方法的不同，进一步可将基于文本内容的检测方法划分为基于文本序列建模的方法和基于文本图建模的方法。

基于文本序列建模方法通常将帖文看作单词序列，然后借助循环神经网络、预训练语言模型等方法对序列的处理优势，揭示假信息内容中潜在的逻辑冲突或叙事不连贯[14]。

基于文本图建模方法通常以文本图的形式刻画文本内在语义关联，并在此基础之上，利用图神经网络对高阶邻域信息进行深度聚合，以增强模型对帖文内容细粒度语义的探索，克服序列模型难以捕捉词汇间长距离语义依赖的缺点[15-17]。例如，文献[16]将文本转换为语义-实体图，其中节点不仅包含原帖文中的词语，而且利用世界知识和语言知识库扩展了语义词，从而丰富了原社交媒体文本的表示，从一定程度上缓解了数据稀疏共现问题；文献[17]将文本转换成句法依存图，并通过设计子图注意力聚合和关键词去偏模块，有效提升了虚假新闻的检测性能。

这些方法虽在虚假信息早期检测方面取得了一定的进展，仍存在一些局限性。例如，现有基于内容的检测方法易受轻微的语义翻转或者虚假细节更改的影响，导致检测性能大幅下降。同时，社交媒体帖文的短文特性以及假信息呈现的多样性与隐蔽性，导致仅依赖单一文本模态数据难以全面捕捉假信息判别的有效线索。

3.1.2　基于多模态语义关联的虚假信息检测方法

社交媒体帖子中的图文数据共同构成了事件的完整叙事，因此二者在内容语义上互为补充：图像为模型理解文本内容提供了直观参照，文本描述则为模型解读图像内涵提供了语义指引。基于多模态语义关联检测方法通过整合文本与图像语义，揭示跨模态信息之间的关联与矛盾，从而能够有效提升多模态假信息识别的性能。其检测的一般流程如图 8 所示。

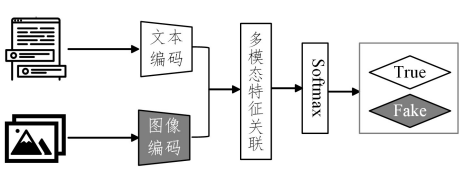


图 8　跨模态语义关联的假信息检测的基本流程

Fig. 8　Basic process of fake news detection for cross-modal semantic association

首先，通过文本编码器和图像编码器分别提取文本和图像特征；然后，将两种模态特征进行关联融合，形成混合模态特征；最后，将混合模态特征用于信息的检测。根据跨模态特征关联方式的不同，将其大致划分为两大类别：基于图文一致性推理和基于图文特征增强。

1）基于图文一致性推理的检测方法

该方法的核心理念在于，通过揭示多模态信息间的语义不一致性，如图像内容与文本描述间的矛盾，来识别信息中的谬误。其关键在于量化图像与文本之间的视觉-语义匹配程度。一种直观且简便的计算策略是：将两种模态的特征各自投射至一个共享的联合特征空间中，然后在此空间内进行相似度度量。例如，Zhou 等[18]提出的相似度感知假信息检测方法（Similarity-Aware Multi-modal Fake News Detection，SAFE）与 Xue 等[19]提出的多模态一致性神经网络（Multi-modal Consistency Neural Network，MCNN）均采用余弦相似度衡量图文特征的一致程度，并将此度量融入损失函数指导模型训练。两者的区别在于特征空间的融合策略：SAFE 创新性地借助图文描述模型，将图像转换为文本描述，随后利用 Text-CNN 技术将文本和图像特征转换到统一的向量空间；而 MCNN 则采取权重共享机制，将文本和视觉特征转换到公共特征空间，实现跨模态的一致性映射。这两种方法重点关注图文全局相似性的量化，但在微观层面的细粒度辨析上存在不足。

虚假信息通常是对真实信息进行局部修改得到的，如对关键事实进行夸张渲染或虚构。为了实现更细粒度的跨模态一致性推理，研究人员开始利用信息内在的结构化（图形）表示来细粒度地描述文本和图像内容，同时实现错误信息的精确定位。例如，Fung 等[20]所设计的“信息外科医生”模型（Information Surgeon Model，Info-Surgeon）采用文献[21]中的抽象语义表示图（Abstract Meaning Representation，AMR）概念，将信息转化为一个多模态知识图谱的结构。该模型利用图注意力与消息传递机制，高效融合不同模态的特征信息。InfoSurgeon 通过分析并预测图中节点间关系的可信度，能够精准识别并揭示信息的真实属性，为虚假信息检测提供了一种新颖而深入的方法。文献[22]对这项任务进行了深入探索，专注于预测文本中的细粒度知识元素和图像的逻辑关系。其核心在于利用多模态 Transformer 模型融合视觉特征、文本语义和 AMR 图的线性化表示，并通过设计一种局部聚集机制，学习跨模态局部信息的上下文关联。该类方法虽然能够细粒度地捕捉图文不一致特性，并增强结果的可解释性，但是 AMR 图的构建既耗时又费力，同时会产生级联误差。

为消除文本和图像模态间存在的语义差异，确保融合特征能够准确反映二者之间的关联，防止由模态间语义鸿沟导致的信息不匹配问题，文献[23]提出了一种图文一致性增强的多模态假信息检测方法（Fake News Detection via CLIP-Guided Learning，FID-CLIP）。该方法利用文本-图像对比预训练模型，生成高质量跨模态特征，并基于此有效量化帖子的跨模态相似性。文献[24]提出了一种多特征并举的假信息检测方法（Bootstrapping Multiview Representations for Fake

News Detection, BMR), 通过随机拼凑真实信息中互不相关的图文数据, 构建图文不一致的样本集, 并与图文一致的样例合并, 形成新数据集。模型利用交叉熵损失学习隐含的跨模态一致性特征, 再经混合专家模型[25]与其他特征加权融合。

2)基于图文特征增强的检测方法

文本模态和视觉模态为虚假信息检测提供了各有侧重、相互补充的信息。因此, 跨模态间信息的相互增强, 可以帮助模型加深对帖子内容语义的理解, 从而改善信息的检测性能。最早, Jin 等[26]提出一种基于神经元级别注意力机制的循环神经网络(Recurrent Neural Network with an Attention Mechanism, attRNN)来融合图文信息。这种方法侧重于多模态内容的单向增强, 即在文本引导下突出重要的图片区域。Song 等[27]首次使用共注意力网络(Cross-modal Attention Residual and Multi-channel Convolutional Neural Networks, CARMN), 以更好地融合图文信息。为了实现图像语义特征的精细化捕获, 需采用先进的图像目标识别技术以高精度地锁定图像中至关重要的实体对象。例如, Wang 等[28]提出的知识驱动的假信息检测方法(Knowledge-driven Multimodal Graph Convolutional, KMGCN)基于 YOLOv3 模型显式提取图片的目标标签, 然后使用图卷积神经网络对文本中的单词和图片目标标签之间的相关性进行建模。类似地, Li 等[29]提出面向实体对齐与融合的多模态假信息检测方法(Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection, EMAF), 该方法利用 Faster-RCNN 模型提取图片的目标标签, 然后使用胶囊网络将文本中的名词和这些目标标签进行融合。尽管这些方法致力于在图文间建立目标级语义对应, 但由于受限于通用目标识别技术, 因此在一定程度上未能充分挖掘多模态数据中蕴含的高阶互补信息。为此, Qian 等[30]提出了一种基于分层上下文注意力网络的多模态假信息检测方法(Hierarchical Multi-modal Contextual Attention Network, HMCAN), 将获得的图像和文本表示反馈到多模态上下文注意网络中, 以融合模态间和模态内的关系, 同时通过设计分层编码网络来捕获丰富的分层语义, 用于假信息检测。此外, 文献[31]提出了一种实体增强的多模态假信息检测框架, 通过引入视觉实体来建模图片的高层语义, 减小多模态的语义鸿沟。

表 2 对基于多模态语义关联的代表性检测方法及其在 Weibo 数据集上[24]的检测性能进行了对比分析。可以发现: 在基于图文一致性推理的方法中, FID-CLIP 和 BMR 的准确率超过了 0.9, 表明通过图文对比预训练模型, 能够显著提升文本和图像之间的语义关联, 进而增强假信息检测的性能; 在图文特征增强方面, EM-FEND 表现最佳, 这得益于其对图像实体语义特征的细致提取, 因此模型能更精确地捕捉文本和图像的核心信息, 从而提高了跨模态关联表征的准确性。

表 2 基于多模态语义关联的虚假信息检测方法的分类与性能对比

Table 2 Classification and performance comparison of fake news detection methods based on multimodal semantic association

特征关联方式	算法	文本特征	图像特征	融合方式	评价指标(准确率)	数据集
图文一致性推理	SAFE[18]	Text-CNN	Img2sentence+Text-CNN	拼接+Muti-loss	0.790	Weibo
	MCNN[19]	BERT	ResNet50	注意力机制+Muti-loss	0.846	Weibo
	InfoSurgeon[20]	BERT+LSTM	图像实体与事件识别	图卷积神经网络+Muti-loss	—	Weibo
	FGVE[22]	OSCAR[22]	OSCAR[22]	多模态 Transformer+注意机制+Muti-loss	—	Weibo
	FID-CLIP[23]	BERT CLIP	ResNet CLIP	CLIP 相似度加权	0.907	Weibo
	BMR[24]	BERT	MAE[23]	混合专家模型[25]+Muti-loss	0.918	Weibo
图文特征增强	attRNN[26]	BiLSTM	VGG19	神经元级注意力	0.808	Weibo
	CARMN[27]	BERT	VGG19、CNN	协同注意力机制、多通道 CNN	0.865	Weibo
	KMGCN[28]	—	YOLOv3	图卷积神经网络	0.886	Weibo
	EMAF[29]	BERT	Faster-RCNN	胶囊网络	0.893	Weibo
	EM-FEND[30]	BERT	VGG19+entity detector+OCR model	共注意力机制	0.904	Weibo
	HMCAN[31]	BERT+分层编码网络	ResNet50	共注意力网络	0.885	Weibo

3.2 基于社交上下文感知的虚假信息检测方法

社交媒体本质上是一个由多元实体(如用户、帖子)和复杂关系(如转发、评论、好友关系)交织而成的异质网络。一条信息一经在社交媒体平台上发布, 通常会嵌入复杂的网络环境中, 随之附带丰富的上下文背景, 具体涵盖发布者特征、传播关系以及其他用户的评论和转发等元素。这些元素对信息的理解与评估具有重要影响。因此, 可以从不同的视角, 融合社交媒体上下文信息, 开展社交媒体上虚假信息的检测任务。根据侧重点不同, 大致可以划分为 3 个方向。

3.2.1 基于用户特征的虚假信息检测方法

该方法主要从信息发布或转发用户的个人资料中抽取基本特征, 深入探索社交媒体用户特征与假新闻之间的相关性。例如, Shu 等[32]基于真实世界的数据集来衡量用户对虚假新闻的信任程度。研究发现, 分享真实新闻的用户往往比分享虚假新闻的用户拥有更长的注册时间。Long 等[33]发现, 在基于内容的检测方法中应用用户个人资料(如政党归属、验证信息和位置)可以提高虚假信息检测上的性能。文献[34]将帖文真实性和用户可信度视为潜在的随机变量, 利用贝叶斯网络模型捕捉新闻真相、用户观点和用户可信度之间的条件依赖关系。

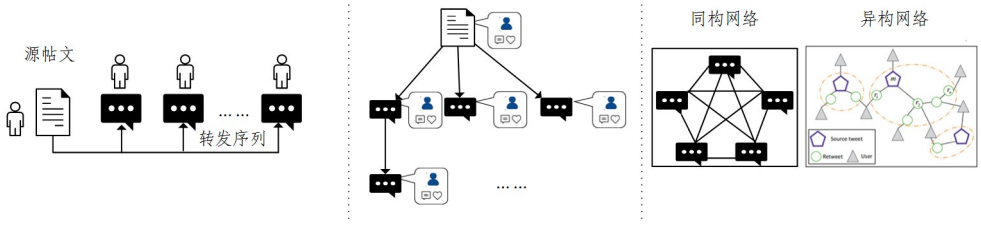
3.2.2 基于评论立场的假信息检测方法

用户评论是社交媒体中判断消息真实性的宝贵资源,

尤其在新话题或事件缺乏基础事实的情况下，评论能为核实新闻真实性提供线索，助力虚假信息的甄别[4]。因此，提炼用户对信息内容的立场观点，可有效提升虚假信息检测效果。例如，Shu 等[35]设计了一种句子-评论共注意力网络，通过注意力机制增强帖文内容与用户评论的深层次语义交互，改进虚假信息模型的检测性能。针对评论质量不一可能导致的噪声干扰问题，Wu 等[36]创新性地设计了一种融合决策树的共注意力网络模型。该模型通过分析用户特性，精选出高价值的评论内容，以此来优化并辅助源信息的真伪判定过程，有效提升了判别模型的准确性和鲁棒性。同时，Zhang 等[37]认为假新闻常常能引起人们的注意，高度唤起或激活人们的情绪，因此在人群中引起的新闻评论（即社会情绪）不应被忽视。然而，文献[38]的研究表明，个体用户的评论，作为其认知独特性的体现，包含非客观性与语义偏见，未必能准确反映客观事实。为此，在将评论作为假信息判别依据时，必须科学过滤噪声数据，提炼高质量评论信息，确保检测结果的可靠性。

3.2.3　基于传播结构的虚假信息检测方法

社交网络中，真伪信息的传播规律与模式存在显著差异。为此，众多研究人员致力于信息传播过程的建模（图 9 给出了 3 种典型的信息传播关系建模方法示例），利用智能算法挖掘信息传播特征，从整体上评估信息的真伪。



(a)基于时间序列的传播网络　(b)基于树结构的传播网络　(c)基于图结构的传播网络

图 9　信息传播过程建模示例

Fig. 9　Example of modeling the process of information dissemination

根据信息传播过程建模视角与技术手段的差异，可将基于传播模式的假信息检测方法细分为 3 类。

1）基于时间序列建模的方法。早期研究者将信息传播过程视为一个时间序列集合，运用循环神经网络[39]或改良的 Transformer 模型[40]来对该序列进行建模。此类方法旨在捕捉信息传播过程中的时序语义与结构特征，通过模型对时间序列中各时间节点信息的处理与学习，揭示传播动态与趋势。

2）基于树结构建模的方法。此类方法将信息传播过程抽象为一个树状结构。其中，源信息被视为树的根节点，而每一次转发或评论行为则生成新的分支节点或叶节点。在此传播结构基础之上，采用递归神经网络[41]、Tree Transformer 方法[42]或图神经网络[43-44]等技术，对树结构中的各个节点进行深度建模，旨在揭示信息传播路径、节点间关系以及传播影响力等复杂特性。

3）基于图结构建模的方法。由于图在建模复杂交互时具有天然的优势，研究者开始转向应用图结构建模信息的传播过程，以揭示信息在网络环境中的传播机制与模式。根据节点与边的类型差异，信息传播图可分为同质传播图和异质传播图。其中，在同质传播图中以单一类型的节点和边来模拟信息传播过程。例如，文献[45]通过自顶向下的传播图以及自底向上的扩散图描述信息传播的过程，利用双向图卷积神经网络学习信息传播的结构特征；文献[46]通过删除边、mask 点以及提取子图的方式对传播图进行数据增强，利用对比学习获取传播图的高阶特征，以便更好地理解信息的传播过程。信息传播过程本身就是一个动态的过程，而以往的方法忽略了信息传播过程的动态变化。对此，文献[47-48]将信息传播过程建模为一张动态图，利用动态图神经网络得到信息传播的动态图表征，取得了相对不错的效果。

尽管同质传播图简化了模型，便于分析单一交互模式下的信息扩散，但其仅关注一种节点类型和关系类型，限制了对多元节点与关系共存情境下复杂信息传播现象的刻画。在社交网络的复杂环境中，信息传播不仅涉及评论、转发等互动行为，还涵盖了参与其中的各类用户节点以及这些用户之间的关注关系等多元类型的边。整合这些多维度的节点与关系数据，对提升信息传播现象的洞察能力和模型的检测性能也至关重要。因此，研究者尝试利用异质图对假信息的传播过程进行建模[49]。

表 3 针对几种典型的基于传播模式的假信息检测方法进行了系统性对比。

表 3　典型的基于传播结构的假信息检测方法的性能对比分析

Table 3　Performance comparison and analysis of typical fake news detection methods based on propagation pattern

算法	源帖文本内容	社交上下文			评价指标		数据集
		用户特征	评论/转发内容	传播结构	准确率	F1	
GLAN[49]	√	√	√	异质图	0.9460	0.9460	Weibo
GCAN[51]	√	√	—	时间序列同质图	0.8770	0.8250	Twitter
BiGCN[45]	√	—	√	同质图	0.8800	—	Twitter
RDE[46]	√	—	√	同质图	0.8800	—	Twitter
HDGCN[50]	√	√	√	异质图	0.9610	0.9600	Weibo
UPFD[52]	√	√	—	传播树	0.8462	0.8465	Politifact
CIPHER[53]	√	√	—	异质图	0.8720	—	Politifact
HG-SL[54]	√	√	—	异质图	0.9005	0.8993	Politifact

从表 3 可以发现：

在 Weibo 数据集上，HDGCN 展现出比 GLAN 更优越的性能，二者均擅长利用异质网络整合多种信息（如评论、用户特征、用户间的交互等）以提升虚假信息检测。HDGCN 的核心优势在于其基于异质图的分层注意力机制，该机制更精细，能深入发掘并利用图中的复杂结构信息。

在 Twitter 数据集上，GCAN，BiGCN 和 RDEA 展现出

相近的检测性能。三者皆在同质传播网络上进行建模，区别在于 GCAN 侧重于利用用户属性特征，而 BiGCN 与 RDEA 则侧重于转发或评论内容。这反映出在 Twitter 数据集中用户特征与内容特征同等重要。

在 Politifact 数据集上，HG-SL 模型的性能表现优于 CIPHER 与 UPFD 模型。这 3 个模型均结合了帖子内容与用户特征进行分析。然而，UPFD 因采用树状传播模型，相较于另外两者，在性能上略显不足，这一差异间接证明了异质图在捕获传播动态的细微层面较树形结构更具优越性。HG-SL 之所以能取得更优异的表现，关键在于模型对用户行为的处理更加精细，有效融合了用户的全局背景与局部互动信息，因此在性能上超过 CIPHER 模型。

随着智能技术的发展，更适配于消息传播的信息整合模型（如基于消息发布时间的序列化模型、基于信息传播轨迹的树结构/图结构模型），更简便的子任务协同训练框架（如多任务学习），使得神经网络模型在虚假信息判别上的性能不断提升，但在信息的早期传播阶段，由于社交上下文信息的缺乏，其早期检测性能不佳。

3.3 基于知识驱动的虚假信息检测方法

社交媒体帖文内容不仅包含反映行文模式与情感色彩的词汇，更包含了揭示主题核心的实体词。引入这些实体相关的背景知识，不仅能有效扩充短文本的语义内涵，还可以为假信息识别提供额外的证据支持。基于此，研究者提出了基于知识驱动的假信息检测方法。该类方法的核心理念在于，通过对待检测信息内容中提取的关键要素与已知事实或常识知识进行对比，来评估信息的真实性[55]。

该类研究通常立足于一个基本假设：信息真实性判断的关键在于其对实体描述和事实陈述的准确性，而非表面的修辞表述。信息文本语境下实体的语义含义与其背景知识间的差异，有助于揭示信息内容的不一致性与可疑之处，为判断信息真伪提供关键依据。此类方法的核心主要包括两大部分：一是如何高效获取可靠的已知事实或常识知识；二是如何精准提取待检测信息特征并与已知事实进行有效对比。依据知识获取方式的差异，大体可以将其划分为：基于显式知识对比和基于隐式知识推理的虚假信息检测方法。

3.3.1 基于显式知识对比的虚假信息检测方法

该类方法基于外部高质量知识库（如维基百科、知识图谱、高质量佐证语料等），利用实体链接技术[1)]获取与文本内实体紧密相关的结构化关联知识和文本类背景知识[56]，具体过程如图 10 所示。针对待检测帖文，首先运用实体链接技术识别其中涉及的实体词语；接着，利用专业实体抽取工具（如利用 tagme 从维基百科抓取英文实体的详细描述信息，或借助高质量知识图谱 API 接口获取结构化知识数据）从外部知识库中检索与识别出与实体紧密相关的事实性、常识性知识。

在获取核心实体与其背景知识的基础上，通过对比待检测信息中实体的上下文语义与外部获取的实体背景知识，揭示潜在的不一致性和虚假线索。当前，主要存在两种知识对比策略。

1）对比网络：设计专门的对比网络架构，以量化并捕获信息内部实体语义与基于外部知识库的实体背景语义之间的差异。例如，文献[56]采用了向量差与点积运算，精准衡量上下文实体语义与基于外部知识的实体语义之间的偏离程度。

2）注意力机制：通过注意力机制，将外部知识库获取的知识与原始信息进行联合注意力处理，借助事实或常识知识聚焦源信息中可能存在的可疑片段。这种方法通过知识引导的注意力分配，强化对潜在虚假信息特征的关注与识别[57]。

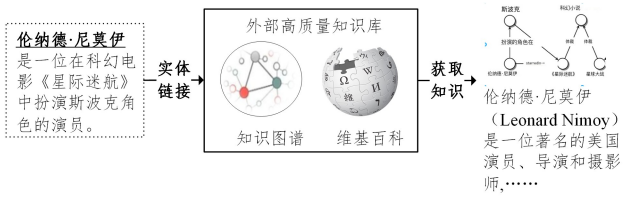


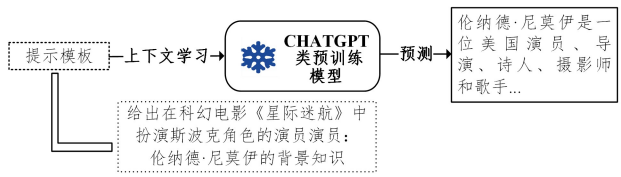
图 10 基于外部高质量知识库的实体知识获取示意图

Fig. 10 Schematic diagram of entity knowledge acquisition based on external high-quality knowledge base

3.3.2 基于隐式知识推理的假信息检测方法

基于 Transformer 架构的大规模预训练语言模型，通过在海量数据集上的预先训练，积累了丰富的事实、常识知识，将这些知识内化于模型参数，展现出强大的语言理解和上下文学习能力。因此，隐式知识获取策略倾向于将此类预训练语言模型视为内嵌式知识库，通过探查模型内部参数，提取所需的事实与常识信息。针对不同类型的大语言模型架构，其内在知识的访问手段亦有所差异，具体如图 11 所示。

（a）基于 BERT 类预训练语言模型的隐式知识获取方法



（b）基于 chatGPT 类预训练语言模型的隐式知识获取方法

图 11 基于预训练语言模型的隐式知识获取方法示意图

Fig. 11 Schematic diagram of implicit knowledge acquisition method based on pre-trained language model

对于以 Transformer Encoder 为代表的 BERT 类预训练语言模型，采用掩码（Mask）的方式调用内部知识。具体而言，利用预训练期间的“掩码”技巧，通过预测隐藏实体词来获取模型内部的事实性知识[58-59]。进一步，结合预训练好的蕴含关系预测模型[55]，可从源信息与由 BERT 模型生成的“证据”句子间识别出潜在的蕴含特征。这些特征随后可被输入多层感知机中，以服务于最终的事实

1) 英文主要为 tagme 实体链接工具（https://sobigdata.d4science.org/group/tagme/）；中文主要是复旦大学开源的知识图谱 API：中文采用复旦大学研发的中文概念图谱 CN-Probase API 接口（http://kw.fudan.edu.cn/apis/cnprobase/）

验证与对比分析，具体过程如图 12 所示。

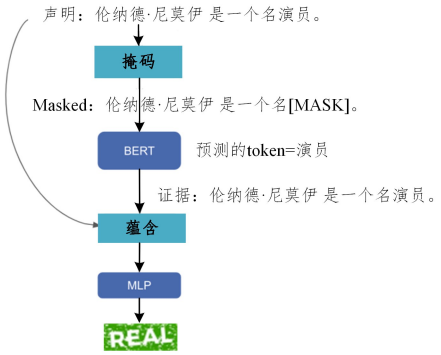


图 12　基于 BERT 隐式知识推理的虚假信息检测方法的整体流程[55]

Fig. 12　Overall process of fake news detection method based on BERT implicit knowledge inference[55]

对于以 Transformer Decoder 为代表的 ChatGPT 类大规模预训练模型，由于其卓越的语言理解能力、强大的上下文学习能力以及文本生成能力，通常采用提示学习方法获取模型内部知识。这种方法通过精心设计的提示信息，引导模型深入检索其内部知识，以实现对事实与常识知识的精准获取及对比[60-61]。图 13 展示了基于思维链（Chain-of-Thought，CoT）的提示信息方法，用于隐式知识推理的虚假信息检测示例。

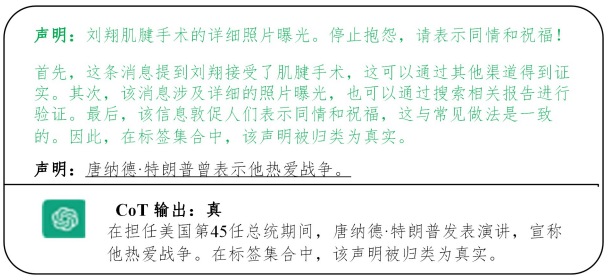


图 13　基于思维链提示的隐式知识推理虚假信息检测方法示例

Fig. 13　Example of implicit knowledge inference fake news detection method based on CoT prompting

4　虚假信息检测的可解释性研究

当前，面向社交媒体的虚假信息检测方法侧重于利用深度学习模型整合更多外部信息，自动挖掘隐藏特征，提高假新闻检测性能。然而，随着模型复杂度的增加，模型内的决策过程也越来越难以解释和验证。相关研究[62]表明，虚假信息的传播能力与事件的重要性和模糊性密切相关。因此，仅仅将信息标记为假通常是不够的，模型还必须自动提供判断依据，以增强判断过程和结果的可解释性。

针对社交媒体中虚假信息的检测可解释性研究集中于两大核心策略：注意力机制驱动的方法与解释生成导向的技术。这些策略旨在增强模型检测结果的可解释性，为检测结果提供判断依据。

4.1　基于注意力机制的可解释假信息检测方法

注意力机制的设计灵感来源于人类大脑在认知外界事物过程中对信息的自动筛选以及对关键信息的高效捕获。其核心理念即为对输入数据的不同组成部分施加差异化的权重，引导模型对关键信息给予更多关注。因此，采用注意力机制的假信息检测解释方法，通过可视化技术展示输入数据各部分分配的注意力权重，突出与判别结果紧密相关的事实词汇、信息属性及可疑用户等元素，从而实现对模型决策过程的直观解读[63]。例如，文献[64]通过可视化技术揭示了模型对虚假信息的注意力分布，发现表达怀疑、生气和其他情绪词要比事件相关词具有更高的注意力权值。文献[36]则创新性地设计了多种共注意力机制，以增强原文本与评论间的语义交流，实现通过评论内容精确定位原文中的可疑信息。反之，根据原文本选取关键评论作为证据支持。文献[51]运用图注意力机制，有效辨识信息传播链中的可疑用户，提升了检测结果的可解释性。为了克服之前基于图神经网络的传播方法未注意到传播网络在增强检测结果可解释方面的能力的问题，文献[65]提出了一种基于子图推理的方法。该方法设计了一个层次路径感知的核图注意力网络，解释传播中哪些子图对于假信息检测是重要的。

尽管基于注意力机制的方法通过突显模型关注的输入部分，有效提升了检测结果的可解释性，但目前针对注意权重分布的合理性和有效性评估大多依赖人工审查（例如文献[35]利用 Amazon Mechanical Turk 平台进行人为评估假信息评论的权重排序的合理性）。因此，为验证注意力权重分布的合理性，需建立一种能够自动评估注意力分布合理性和有效性的方法。

4.2　基于解释生成的假信息检测方法

基于解释生成的假信息检测方法的核心理念在于从提供的证据材料中生成详尽的结果判定的解释说明。例如，为了从候选事实生成易于人类理解的解释说明，文献[66]提出了一种基于知识图谱和文本语料的可解释假信息检测方法。该方法利用 Horn 子句形式编码背景知识，将抽象的候选事实编织成直观清晰的解释，极大地增强了模型结果的可解释性。这种方法主要在集成外部知识库的场景中展现出效用，但在应对基于语言模型推理结果时的效能不足，无法有效提供解释。为此，文献[67]提出了一种基于 Transformer 架构的解释生成方法，侧重于信息的精炼提取与高层次抽象，进而直接生成关于数据真实性的清晰阐释。然而，该方法主要通过提取判别结果说明的摘要来生成解释，缺乏直接依据原始证据进行逻辑推演的能力，无法展现从证据到结论的思维形成过程。为此，文献[68]提出了一种自动事实核查系统 Ta'keed。图 14 描述了该系统的整体架构，其主要包含两大组件：信息检索和基于大型语言模型的声明验证模块。给定待核查的信息，首先，利用搜索引擎检索相关的证据片段，选取排序在 top3 的结果；然后，借助大模型强大的自然语言理解、生成以及推理能力，基于检索到的证据和待检测的推文生成提示，诱导大模型结合外部证据对给定信息的真实性进行判别，并生成对应的解释。

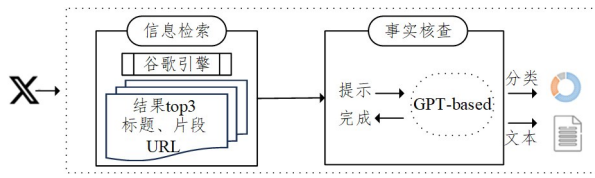


图 14　Ta'keed 整体架构示意图

Fig. 14　Overall architecture of Ta'keed

目前，对于生成的解释，主要利用摘要评价指标（如ROUGE[1)]和相似性度量等）测评解释生成的准确性与语义的完整性。这种评估方法的局限性在于需要有一个解释的标准进行对比，然而这种标准数据的构建需要花费大量的人力。

5 社交媒体虚假信息检测面临的挑战与机遇

5.1 当前虚假信息检测任务面临的挑战

生成式人工智能的迅速发展，引发了检测技术与生成技术之间潜在的“军备竞赛”。抗击虚假信息的检测技术尽管已取得了一定的成效，但仍面临一些挑战和问题。

1）现有检测技术难以全面覆盖各类精心设计的虚假信息。

目前，社交媒体中的虚假信息呈现出复杂多样的形态，要求检测模型不仅需具备强大的多模态融合与深度语义理解能力，能够精准识别并量化图文间的语义偏离与逻辑冲突；而且需具备对视觉欺骗技术如图像篡改、合成、伪造等的敏锐洞察力；此外，还需对情感倾向、立场偏见、社会心理效应等软性因素具有高度敏感性，以揭示信息背后可能的操纵意图与传播策略。因此，如何构建一个既能在宏观层面把握多模态信息的全局关联，又能在微观层面深入剖析每个模态内部细节的复杂虚假线索的多维度、多层次模型，仍是社交媒体虚假信息检测领域亟待解决的难题。

2）面对需要事实验证的虚假信息检测任务，对新兴领域或快速演变事件的知识库的即时更新与有效支持成为难点。

由于知识库更新滞后或信息不全，基于知识驱动的假信息检测系统往往面临性能瓶颈，特别是在诸如早期新冠肺炎症状等快速发展的议题上，知识库初始构建时可能已包含噪声或偏差，这些因素会直接影响基于知识驱动的假信息检测的准确性和可靠性。因此，确保知识库既能与时俱进并实时反映最新知识状态，又能保持高度的准确性和一致性，避免过时、错误或有偏信息对事实核查性能的消极影响，是提升知识驱动的假信息检测效能的关键。

3）当前的检测技术在处理人类创作与机器制造的虚假内容时，表现出较大的性能差异。

随着生成式人工智能技术的兴起彻底革新了数字内容的创作景观，信息产生方式正经历范式的转变，如图 15 所示。我们正处于人类/机器生成的真/假信息混合的社交生态中，这就要求假信息检测器能够同时有效处理机器生成与人类撰写的假信息。然而，当前的虚假信息检测技术在处理人类创作与机器制造的虚假内容时，表现出较大的性能差异[69]。这一局限性部分归咎于研究设计上的偏颇——或是完全基于人类编写信息的假设，或是简单化地将所有机器生成的信息一概视为虚假。

鉴于此，亟待开发新一代假信息检测工具，它应当深刻理解并能有效区分 4 种信息类型（机器转述的事实消息、机器原创的虚构信息、人为编造的假信息以及真实的人类报道）之间的细微差异。这将是智能生成时代推进信息真实性鉴别技术进步的关键步骤。

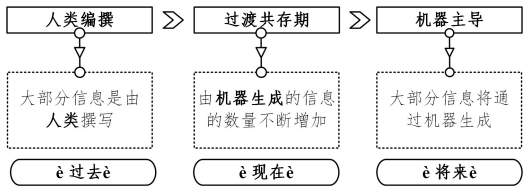


图 15 信息产生方式的演变

Fig. 15 Evolution of the way of information generation

5.2 大模型驱动虚假信息检测新范式

自回归式的大规模预训练模型由于其卓越的语言理解能力、强大的上下文学习能力以及文本生成能力，在很多自然语言处理任务上，无需调整模型参数，仅通过推理方式，即可获得较为出色的表现。因此，大模型为虚假信息检测技术研究提供了新的机遇。目前，可采用 3 种类型的范式将大模型应用于虚假信息检测。

1）基于大模型内部知识进行检测

该方法的关键在于高效的提示和推理策略设计。通过将定制的提示融入待检测的虚假信息内容，进而依托大模型出色的自然语言理解能力和其在庞大的预训练数据中积累的深层次知识进行深度剖析（见图 11(b)）。这种方法尤为适用于那些可通过语言特征（如情感倾向、逻辑矛盾）而非直接事实查证来揭露的虚假信息，以及那些在模型预训练与微调过程中已充分涉及相关背景知识的情形。

2）基于大模型驱动检索增强的虚假信息检测

面对需要事实验证的虚假信息检测任务，当大模型内部知识不足以覆盖所有必需背景知识时，采用大模型驱动的检索增强策略显得尤为重要。该策略通过动态检索和整合来自数据库或网络的补充信息，为大模型提供必要的验证上下文，确保即便在模型原生知识局限下，也能精准高效地执行信息真伪辨别任务。

图 16 给出了基于大模型驱动检索增强的基本流程。首先，将待验证的内容转换成一系列查询；然后，通过调用 API 从数据库或互联网检索相关文档，并将其作为上下文融入 Prompt 提示中；最后，借助大模型强大的上下文学习和推理能力，对待检测信息进行深度剖析与判别，从而提升检测的准确性和全面性。

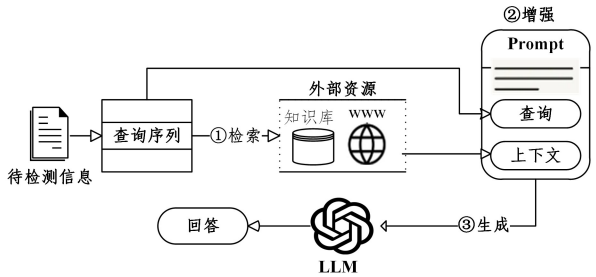


图 16 基于大模型驱动的检索增强示意图

Fig. 16 Schematic diagram of retrieval enhancement based on large model driven

3）大模型作为虚假信息检测的“咨询顾问”

大模型依赖自身的知识储备以及强大的语言理解能力，

1) ROUGE(Recall-Oriented Understudy for Gisting Evaluation)通过比较算法生成的内容与参考内容之间的重叠情况，来评估生成文本的质量。

能够洞察假信息并提供多视角的分析。然而，由于无法正确选择和整合这些理由，其直接作为检测工具的效能受到限制。因此，一种更行之有效的策略应运而生：将大模型定位为辅助决策的“咨询顾问”，而非直接的“裁判者”。图 17 给出了大模型作为虚假信息检测的“咨询顾问”的研究框架。在此框架下，大模型专注于为虚假信息检测模型提供丰富多元的视角和深度解析，助力模型从各个角度审视问题，从而作出更为精准和全面的判断。这种方法旨在充分利用大模型的强项，即生成多角度的洞见与论证，以增强整体检测系统的判断力与可靠性。

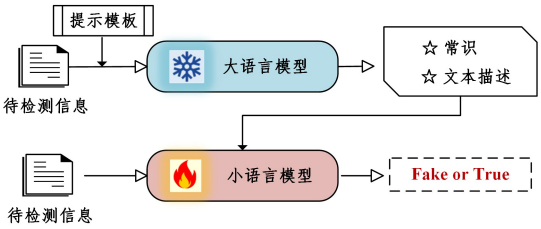


图 17　作为虚假信息检测的“咨询顾问”的大模型的研究框架

Fig. 17　Research framework of large model serves as “consultants” for fake news detection

6　总结与展望

本文针对智能生成时代下社交媒体虚假信息的检测技术研究展开综述。首先，在理论层面，探讨了虚假信息的内涵本质、产生机理以及表现形态，并对检测任务进行了形式化描述；其次，在技术层面，聚焦内容语义关联、社交上下文感知和知识驱动三大模块，对比梳理典型检测方法；并在此基础之上，深入探究增强检测算法可解释性的最新研究成果；进一步，从对抗博弈角度，细致剖析社交媒体虚假信息检测技术面临的挑战以及大模型驱动下虚假信息检测技术潜在的突破点。前人的研究为社交媒体虚假信息检测奠定了坚实基础，但这一领域仍有许多值得深入研究的问题。本文将从标准数据集构建、大模型驱动检测技术开发以及跨学科合作 3 个方面，提出一些值得进一步探讨的研究方向。

1）高质量、多样化标准数据集建设

数据是智能化虚假信息检测方法研究的基础。当前社交媒体上的虚假信息特征复杂多样，其不仅限于文本模态，还包括图像和视频等视觉模态以及复杂的社交上下文信息；在内容上涵盖多个领域（如政治、军事、健康、娱乐等），并表现出显著的语言多样性。例如，Twitter 中关于“俄乌冲突”话题的虚假信息涉及多达 11 种不同的语言。然而，现有的大多数据集局限于特定领域、单一语言或模态，且数据规模有限。同时，这些数据主要来源于传统社交媒体（如 Twitter、微博等），缺少新兴社交平台（如抖音等）等相关数据，故无法全面体现社交媒体虚假信息检测领域所面临的最新挑战。

一个优质的标准数据集应能够全面反映当前真实社交场景中虚假信息的复杂性和多样性。为此，在未来标准数据集建设研究中，应立足于主流社交媒体，确保数据在类型上涵盖多种模态，在内容上覆盖多个领域，在语言使用上涉及多种语言，同时混合人工、机器等多种产生方式生成的虚假信息，为开发更加通用、高效、智能化的虚假信息检测方法奠定基础。此外，数据集中的标签多由事实核查机构通过人工标注完成。随着数据规模的增大，这一过程所需的时间和人力资源不断增多，可以采用迁移学习的策略，利用现有的先进检测技术进行预标注，随后再由人工进行校对和修正。这种方法不仅能够提升数据标注的准确性，还能有效节约人力和时间成本。

2）基于大模型驱动多智能体协同的虚假信息检测框架研究

一个高效且通用的检测模型不仅需要强大的多模态数据融合与深度语义理解能力，还需要对风格、事实常识、视觉欺骗、立场偏见等多维度虚假信息线索具备敏锐的洞察力。大模型凭借其强大的自然语言理解和推理能力，在对抗假信息方面展现出巨大的潜力。为此，结合群体智能思想，以大模型为核心，设计多智能体协同的虚假信息检测框架，可增强多维度虚假线索的洞察能力。具体而言，通过精心设计模板以及相关技能的学习（如跨模态事实知识检测与整合能力），创建多个智能体，每个专责检测虚假信息的一个方面，例如文本风格、事实推理、社交背景、视觉操纵等。在此基础之上，开发一种协同机制使多个智能体能够相互协作（例如借鉴胶囊网络的动态路由机制），形成协同检测网络。

3）多学科知识协同的防御体系构建

随着智能生成技术的进步和信息的快速传播，构建一套有效的虚假信息防御体系显得尤为紧迫。然而，利用智能技术检测和减缓虚假信息的传播，仅是应对社交媒体虚假信息挑战的一部分。要全面解决这一问题，其还需跨学科合作，特别是社会科学、心理学、认知科学等多个领域知识的支撑。例如，探究人类信任假信息的原因和心理机制，进行针对性引导，对于增强个体主动防范意识至关重要；基于用户的认知偏差，优化检测算法，有助于提升对假信息的检测性能；分析人类的易感性，并将该特征融合到社交媒体平台的推荐算法中，有助于防止信息的回声效应，进而减缓假信息传播的影响等。因此，以社会、认知、心理等学科知识为内在动力，以智能技术为外在助力，通过跨领域知识协同，是未来构建虚假信息坚固防御体系的关键。

参 考 文 献

[1] HU Z R, HUANG C X, YAN S J. China New Media Development Report No. 14(2023) News, Communication [M]. Beijing: Social Sciences Academic Press, 2023.

[2] GUO B, DING Y S, YAO L N. The future of false information detection on social media: New perspectives and trends[J]. The Future of False Information Detection on Social Media: New Perspectives and Trends, 2020, 53(4): 1-36.

[3] ZHANG Z Y, JIN J C, LI F, et al. A Review of Online Social Network False Information Detection, Transmission, and Control Research from the Perspective of Artificial Intelligence[J]. Journal of Computer Science and Technology, 2021, 44(11): 2: 2261-2282.

[4] BHATTACHARJEE A, SHU K, GAO M, et al. Disinformation in the Online Information Ecosystem: Detection, Mitigation and Challenges[J]. Journal of Computer Research and Development, 2021, 58(7): 1353-1365.

[5] ALAM F, CRESCI S, CHAKRABORTY T, et al. A Survey on Multimodal Disinformation Detection[C]// Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2022: 6625-6643.

[6] HARDALOV M, ARORA A, NAKOV P, et al. A Survey on Stance Detection for Mis- and Disinformation Identification [C]// Findings of the Association for Computational Linguistics: NAACL. 2022: 1259-1277.

[7] VOSOUGHI S, ROY D, ARAL S L. The spread of true and false news online[J]. Science, 2018, 359(6380): 1146-1151.

[8] ZELLERS R, HOLTZMAN A, RASHKIN H, et al. Defending against neural fake news[C]// Advances in Neural Information Processing Systems. 2019: 9054-9065.

[9] FUNG Y, THOMAS C, REDDY R G, et al. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection[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). 2021: 1683-1698.

[10] SHU K, LI Y, DING K, et al. Fact-enhanced synthetic news generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 13825-13833.

[11] HUANG K H, MCKEOWN K, NAKOV P, et al. Faking fake news for real fake news detection: Propaganda-loaded training data generation[J]. arXiv: 2203. 05386, 2022.

[12] WANG W Y, CHANG C Y, PENG W C. Style-News: Incorporating Stylized News Generation and Adversarial Verification for Neural Fake News Detection[C]// Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. 2024: 1531-1541.

[13] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7008-7024.

[14] TRUEMAN T E, KUMAR A, NARAYANASAMYP, et al. Attention-based C-BiLSTM for fake news detection[J]. Applied Soft Computing, 2021, 110: 107600.

[15] WANG Y, WANG L, YANG Y, et al. SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection[J]. Expert Systems with Applications, 2021, 166: 114090.

[16] XU F, LI M H, HUANG Q, et al. Knowledge Graph Driven Graph Convolutional Neural Network Rumor Detection Model [J]. Chinese Science: Information Science, 2023, 53(4): 663-681.

[17] LIANG X, ZHANG Q, SHI C, et al. MSynFD: Multi-hop Syntax aware Fake News Detection[C]// Proceedings of the ACM on Web Conference 2024. 2024: 4128-4137.

[18] ZHOU X Y, WU J D, ZAFARANI R. SAFE: Similarity-Aware Multi-modal Fake News Detection[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2020: 354-367.

[19] XUE J, WANG Y, TIAN Y, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5): 102610.

[20] FUNG Y, THOMAS C, REDDYR G, et al. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection[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). 2021: 1683-1698.

[21] BANARESCU L, BONIAL C, CAI S, et al. Abstract meaning representation for sembanking[C]// Proceedings of the 7th linguistic Annotation Workshop and Interoperability with Discourse. 2013: 178-186.

[22] THOMAS C, ZHANG Y, CHANG S F. Fine-grained visual entailment[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 398-416.

[23] ZHOU Y, YANG Y, YING Q, et al. Multimodal fake news detection via clip-guided learning[C]// 2023 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2023: 2825-2830.

[24] YING Q, HU X, ZHOU Y, et al. Bootstrap multi-view representations for fake news detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 5384-5392.

[25] MA J, ZHAO Z, YI X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1930-1939.

[26] JIN Z W, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017: 795-816.

[27] SONG C, NING N, ZHANG Y, 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.

[28] WANG Y, QIAN S, HU J, et al. Fake news detection via knowledge-driven multimodal graph convolutional networks[C]// Proceedings of the 2020 International Conference on Multimedia Retrieval. 2020: 540-547.

[29] LI P, SUN X, YU H, et al. Entity-oriented multi-modal alignment and fusion network for fake news detection[J]. IEEE Transactions on Multimedia, 2021, 24: 3455-3468.

[30] QIAN S, WANG J, HU J, 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. 2021: 153-162.

[31] QI P, CAO J, LI X, 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. 2021：1212-1220.

[32] SHU K，WANG S，LIU H. Understanding user profiles on social media for fake news detection[C]// 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE，2018：430-435.

[33] LONG Y，LU Q，XIANG R，et al. Fake news detection through multi-perspective speaker profiles[C]// Proceedings of the Eighth International Joint Conference on Natural Language Processing(Volume 2：Short papers). 2017：252-256.

[34] YANG S，SHU K，WANG S，et al. Unsupervised fake news detection on social media：A generative approach[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：5644-5651.

[35] SHU K，CUI L，WANG S，et al. defend：Explainable fake news detection[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019：395-405.

[36] WU L，RAO Y，ZHAO Y，et al. DTCA：Decision tree-based co-attention networks for explainable claim verification[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：1024-1035.

[37] ZHANG X，CAO J，LI X，et al. Mining Dual Emotion for Fake News Detection[C]// Proceedings of the Web Conference 2021. 2021：3465-3476.

[38] WU L，RAO Y，LAN Y，et al. Unified Dual-view Cognitive Model for Interpretable Claim Verification[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). 2021：59-68.

[39] LIU Y，WU Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[40] KHOO L M S，CHIEU H L，QIAN Z，et al. Interpretable rumor detection in microblogs by attending to user interactions[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：8783-8790.

[41] MA J，GAO W，WONG K F. Rumor detection on twitter with tree-structured recursive neural networks[C]// Association for Computational Linguistics. 2018：1980-1989.

[42] MA J，GAO W. Debunking Rumors on Twitter with Tree Transformer[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020：5455-5466.

[43] HU D，WEI L，ZHOU W，et al. A rumor detection approach based on multi-relational propagation tree[J]. Journal of Computer Research and Development，2021，58(7)：1395-1411.

[44] CUI C，JIA C. Propagation Tree Is Not Deep：Adaptive Graph Contrastive Learning Approach for Rumor Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：73-81.

[45] BIAN T，XIAO X，XU T，et al. Rumor detection on social media with bi-directional graph convolutional networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：549-556.

[46] HE Z，LI C，ZHOU F，et al. Rumor detection on social media with event augmentations[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：2020-2024.

[47] SONG C，SHU K，WU B. Temporally evolving graph neural network for fake news detection[J]. Information Processing & Management，2021，58(6)：102712.

[48] SUN M，ZHANG X，ZHENG J，et al. Ddgcn：Dual dynamic graph convolutional networks for rumor detection on social media[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022：4611-4619.

[49] YUAN C，MA Q，ZHOU W，et al. Jointly embedding the local and global relations of heterogeneous graph for rumor detection[C]// 2019 IEEE International Conference on Data Mining(ICDM). IEEE，2019：796-805.

[50] KANG Z，CAO Y，SHANG Y，et al. Fake news detection with heterogenous deep graph convolutional network[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham：Springer International Publishing，2021：408-420.

[51] LU Y J，LI C T. GCAN：Graph-aware co-attention networks for explainable fake news detection on social media[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：505-514.

[52] DOU Y，SHU K，XIA C，et al. User preference-aware fake news detection[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021：2051-2055.

[53] SUN L，RAO Y，LAN Y，et al. Hg-sl：Jointly learning of global and local user spreading behavior for fake news early detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023：5248-5256.

[54] GHOSH S，MITRA P. How early can we detect? detecting misinformation on social media using user profiling and network characteristics[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham：Springer Nature Switzerland，2023：174-189.

[55] SHI B，WENINGER T. Discriminative predicate path mining for fact checking in knowledge graphs[J]. Knowledge-based Systems，2016，104：123-133.

[56] HU L，YANG T，ZHANG L，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). 2021：754-763.

[57] VO N，LEE K. Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics：Main Volume(Online). Association for Computational Linguistics，2021：965-975.

[58] SOLEIMANI A，MONZ C，WORRINGM. Bert for evidence re-

trieval and claim verification[C]// Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14—17, 2020, Proceedings, Part II 42. Springer International Publishing, 2020: 359-366.

[59] LEE N, LI B Z, WANG S, et al. Language Models as Fact Checkers? [C]// Proceedings of the Third Workshop on Fact Extraction and VERification(FEVER). 2020: 36.

[60] HU B, SHENG Q, CAO J, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 22105-22113.

[61] ZHANG X, GAO W. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method[C]// Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics(Volume 1: Long Papers). 2023: 996-1011.

[62] ROSNOW R L. Inside rumor: A personal journey[J]. American Psychologist, 1991, 46(5): 484.

[63] BASTINGS J, FILIPPOVA K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? [C]// Proceedings of the Third Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP. 2020: 149-155.

[64] CHEN T, LI X, YIN H, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]// Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22. Springer International Publishing, 2018: 40-52.

[65] YANG R, WANG X, JIN Y Q, et al. Reinforcement subgraph reasoning for fake news detection[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 2253-2262.

[66] GAD-ELRAB M H, STEPANOVA D, URBANI J, et al. Exfakt: A framework for explaining facts over knowledge graphs and text[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019: 87-95.

[67] ATANASOVAP. Generating fact checking explanations[M]// Accountable and Explainable Methods for Complex Reasoning over Text. Cham: Springer Nature Switzerland, 2024: 83-103.

[68] ALTHABITI S, ALSALKA M A, ERIC A. Ta'keed: The First Generative Fact-Checking System[C]// Computer Science & Information Technology(CS & IT). AIRCC, 2024.

[69] SU J, CARDIE C, NAKOVP. Adapting Fake News Detection to the Era of Large Language Models[C]// Findings of the Association for Computational Linguistics: NAACL 2024. 2024: 1473-1490.

CHEN Jing, born in 1990, Ph. D, lecturer. Her main research interests include natural language processing, social network analysis and knowledge engineering.

(责任编辑:柯颖)