社交媒体虚假信息检测技术研究综述

许旻辰[1]，屈丹[1,2]，司念文[1,3]，彭思思[1]，陈雅淇[1]

（1.信息工程大学信息系统工程学院，河南 郑州 450000；
2.先进计算与智能工程（国家级）实验室，河南 郑州 450000；3.清华大学电子工程系，北京 100084）

摘　要：实现及时有效的虚假信息检测有助于遏止虚假信息传播，降低社会危害。目前已有大量深度学习方法被用于虚假信息检测，总结现有研究的检测原理和检测范式对于明确技术优化方向至关重要。因此，结合虚假信息检测的原理和实现路径对现有研究进行全面综述，并首次对大语言模型在该领域的应用进行总结对比。首先，介绍虚假信息检测任务的相关概念，并汇总分析常用虚假信息检测数据集的数据结构；然后，根据检测原理和实现方式，分别介绍如何通过语义特征表示、辅助任务设计、内部知识推断和事实核查来检测文本和多模态虚假信息，将其细化为 10 个子类别，并总结分析各个子类别检测方法的潜在特性；最后，对基于深度神经网络和大语言模型的虚假信息检测范式进行总结，对比两种检测范式的代表性方法在 7 个虚假信息检测数据集中的检测性能，并归纳大语言模型检测虚假信息的优势和局限性，展望大语言模型给虚假信息检测领域带来的机遇与挑战，为后续研究提供参考。

关键词：深度学习；自然语言处理；虚假信息检测；大语言模型；事实核查

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0070287

Technologies for Detecting Disinformation in Social Media: A Comprehensive Review

XU Minchen[1], QU Dan[1,2], SI Nianwen[1,3], PENG Sisi[1], CHEN Yaqi[1]

(1.School of Information Systems Engineering, Information Engineering University, Zhengzhou 450000, Henan, China;
2.Laboratory for Advanced Computing and Intelligence Engineering, Zhengzhou 450000, Henan, China;
3.Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

【Abstract】Timely and effective disinformation detection is crucial for curbing the spread of disinformation and minimizing social harm. Numerous deep learning methods have been employed for disinformation detection. Summarizing the detection principles and paradigms of existing research is essential for identifying directions for technical optimization. Therefore, this paper comprehensively reviews existing research based on the principles and implementation paths of disinformation detection, and for the first time, summarizes and compares the applications of large language models in this field. First, the relevant concepts of disinformation detection tasks are introduced and the data structures of commonly used disinformation detection datasets are summarized. Then, based on detection principles and implementation methods, the paper presents ways to detect textual and multimodal disinformation through semantic feature representation, auxiliary task design, internal knowledge inference, and fact verification, refining them into ten subcategories and summarizing the potential characteristics of detection methods for each subcategory. Finally, the paper summarizes disinformation detection paradigms based on deep neural networks and large language models, compares the detection performance of representative methods from these paradigms across seven disinformation detection datasets, and highlights the advantages and limitations of large language models in detecting disinformation. It also presents the anticipated opportunities and challenges brought about by large language models in the field of disinformation detection, providing a reference for future research.

【Key words】deep learning; natural language processing; disinformation detection; large language models; fact checking

基金项目：国家自然科学基金（62171470）；河南省中原科技创新领军人才项目（234200510019）；河南省自然科学基金面上项目（232300421240）。

作者简介：许旻辰，男，硕士研究生，主研方向为虚假新闻检测、自然语言处理；屈丹（通信作者），教授、博士；司念文，讲师、博士；彭思思、陈雅淇，博士研究生。

收稿日期：2024-08-23　　**修回日期：**2024-11-28　　**E-mail：**qudan_xd@163.com

计算机工程 www.ecice06.com

0 引言

得益于互联网技术的全面普及，社交媒体用户数量实现迅速增长。以微博为例，截至 2023 年 9 月，月活跃用户数量已达 6.05 亿[1]。社交媒体平台的流行为信息传播带来了极大便利，但由于社交媒体存在用户基数大、受众群体广、信息传播成本低等特点，极易成为假新闻、谣言等虚假信息传播的温床。目前，社交媒体平台的信息审核机制无法实现及时有效的虚假信息监管，对社会稳定、国家安全造成潜在威胁。从提高网络空间环境治理能力的角度出发，迫切需要实现虚假信息检测以解决现实问题。

新闻传播学理论和机器学习等技术的进步推动虚假信息检测任务不断向前发展，目前，已有多篇综述文献对社交媒体虚假信息检测技术进行了总结。文献[2-3]将虚假信息检测方法整体归纳为依赖内容特征以及上下文特征的方法。鉴于虚假信息在主题场景、内容形式等方面的多样性，研究人员针对多模态信息[4]、新冠肺炎[5]等特定场景的虚假信息检测方法进行了总结。虚假新闻检测作为虚假信息检测的任务形式之一，文献[6]介绍了虚假新闻检测的在线工具和相关组织，并从新闻内容和社会背景 2 个角度总结检测方法。文献[7]从虚假新闻的传播意图特征入手，将虚假新闻的传播意图分为误导公众、操纵舆论、吸引注意 3 类。文献[8]从特定模型出发，按照知识驱动、传播特征和社会情境 3 个方面，对基于图的虚假新闻检测方法进行分类。

现有综述工作虽然从不同维度对虚假信息检测方法进行了梳理，但仍缺乏对检测原理和实现路径的系统性整合。为此，本文从虚假信息产生的基本过程出发，围绕不同检测方法的判断依据和实现方式，对社交媒体虚假信息检测技术进行系统梳理与整合。

1 框架结构

为探究不同检测方法提出的动机，本文基于虚假信息产生的基本过程，对虚假信息检测的依据进行分析。如图 1 所示，社交媒体信息产生的基本过程涉及 3 个关键对象：首先是背景信息，包括现实事件或特定知识，为社交媒体信息提供内容基础；其次是信息发布者，根据背景信息创建社交媒体信息；最后是社交媒体用户，其获取并传播社交媒体信息。在这一过程中，通过提取社交媒体信息中的声明，并根据事实或相关知识进行验证推理，可以直观地实现虚假信息检测。此外，新闻传播学的实证研究表明[9]，信息的内容特征以及社交网络特征能够间接

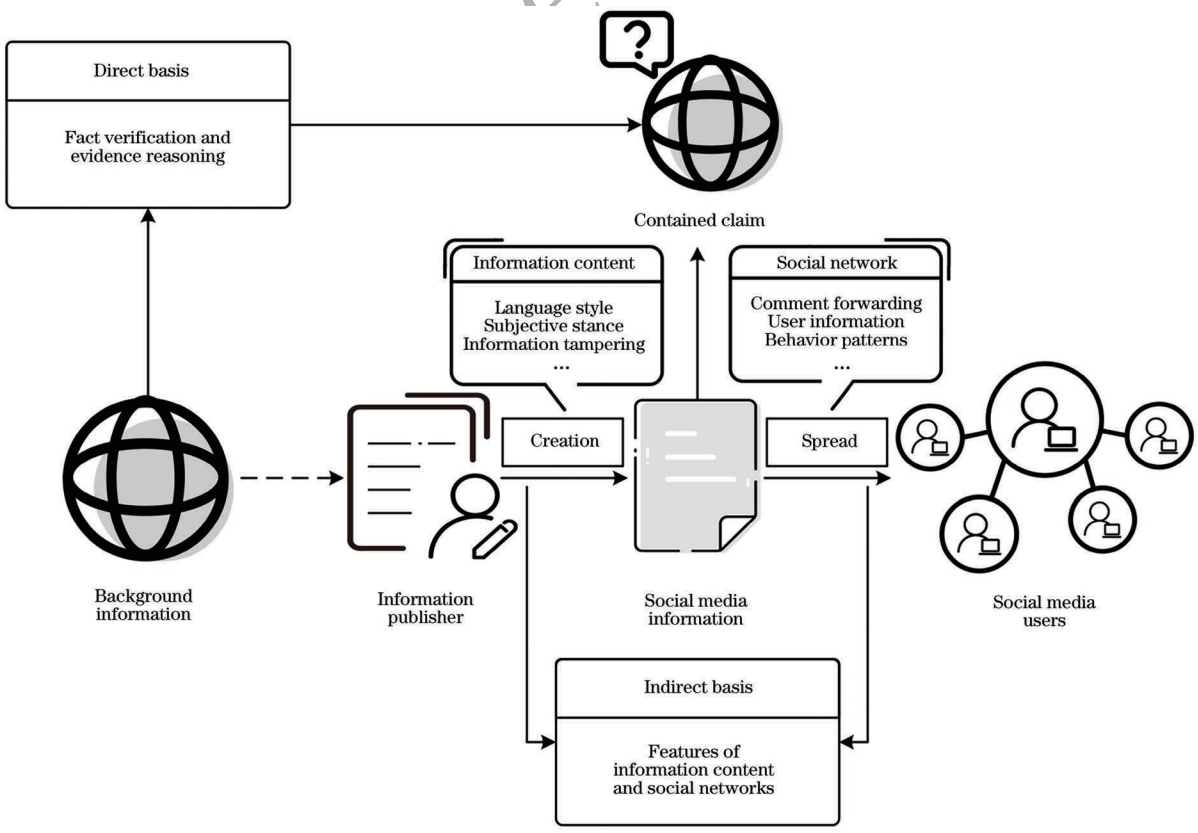

图 1 社交媒体虚假信息检测依据分析

Fig. 1 Analysis of the basis for detecting disinformation in social media

实现虚假信息检测，例如信息发布者的语言风格和主观立场、社交媒体用户的转发评论和基本信息等。

综合上述检测依据分析，本文结合虚假信息检测的判断依据和实现路径对已有研究工作进行分类，并首次对比分析了大语言模型进行虚假信息检测的优势和局限性。如图 2 所示，本文将信息内容特征与社交网络特征作为间接检测依据，将事实验证和证据推理作为直接检测依据。首先从获取上述检测依据的实现路径出发，将现有虚假信息检测实现路径分为语义特征表示、辅助任务设计、内部知识推断和事实核查 4 类；然后，对各个实现路径的具体方法进行分类细化，并总结具体方法；最后，对比分析大语言模型检测虚假信息的优势和局限性，从而为大语言模型时代虚假信息检测技术的研究发展提供参考。

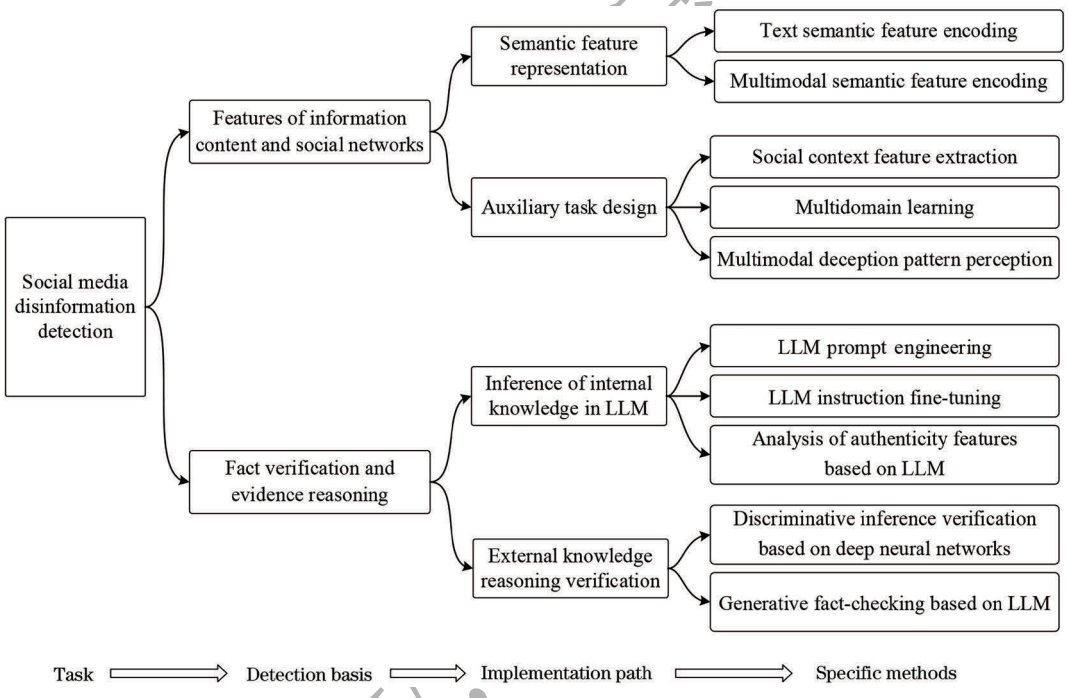

图 2 虚假信息检测方法分类

Fig. 2 Classification of disinformation detection methods

2 社交媒体虚假信息检测

2.1 相关概念及任务定义

在总结相关研究工作前，需要明确虚假信息的概念以及虚假信息检测任务的定义。本文根据虚假信息不准确、存在误导性的本质特征，参照国际电信联盟和联合国教科文组织对虚假信息的定义[10]，将用户有意或无意间发布的可能造成具体伤害的虚假或误导性内容划分为虚假信息，以尽可能全面地覆盖虚假信息检测的任务范畴，其具体表现形式包括虚假新闻和已经证伪的谣言信息。虚假新闻是指为达到某一目的而发布不实信息以欺骗他人的报道[7]，一般是由媒体机构发布的、未经证实或有意误导的新闻信息。与虚假新闻不同，谣言信息一般来自个人用户，是关于社会中某事件的特殊陈述或断言信息[11]。

综合上述概念，虚假信息检测本质上可以定义为一项分类任务，旨在鉴别信息内容的真实性，防止不准确或误导性的信息造成潜在危害。自然语言处理领域中的谣言检测、虚假新闻检测以及自动事实核查任务均可视为虚假信息检测任务在不同现实需求下的具体任务形式。

2.2 常用数据集

如表 1 所示，本文对常用的虚假信息检测数据集进行了统计分析。虚假信息检测数据集的内容可分解为新闻文本、帖子声明、社交上下文信息、分类证据以及解释说明这 5 个组成部分。其中，新闻文本和帖子声明是数据集中的待检测对象，主要来源于微博、Twitter、Politifact、Snopes 等社交媒体平台和事实核查网站；社交上下文信息和分类证据是进行真实性判断的依据之一；解释说明则是对检测样本分类结果的说明，用以提高检测结果的可信度，以及警示用户，干预虚假信息传播。

从数据集结构角度分析，考虑到基线检测方法或研究动机的不同需要，数据集结构上存在一定差异。以 LIAR[15] 数据集为例，最初的 LIAR[15] 数据

表 1 虚假信息检测数据集统计

Table 1 Statistics of disinformation detection datasets

Dataset	Sample size	Category count	Data source	Dataset structure					
				News content	Claim	Image	Social context	Evidence	Explanation
Twitter, Weibo[12]	1 154,4 649	2	Twitter, Weibo	—	√	√	√	—	—
PHEME[13]	4 842	3	Twitter	—	√	√	√	—	—
Twitter15, Twitter16[14]	1 490,818	2	Twitter	—	√	—	√	—	—
LIAR[15]	12 836	6	Politifact	—	√	—	√	—	—
LIAR-PLUS[16]	12 590	6	Politifact	—	√	—	√	—	√
LIAR-RAW[17]	12 590	6	Politifact	—	√	—	√	√	√
RAWFC[17]	2 012	3	Snopes	—	√	—	—	√	√
BuzzFeed[18]	2 263	4	Facebook, Twitter, etc.	√	—	—	√	—	—
FEVER[19]	185 445	3	Wikipedia	—	√	—	—	√	—
ISOT[20]	44 898	2	Reuters and other news websites	√	—	—	—	—	—
FA-KES[21]	804	2	Syrian Violations Documentation Center	√	—	—	—	—	—
FakeNewsNet[22]	23 921	2	Politifact, GossipCop	√	—	√	—	√	—
COVID-Social, COVID-Scientific[23]	340,172	2	Politifact, Kaggle, etc.	—	√	—	—	√	—
Twitter-COVID19, Weibo-COVID19[24]	400,399	2	Twitter, Weibo	—	√	—	√	—	—
WatClaimCheck[25]	33 697	3	Fact-checking websites	√	—	—	√	√	√
CHEF[26]	10 000	3	China Anti-Rumor Center, etc.	—	√	—	—	√	—
FakeNewsAMT[27]	980	2	News websites	√	—	—	√	√	—
Euvsdisinfo[28]	14 497	2	EUvsDisinfo database	√	—	—	√	√	—

集受限于当时的模型语义理解能力，并且基线检测方法不涉及事实核查，除检测对象外，仅包含信息发布时间、环境背景等社交上下文信息。LIAR-PLUS[16]在此基础上收集了事实核查人员的核查文章摘要作为检测证据之一。之后，为了满足构建事实核查系统的需要，LIAR-RAW[17]加入了通过搜索引擎得到的证据信息。此外，针对具有重大影响的社会热点事件，也提出了多个特定主题的数据集，例如针对新型冠状病毒感染的肺炎事件背景提出的 COVID-Social[23]、Twitter-COVID19[24]数据集，针对叙利亚战争事件背景提出的 FA-KES[21]数据集。

从信息真实性分类标准的角度分析，常见的分类尺度包括二分类、三分类和六分类。二分类是基础的分类标准，根据检测目标的真实性将样本分为“真实”和“虚假”两类。三分类方法在此基础上增加了“无法判断”这一类别，提高了实际应用中人工修正模型检测结果的效率。六分类方法是事实核查网站 Politifact 制定的分类标准，包括“真实”“基本真实”“半真半假”“基本错误”“虚假”“完全编造”6 类标签，对信息真实性程度进行了更细粒度的划分。

目前，虚假信息检测领域已经拥有了丰富的训练数据，涵盖多个事件主题，并具有多样的数据结构，为虚假信息检测技术的发展提供了优质数据基础。但是，大语言模型在虚假信息检测领域的广泛应用也为数据集的构建提出了新要求，例如，收集大语言模型训练数据终止日期之后的虚假信息和更为复杂的文本信息等，从而进一步研究大语言模型进行虚假信息检测的能力边界。

3 基于信息内容特征与社交网络特征的虚假信息检测

基于信息内容特征与社交网络特征的虚假信息检测方法，旨在提取信息内容和传播动态的深层特征，并将其映射为真实性标签。由于社交媒体信息的复杂性，一类研究侧重于提升模型对各个模态语义特征的表示能力；另一类则通过实证分析设计辅助任务，融入虚假信息检测任务的先验知识以增强检测效果。

计算机工程 www.ecice06.com

3.1 基于语义特征表示的虚假信息检测

社交媒体信息常见的表现形式包括纯文本和图文信息，主要涉及文本和图像 2 类模态特征。根据检测方法涉及模态特征表示范围的不同，将其分为基于文本语义特征的虚假信息检测方法和基于多模态语义特征的虚假信息检测方法。

3.1.1 基于文本语义特征的虚假信息检测方法

文本作为最常见的信息载体，能够进行高效的信息传播，因此，基于文本单模态语义特征的检测方法受到广泛关注。早期工作利用深度神经网络对社交媒体信息进行文本语义特征编码，替代人工特征工程检测，实现了检测性能的进步。文献[12]于 2016 年提出了利用循环神经网络(RNN)进行谣言检测的方法，对比分析了基础 RNN[29]、长短时记忆网络(LSTM)[30]、门控循环单元(GRU)[31]在谣言检测任务中的性能。利用 RNN[29]模型的记忆能力，对社交媒体中的文本信息进行建模表示，捕获相关帖子上下文信息随时间的变化，从而学习到文本语义信息的深层表示。

文献[12]的研究证明，文本模态的语义特征编码能够实现有效的虚假信息检测，但是由于 RNN[29]难以捕捉文本上下文的长程依赖，限制了模型的能力。针对此问题，文献[32]提出通过结合卷积神经网络(CNN)的注意力残差网络进行虚假信息检测，利用注意力机制捕捉文本信息的长程依赖，并应用窗口大小可变的 CNN 来选择重要的局部特征。此外，对于长文本新闻信息的上下文句子信息交互，文献[33]将每个新闻文档构建为全连通图，将句子的向量化表示作为图中的节点信息，句子两两之间的相似度作为边的权重，最后利用图注意力网络(GAT)[34]进行信息聚合，得到新闻文档的特征表示。

社交媒体信息的真实性可以与多种语义特征相关联，因此，多任务训练的思想被用于提高模型的虚假信息检测能力。文献[35]采用 Transformer 构建特征编码器，并根据虚假信息容易引发用户争议这一客观现象，设置了立场检测与谣言检测两项任务，提升模型的虚假信息检测能力。文献[36]进一步设置了新闻语句偏见检测、虚假新闻检测、谣言检测、诱惑式标题检测 4 个任务，统一了谣言检测和虚假新闻检测这 2 类主要的虚假信息检测任务。

以 Transformer 的双向编码器表示(BERT)[37]为代表的预训练语言模型凭借 Transformer 架构中的自注意力机制，被广泛用于强化虚假信息检测任务中的文本语义信息表示。FakeBERT[38]将预训练语言模型与具有不同核大小和滤波器的单层 CNN 相结合，以进行虚假信息检测。文献[39]直接微调预训练语言模型进行虚假信息检测，发现 BERT[37]等 9 个常用预训练语言模型微调后的谣言检测性能可以媲美甚至优于当时性能最佳的其他基线方法。得益于预训练语言模型较强的语义表示能力，文献[40]利用提示学习的方法，将谣言检测的分类任务转变为基于提示的完形填空任务，并采用预训练语言模型分层编码的方式，将较浅层的语法编码部分与语义编码部分分离，只对语义编码部分的模型进行微调，实现了零样本跨语言跨域的谣言检测。

除了改进模型结构以外，添加特征工程以丰富文本语义特征也是提升虚假信息检测性能的方法之一。文献[41]利用文本复杂性评估工具复杂性轮廓生成器(CoCoGen)[42]，增加了 6 种可解释性特征(句法复杂性特征，词汇的密集性、复杂性和变异性特征，信息论特征，n 元频率特征，语词分析和单词计数，词频测量)用于训练。文献[43]提取社交媒体信息中的情感词，并将其加入到文本表示特征中，从情感分析的角度丰富文本信息。

扩展特征工程的信息范围，将外部新闻环境的语义信息作为额外特征，能够帮助模型更好地掌握新闻事件的背景，从而得到更准确和全面的语义表示。文献[44]收集了长时间跨度(2010—2021 年)的主流虚假信息检测数据，并从新闻流行性和新颖性 2 个角度实现新闻环境的对比感知，丰富虚假新闻检测的判断依据。与新闻环境感知的思想类似，文献[45]认为多篇新闻文章间可能包含互补或矛盾的信息，由此提出了跨文档虚假信息检测方法，通过信息抽取，以知识图谱的方式表示文本语义信息，并将各个文档的知识图谱通过共指消解[46]连接为跨文档的事件知识图谱。最终，在事件和单个文档 2 个层面进行虚假信息检测，利用事件级知识图谱特征提高单个样本的虚假信息检测效果。

上述研究表明，文本语义特征能够作为虚假信息检测的依据，通过提升模型语义特征捕捉能力、添加特征工程强化特征表示等策略，有效改进了模型的检测性能。尽管如此，此类方法仅依靠学习文本语义特征到真实性标签的浅层映射来实现分类，由于判断依据较为单一，对于语言风格等方面迷惑性较强的虚假信息，其检测能力受限。

3.1.2 基于多模态语义特征的虚假信息检测方法

提取并融合社交媒体信息中图像、音频、文本等多个模态的语义信息，是检测多模态虚假信息的基

计算机工程 www.ecice06.com

本方法。文献[47]提出的 att-RNN 网络首次利用深度神经网络替代人工特征工程,实现多模态虚假信息检测,提升了检测性能。att-RNN 分别通过 LSTM 和视觉几何小组(VGG)的 19 层网络结构(VGG-19)提取文本和图像的模态特征,并利用注意力机制捕获模态特征之间的关系,得到与文本信息存在隐式关联的图像特征,最终将文本和图像特征拼接后进行多模态虚假信息检测。在此基础上,SINGHAL 等进一步强化了单模态特征的表示能力,先后在文献[48-49]中提出 SpotFake 框架和 SpotFake+框架,分别利用 BERT 和 XLNet[50] 提取文本模态信息特征,与图像语义特征拼接后进行分类检测。上述早期工作在进行多模态特征表示时,仅对各模态特征进行简单拼接或加权,容易导致多模态信息表示偏差,无法满足模态特征间进行复杂交互的要求,限制了检测性能的提升。

模态特征间的复杂交互是多模态语义特征编码面临的主要挑战,主要体现在模态对齐和模态融合 2 个特征交互过程中。模态对齐过程要求在不同的数据模态特征之间建立对应关系,提高模态特征间的语义一致性;模态融合过程要求整合多模态特征表示,充分利用模态特征间的互补性质。早期工作中对模态特征的简单拼接或加权容易导致模态信息表示偏差,无法捕获模态间潜在的语义关联,导致模型无法充分利用多模态数据的丰富信息。

在实际研究中,注意力机制凭借动态特征选择的优势,在多模态特征融合中能够自然地实现特征对齐,被广泛用于多模态特征交互。文献[51]提出了一种多模态协同注意力网络,堆叠多个协同注意力层进行模态间特征的双向交互,融合空间域、频域和文本 3 个模态的语义特征。文献[52]提出的层次化多模态上下文注意力网络利用 BERT 中间层的隐藏状态丰富语义信息,并构建上下文 Transformer 模块实现模态内和模态间的语义特征交互,该模块由 2 个参数独立的 Transformer 组成,分别负责模态内和模态间的语义特征交互。文献[53]引入了对比语言-图像预训练(CLIP)[54] 提取跨模态对齐的图文特征,并利用文本和图像之间的相关性加权得到融合特征,最终通过一个模态感知的注意力层衡量文本、图像和融合特征的重要性,得到用于分类的多模态特征表示。

此外,有研究通过学习跨模态的一致性特征表示来促进特征交互,而不仅仅依赖注意力机制。文献[55]提出端到端的多模态变分自编码器检测多模态虚假信息。该方法首先利用 LSTM 和 VGG-19 作为编码器分别提取文本和图像特征,并经过全连接层形成多模态特征,之后解码器从采样后的多模态特征中重构数据,设计重构损失迭代优化编码器和解码器,学习模态间的共享表示。文献[56]提出了一种二阶段检测算法,第一阶段设计图文样本匹配任务构造图文对的正负样本,其中正样本表示图像和文本信息来自同一多模态信息,在此基础上,提出跨模态知识蒸馏函数,训练文本和图像编码器,提高正样本对的语义相似度并降低负样本对的语义相似度,实现模态特征对齐。文献[57]首先改进了多门混合专家网络得到多模态特征的初步融合,并优化单模态特征表示,之后将单模态特征单独用于真实性分类,以此判断各模态特征重要性并进行加权,最后,采用类似其他工作的跨模态一致性学习方法,构造图文信息匹配任务,学习模态特征一致性表示,作为最终的决策依据之一。

3.2 基于辅助任务设计的虚假信息检测

研究人员通过实证研究得到社交媒体虚假信息的多种现实特征,因此,部分工作从虚假信息检测的实证研究结论出发设计辅助任务,引导模型学习特定模式特征。根据辅助任务所学习的实证特征,将其分为基于社交媒体上下文信息、基于多领域训练、基于多模态欺骗模式的 3 类虚假信息检测方法。

3.2.1 基于社交上下文信息的虚假信息检测方法

部分虚假信息由于内容较短或是语言风格与真实信息较为接近,只依靠语义特征难以进行有效检测。针对此问题,将社交上下文特征提取作为虚假信息检测的辅助任务,能够帮助获取更全面的特征视角,提升虚假信息检测性能。该方法属于文本虚假信息检测任务的主流方法,并且在低资源虚假信息检测任务中也展现出较大潜力。

社交上下文信息是指信息在社交网络中传播和交互所产生的外部特征,包括信息的传播结构、用户行为和来源可信度等。其中,结合信息的传播结构特征以及语义特征进行虚假信息检测得到了广泛研究。文献[58]将谣言信息的传播路径表示为按发帖时间排序的文本序列,利用 GRU 和 CNN 提取传播序列特征,并限制推文评论信息的时间范围,以提升谣言早期传播阶段的检测性能。文献[59]采用自上而下建模以及自下而上建模 2 种方式构造信息传播树,模拟信息扩散和信息聚合的传播过程,并通过树结构递归神经网络提取传播结构特征以及文本语义特征用于谣言检测。文献[58-59]将信息传播结构特征用于检测任务的早期探索,图 3 所示的传播树结构(根节点为原始信息,子节点为评论信息,边代

计算机工程 www.ecice06.com

表互动关系)对之后的社交上下文特征提取产生了深远影响。构建与图 3 类似的传播树并通过图神经网络进行特征处理,是目前主流的社交上下文特征提取方法。早期研究由文献[60]应用于谣言检测任务中,利用图卷积神经网络(GCN)[61]将虚假信息检测转为图分类任务。与文献[59]的思路类似,文献[60]按照自上而下和自下而上 2 种方式构建传播树,并拼接 2 类传播树,由 GCN 提取的图特征用于分类检测。尽管该方法实现了信息传播结构特征和语义信息的深度融合,但仍存在传播树忽视远距离信息交互、对用户间关系等社交上下文信息利用不足等问题。

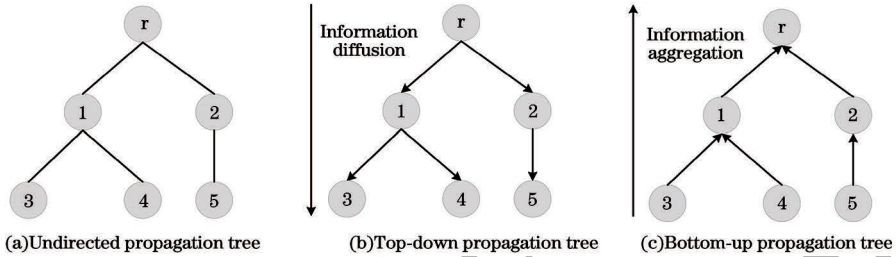

图 3 3 类常见的传播树结构

Fig. 3 Three common types of propagation tree structures

针对传播树忽视信息横向传播以及远距离信息交互的问题,研究人员从传播树构造以及特征提取方式 2 个角度提升检测性能。文献[62]对传播树的构造方法进行改进,连接每个信息传播分支内的兄弟节点,并采用无向图的方式模拟信息的横向传播交互,最终利用 GAT 对信息内容与传播树进行编码。文献[63]则改进了传播结构特征提取方法,将传播树按时间顺序展开为文本序列,利用自注意力机制模拟信息之间的交互。同时,为了减少传播树展开后结构信息的丢失,定义了 5 种节点对之间的相对结构关系,即另一个节点为当前节点的父节点、子节点、前节点(时间戳早于当前节点)、后节点(时间戳晚于当前节点)或节点本身。将这 5 种结构关系显式地表示为注意力机制计算过程中 2 个额外的可学习向量,使得结构信息和内容信息都可以在注意力机制中进行传播和交互。

传播树的构建只是将社交媒体信息传播的结构特征与语义特征进行了整合,面对越来越复杂的虚假信息伪造技术,结合账号信息、用户行为等其他社交上下文信息能够进行更有效的检测。文献[64]为了整合多层次的社交上下文信息,构建了新闻信息异构图,由 3 种类型的节点(新闻发布者、新闻信息、参与传播的用户)和 4 种类型的边(新闻发布者之间的相互引用关系、新闻信息和发布者之间的发布关系、用户和新闻之间的评论关系、用户之间的社交关系)组成。如图 4 所示,文献[65]构建了带有用户信息的异构传播树。首先,根据社交媒体信息的传播关系构建推文传播网络,之后爬取参与了信息传播的账号信息,包括用户的关注列表和粉丝数等元特征,根据社交网络关系构建用户社交网络,最后按照账号和推文之间的对应关系,将参与了信息传播的用户节点连接到推文传播网络中。该方法实现了谣言检测领域多种社交上下文信息的整合,能够提升突发事件等低资源场景下模型的检测能力。

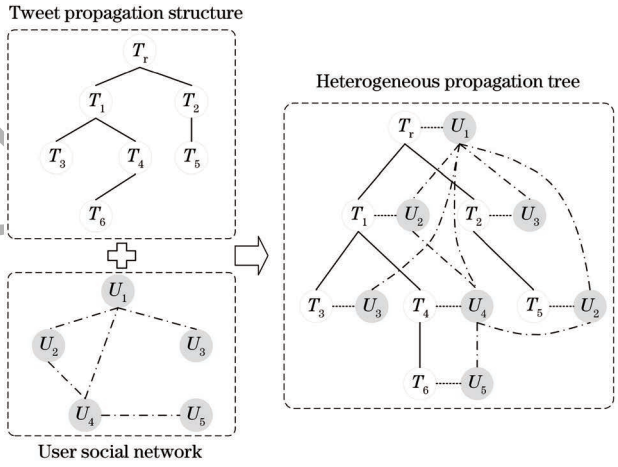

图 4 带有用户信息的异构传播树

Fig. 4 Heterogeneous propagation tree with user information

以上基于社交上下文信息的虚假信息检测方法,没有考虑用户在不同新闻事件中的全局行为特征。为了体现用户行为特征与新闻真实性之间的全局联系,文献[66]在虚假信息检测领域引入了超图的概念,以学习用户的全局传播行为特征。具体来说,文献[66]定义超图中的节点为参与当前数据集中所有新闻传播交互的用户,每条超边将参与同一新闻传播的用户连接在一起,得到表示全局新闻传播的超图。需要注意的是,这项研究中对超图的构建不包含新闻的文本信息,模型仅通过用户的传播行为特征对虚假新闻进行检测,实验结果也证明了在没有文本信息和用户信息的情况下,仅通过传播模式也能够进行有效的虚假信息检测。文献[67]从回音室效应出发捕捉用户的全局行为模式,构建以新闻作为节点的全连通图,并根据参与新闻传播的

共同用户数对图中的边进行加权，得到加权无向新闻参与度图，图中的节点度体现了相似用户群体的行为，能够帮助判断新闻的真实性。

尽管社交上下文特征提取是检测虚假文本信息的主流方法之一，但在多模态虚假信息检测中的研究相对较少。文献[68]提出的多模态特征增强注意力网络(MFAN)首次在多模态虚假信息检测任务中整合了社交上下文特征。与虚假文本信息检测不同，MFAN 需要将图像、文本和传播图作为 3 类模态信息进行多模态特征融合。MFAN 首先通过传播图隐藏链接推断增强传播图特征，并利用图像特征增强文本特征，之后引入模态对齐模块，对增强后的文本和传播图特征进行对齐优化。最后，将 3 个模态特征两两组合经过跨模态协同注意力模块进行特征融合，用于检测分类。文献[69]提出了一种新颖的生成模型检测框架，基于扩散模型生成模拟的用户交互序列，从中学习新闻的传播全局结构和时间深度特征，与新闻多模态特征拼接后进行虚假新闻检测，由于无需真实的社交上下文信息，因此提升了此类检测方法的时效性。

提升实际信息传播状态的模拟能力或扩展社交上下文信息，均能有效提升模型对虚假信息的检测能力。通过整合传播结构信息、用户信息、用户传播行为模式等社交上下文信息，可以较大程度地发挥社交媒体中用户群体智慧的作用，提升在低资源场景以及复杂信息伪造技术下模型的检测能力。但是，由于是对信息潜在的传播模式进行分析，这类方法缺乏良好的检测结果可解释性，并且由于社交上下文信息囊括的特征种类较为复杂，在实际建模表示时受限于模型的表示能力，容易产生信息损失。

3.2.2 基于多领域学习的虚假信息检测方法

实证研究发现，不同主题的社交媒体信息之间在词汇使用和传播模式等方面存在显著差异[70]，导致虚假信息检测模型在处理未见主题的信息时，检测性能易受到领域偏差影响而下降。针对此问题，研究人员将社交媒体信息按照事件主题划分为不同的领域，提出多领域学习作为虚假信息检测的辅助任务，着重学习领域间共享或领域特定的分类知识，以提高在多个领域上的模型检测泛化性能。

构造事件判别器进行对抗训练，学习虚假信息检测的领域共享知识是多领域学习的策略之一。文献[71]首次提出了针对未见主题的虚假新闻检测任务，针对不同主题的多模态新闻，提出的事件对抗神经网络利用事件主题判别器进行对抗训练，去除领域特定的特征表示，从而学习领域共享的可迁移检测特征。但是，这类通过对抗训练实现的跨领域检测能力泛化由于过于关注领域共享特征的学习，而忽视了领域特定知识的学习，限制了模型的检测性能。

研究人员设计了多种模型结构，使其能够学习虚假信息检测的领域特定和领域共享知识，提升多领域训练的有效性。文献[72]基于混合专家模型提出了一个有效的多领域虚假新闻检测框架，由多个专家网络分别提取新闻表征，并利用一个可学习的领域嵌入向量与新闻的句子级嵌入向量共同控制领域门，实现对专家网络中领域特定知识的自适应聚合。文献[73]提出了一种由一个领域特征记忆库和多个领域事件记忆库组成的领域记忆库，其中领域特征记忆库旨在自动捕获和存储领域特征，领域事件记忆库旨在存储各领域中所有的新闻特征，给出输入信息的潜在领域标签分布，尽可能贴近现实中新闻数据可能涉及多个领域标签的现象。在提取输入信息的多视角特征后，由领域记忆库强化领域信息，并通过领域适配器聚合领域特征进行虚假文本信息检测分类。文献[70]为降低多领域训练的标注成本，提出了一种无监督领域发现算法，将具有高用户相关性、标题相关性的新闻簇作为一个新闻领域，计算新闻属于各个领域的概率，将其拼接得到领域分类软标签，之后利用局部敏感哈希算法选择一组样本，人工对其进行真实性标注。

上述方法全部依赖于一个前提条件，即需要目标域检测样本的标注信息，无法满足突发事件等缺乏标注信息的场景需要，因此，部分研究利用无监督训练方法进行多领域学习，实现无监督跨域自适应。文献[74]通过特征解耦的方法强化多领域间共享知识的学习，仅利用领域共享知识实现对未见领域新闻的真实性判断，由于领域特定知识并未用于目标域样本的检测，模型需要大量源域数据进行训练以获得最优的检测性能，或当目标域与源域显著不同时，该方法能够展现出性能优势。文献[75]提出了一种基于对比学习和交叉注意力的无监督跨域谣言检测方法，将目标域数据经过聚类得到伪标签后，利用域内和域间对比学习聚类同类样本，并强化域间共享分类知识，通过交叉注意力机制对齐域间特征分布，从而将源域的领域特定知识迁移至目标域，最终结合域间共享知识和对齐后的域内特定知识实现对目标域的无监督域适应。

3.2.3 基于多模态欺骗模式的虚假信息检测方法

多模态虚假信息在创作过程中容易产生模态间

计算机工程 www.ecice06.com

相互矛盾或异常的内容，为虚假信息检测提供了可感知的模式线索。因此，越来越多的研究工作从多模态虚假信息创作过程的欺骗模式分析出发，提升检测效果。

模态间语义信息的不一致性是一类常见的多模态欺骗模式感知方式，部分研究将多模态信息统一到文本特征空间中进行表示，以此衡量模态语义不一致性。文献[76]首次将模态间语义一致性用于虚假信息检测，利用图片描述生成模型将图像模态信息统一到文本模态，在文本特征空间中评估模态间语义相似性。文献[77]提出了一种实体增强的多模态虚假新闻检测框架，将模态间语义的不一致性评估细化至实体对象，提取出文本和视觉实体后，同样统一到文本特征空间中，以衡量多模态不一致性。

其他研究将多模态特征统一映射到共享语义空间表示，以此衡量模态语义的不一致性。文献[78]引入了评论信息与新闻内容之间的语义差异评估，在得到新闻的单模态和多模态融合语义特征后，通过一致性约束推理层，从 2 个角度进行语义不一致性建模：第一个角度为评论语义和新闻单模态以及多融合语义间的差异；第二个角度为单模态特征对多模态新闻内容造成的语义偏差，强化了多模态语义不一致性的感知能力。文献[79]首先以真实样本作为训练子集，通过图文样本匹配任务引导模态间特征对齐至共享语义空间，之后利用 2 个模态特定的变分自编码器近似得到单模态特征分布之间的 KL 散度，以此作为衡量模态间语义不一致性的分值，最后利用模态间语义不一致性分值自适应控制跨模态特征和单模态特征对检测任务的贡献，当单模态特征间的不一致性较强时，强化跨模态特征的重要性。文献[80]提出的 Event-Radar 框架将模态间语义不一致性作为多视图学习的特征之一，首先通过 CLIP 提取图片和文本的多模态特征，并构建多模态图作为多模态信息的表示方式，其中，节点为文本字符特征和视觉实体特征。之后，分别提取多模态图中与文本和图片对应的事件子图，将文本节点细化为主谓宾状语信息。图像节点为与文本实体相似度最高的视觉实体特征。最后，通过 GCN 提取文本和图像子图的图特征，构造比较函数，衡量模态间事件表示的不一致性。

判断图像信息是否经过人为篡改是检测虚假信息的另一关键线索。文献[81]通过共享部分模型参数实现模态间特征对齐，引入误差水平分析算法作为图像篡改特征提取模块，突出伪造图像的恶意拼接和重压缩特性。文献[82]总结了 3 类多模态信息的欺骗模式，即图像篡改、模态间语义不一致和历史图像重放，由此提出了一种神经符号潜变量模型，各类欺骗模式分别对应一个模型分支，将每种欺骗模式表示为一个需要推断的二值可学习潜变量，若最终捕捉到一个或多个欺骗模式，则意味着检测对象为虚假信息。文献[83]提出了 4 种细粒度的多模态信息伪造类别，包括人脸替换、人脸情绪修改、文本信息替换和文本情感属性篡改，在检测输入多模态信息的真实性的同时，对信息伪造类别进行细粒度分类，并定位被篡改的图像边界框或文本内容。文献[84]在检测图像篡改特征的同时，引入篡改操作意图判断，认为如果真实信息的图像被篡改，其意图必须是无害的，如果虚假信息的图像被篡改，其意图可能是有害或无害的。将篡改痕迹检测和篡改意图检测这 2 项任务构造为包含大量正例样本和无标签样本的无监督学习场景，之后利用正样本和无标签学习方法进行任务训练。

4 基于事实验证和证据推理的虚假信息检测

基于事实验证和证据推理的虚假信息检测技术旨在引入客观证据，分析信息的逻辑准确性或判断信息真实性，以此识别虚假信息。根据知识证据的来源不同，一类研究专注于利用虚假信息检测的先验知识，引导大模型根据内部知识进行逻辑推断；另一类则通过设计事实核查框架，引入知识信息验证检测对象的事实准确性。

4.1 基于大模型内部知识推断的虚假信息检测

自大语言模型问世以来，利用大语言模型检测虚假信息的研究热度迅速上升，已成为该领域不可忽视的新范式。尽管大语言模型内部知识丰富、语义理解能力强等特点与虚假信息检测任务的需求相契合，但由于缺乏虚假信息检测任务的训练，导致难以充分发挥其潜力。因此，现有研究主要从提示工程优化、模型参数指令微调和真实性分析 3 个方面，利用大模型的内部知识进行虚假信息检测。

4.1.1 基于大语言模型提示工程的虚假信息检测方法

早期对大语言模型在虚假信息领域应用的探索，大多集中于通过提示工程测试其相关能力。文献[85]利用包括虚假信息检测在内的多个自然语言处理任务，对 ChatGPT 的性能进行测试。在不提供任何指令信息说明的情况下，直接输入待检测信

计算机工程 www.ecice06.com

息，人工对模型输出进行理解统计，得到 ChatGPT 对随机选取的 100 条新冠疫情相关信息的检测性能，取得了与此前最优方法相当的准确率。受限于没有提供任务指令信息以及测试样本数较小，文献[85]的测试结果不能完全体现大语言模型在直接提示下的零样本检测效果。在给定任务指令信息并限定输出格式后，文献[86-87]以 GPT-3.5 和 GPT-4 作为主要模型，对多个数据集进行了全面的零样本以及小样本情境学习下的检测性能测试。实验结果表明，大语言模型进行零样本和小样本虚假信息检测时，输出解释信息能提高检测性能，但是与文献[85]的检测结果不同，其准确率普遍低于此前深度神经网络的监督方法。此外，文献[88]对 ChatGPT 的虚假信息检测判决一致性实验体现出，大语言模型在不同数据集以及多轮判决中展现出的零样本检测性能不够稳定。

通过提示工程设计针对虚假信息检测的提示模板，能够有效提升其虚假信息检测能力。文献[88]采用常规思维链(CoT)[89]提示，发现大语言模型在 RAWFC 和 LIAR[15]数据集上的检测性能反而会因为忽略检测要素或模型幻觉而弱于标准提示，表明需要针对虚假信息的内在特点设计有针对性的提示方法。文献[90]针对虚假新闻中与事实性紧密相关的要素信息，即人物、地点、时间、事件，设计虚假信息检测的 CoT 提示[89]，提示大语言模型抽取相关要素并调用大语言模型内部参数化知识，引导大语言模型进行信息可信度的逻辑推理，有效提高了 GPT-4 进行虚假信息检测的性能。文献[87]利用大语言模型的语义理解和推理能力，由虚假新闻样本反推出虚假新闻的 9 项特征，基于此提出了原因感知提示方法，用于虚假信息检测。文献[91]对 ChatGPT 在虚假信息检测任务中的泛化性和不确定性问题进行深入研究，充分利用大语言模型的生成能力，将二分类任务转化为对信息真实性打分的软分类任务，并且对检测的错误样例、解释信息合理性等方面进行详细分析。与之类似，文献[92]采用的软分类提示模板要求大语言模型先输出分析过程，再进行分类打分，最终按照 JSON 格式输出结果。文献[76]对闭源大语言模型(各个版本的 GPT-3.5、GPT-4)以及开源大语言模型(LLaMA2[93]、zephyr[94]等)在类似提示方法下的检测性能进行对比，结果显示，开源大语言模型 zephyr[94]在该提示方法下表现出接近 GPT-4 的检测效果，同样具有较大应用潜力。

提示工程是提升大语言模型分析信息真实性能力的重要方法。为了测试大语言模型检测虚假信息的能力，早期研究采用较为简单的标准提示方法，使模型直接给出分类结果，在此基础上，要求给出分析过程或进行软分类，能够提升其检测性能。常用的提示工程方法包括设计任务相关思维链、提供小样本示例进行语境学习以及从虚假信息语义特征出发设置特定的分析角度等。通过提示工程，能使大语言模型进行零样本虚假信息检测，且对不同的事件主题有较好的泛化能力。尽管如此，此类方法仍然存在多轮检测判决不稳定、语义分析证据整合不准确等问题，影响其检测效果。表 2 总结了部分具有代表性的虚假信息检测提示模板。

4.1.2 基于大语言模型指令微调的虚假信息检测方法

对大语言模型虚假信息检测任务提示工程的探索基本都以闭源大语言模型 GPT-3.5 或 GPT-4 作为主要模型，面对社交媒体中的海量信息，其检测成本较高并且可能有潜在的数据隐私安全问题。出于以上考虑以及为了更好地激发大语言模型检测虚假信息，研究人员对开源大语言模型在虚假信息检测任务中的指令微调展开了探索。文献[96]分析了使用参数微调的大语言模型进行虚假新闻检测的潜力，利用专家或 ChatGPT 构建指令集，并采用低秩适配器算法(LoRA)[97]对 LLaMA2[93]进行高效参数微调。实验测试结果表明，微调后的大语言模型能够从揭露虚假信息、叙事风格、事实核查、操纵性分析、实体及情绪信息等几个角度进行文本分析，并且给出对于虚假信息的预测结论。文献[98]采用了另一种构建指令数据集的方法，将指令微调与事实核查相结合，利用搜索引擎工具检索每个检测样本的外部证据，从而构造“指令-证据-检测文本”样式的指令数据，同样利用 LoRA[97]进行高效参数微调，增强大语言模型对于虚假新闻判断逻辑的理解能力，并且通过设置指令数据集对输出内容的格式要求，提高大语言模型的指令跟随能力，便于结果统计。

指令微调效果的好坏，关键在于如何构建指令微调数据集。对于虚假信息检测任务，常用的虚假信息检测相关指令微调数据构建方法包括直接选用现有数据集、收集 GPT-4 或人类专家的新闻分析文本、收集支持样本真实性类别的证据和背景信息等，同时指令微调数据应具有严格的响应输出格式要求，以便后续进行检测和分析结果的高效统计。通过高效参数微调，在降低微调成本的同时，使大语言模型能够快速适应下游任务。

计算机工程 www.ecice06.com

表 2 大语言模型虚假信息检测提示模板

Table 2 Prompt template for disinformation detection with large language models

Prompt category	Literature	Prompt template
Binary classification criteria prompt	Literature [90]	Act as a disinformation detector to analyze the following news article [NEWS]. Does this news article contain any misleading information? Please respond with (1) an analytic process, and (2) "Yes" or "No".
	Literature [91]	Consider for yourself the truth of the following news. You should give "real" or "fake" answers in order of number and not give an "unclear" answer. You can give two answers: "real" or "fake" in only one word without giving any reasons or repeating the original text. Here is the news: …
	Literature [91]	Provide a number 0 or 1, where 0 represents false and 1 represents true. Do not provide any explanations, only respond with the number.
Three-class classification criteria prompt	Literature [91]	Provide a number 0 or 1, where 0 represents false and 1 represents true. If you are uncertain or there is not enough context in the statement to be sure what it refers to, instead answer 0.5. Do not make assumptions. Do not provide any explanations, only respond with the number.
Soft classification criteria prompt	Literature [91]	Provide a score from 0 to 100, where 0 represents definitively false and 100 represents definitively true. Do not provide any explanations, only respond with the numerical score.
Chain-of-thought prompt	Literature [91]	Your task is to provide a score from 0 to 100, where 0 represents definitively false and 100 represents definitively true, but you must not state your score until you've presented a thorough analysis. Do not begin your response with a number. First write your analysis, then write a vertical bar "|", then finally state your score.
	Literature [95]	Given a "passage", please think step by step and then determine whether or not it is a piece of misinformation. You need to output your thinking process and answer "YES" or "NO". The "passage" is: [passage].
	Literature [89]	1. Extract all the characters, place names, time stamps, and key events from the provided text: [NEWS]. 2. Assess the factualness of the extracted events. Show your analytic process. 3. Assess the relationship between all characters, place names, time stamps, and key events. Show your analytic process. 4. Based on your analysis from steps 2 and 3, does this news article contain any misleading information or mismatched relationships? Show your analytic process and respond with "Yes" or "No".

4.1.3 基于大语言模型真实性特征分析的虚假信息检测方法

文献[99]通过实验发现,大语言模型可以通过提示检测出部分虚假新闻,并提供丰富的多视角理论分析,但其检测效果仍然低于此前微调 BERT 等小语言模型的监督方法。该文分析认为,大语言模型虽然具有强大的语境学习和文本分析能力,并且蕴含大量的世界知识,但是由于其预训练过程中缺乏虚假信息检测的相关训练,因此无法正确地整合理论分析结果进行预测。基于上述实验结果和理论分析,文献[99]认为当前的大语言模型可能无法在虚假新闻检测中替代微调的小语言模型,但可以通过提供多视角的指导性理论来为小语言模型提供良好的预测基础。文献[99]提示大语言模型从文本语言风格和常识推理 2 个角度分别生成分析信息,利用交叉注意力机制对分析信息进行加权,输入小语言模型,从而联合大语言模型与小语言模型实现共同检测,取得了较好的检测性能。

为提高深度神经网络汇总大语言模型分析结果的效果和可信度,文献[100]提出了 TELLER 框架。具体来说,TELLER 由认知系统和决策系统组成。认知系统中的大语言模型根据输入信息以及问题模板,得到与信息真实性相关的逻辑判断真值;决策系统利用神经-符号模型,从算法角度实现了决策结果的可信性,能够端到端地从一组逻辑判断真值中自动学习逻辑规则,捕获逻辑判断和真实标签之间的可泛化关系。此外,文献[101]将虚假信息检测任务转化为弱监督任务,总结了 18 条专家判断信息真实性依据的可信度标志,每条可信度标志能在一定程度上反映信息的真实性。然后,设计各个可信度标志对应的提示信息,利用大语言模型得到各个可信度标志的分析表述,再由小型语言模型对所有可信度信号进行整合,得到二分类的真实性标签。

基于大语言模型真实性分析的检测方法,利用大语言模型的语义分析和逻辑推理能力,将其作为分类任务中的特征提取器或子任务标注器。最终,凭借深度神经网络,特别是预训练语言模型具有一定语义理解能力且便于参数训练的特点,通过判决式模型整合多角度真实性分析结果,将其映射为真

计算机工程 www.ecice06.com

实性标签。

4.2 基于外部知识事实核查的虚假信息检测

虚假信息检测任务的核心在于判断其陈述的事实信息是否真实，基于外部知识进行事实核查能够准确且直观地反映信息的真实性，并且由于外部知识资源丰富，此类检测方法在领域泛化性和可解释性方面具有天然优势。根据事实核查方法的不同实现方式，现有研究可分为判别式模型推理核查和大语言模型生成式事实核查两类。

4.2.1 深度神经网络判别式事实核查方法

知识图谱是事实核查中常见的知识表现形式，可以通过提取实体对以及实体间关系的三元组构建知识图谱。早期基于知识图谱的核查方法依赖人为定义的任务流程，难以处理复杂信息。文献[102]提出将核查任务转化为知识图谱的条件路径搜索问题，利用维基百科提取构建知识图谱，并通过知识图谱中实体间连接路径的节点度表示其信息量。然后，将待检测对象抽象为实体对，在知识图谱中搜索实体间信息量最大的联通路径，根据其信息量大小判断待检测对象的可信度。

为强化知识图谱中信息的表示能力，部分研究利用图神经网络提取出知识图谱中包含的专家信息，以实体特征对比或背景知识补充的方式进行事实核查。文献[103]将检测对象的上下文信息构建为文档图，并利用图神经网络提取特征与外部知识图谱的特征进行对比，结合比较特征以及新闻信息表征实现虚假新闻检测。在特征对比阶段，分别提取新闻中集合语义信息的实体表示与知识图谱中对应实体的特征表示，设计实体对比网络以提高模型对特征差异的表达能力。文献[104]采用背景知识补充的方式核查信息真实性，提取新闻文本中的实体信息，并在外部知识图谱中找到对应节点，将外部知识进行编码后，与文本语义特征相结合进行虚假信息检测。文献[105]进一步利用具有注意力机制的 LSTM 作为解码器，根据知识图谱和输入信息，利用复制机制生成自然语言解释，以提高检测结果的可信度。

上述研究实现了外部知识图谱在虚假信息检测中的应用，但这些研究通过维基百科等大规模知识库获取的知识图谱具有一定的信息滞后性，且不能保证其中包含的知识与待检测目标信息高度相关，过多不相关的知识会给检测过程带来噪声。针对知识图谱的可靠性、时效性、领域相关性等问题，文献[106]通过从训练数据中提取实体关系三元组来构建多关系知识图谱，然后应用组合图卷积网络学习相应的节点和关系嵌入，与文本语义特征相融合以检测虚假信息。

知识图谱类的外部知识，其信息形式多为实体间关系的三元组，信息表达能力有限，在面对复杂的文本内容时，将信息聚合到实体的特征表示会造成信息损失，对虚假信息检测产生不利影响。因此，找到能够作为反证的外部证据是检测虚假信息的重要途径。典型方法（如文献[107-108]的实现形式）可以总结为图 5 所示的 3 个阶段：首先，根据检测目标内容检索相关文档；之后，利用证据提取器，在相关文档中提取并排序证据相关段落；最后，根据相关证据检测虚假信息。文献[107]提出的证据感知虚假信息检测方法，通过搜索与新闻文本相关的网络文章，并利用注意力机制对文章内容进行加权，最终聚合所有关于信息来源、网络文章内容、注意力权重和网络文章可信度的特征来检测虚假信息。文献[109]改进了提取证据信息的长距离依赖建模和证据信息冗余消除能力，采用图神经网络对检测信息和证据文本进行建模（单词作为节点，边表示单词之间的共现关系）。此外，在图结构特征的学习过程中，会根据语义结构丢弃不重要的节点，减少证据文本中冗余信息产生的负面影响。

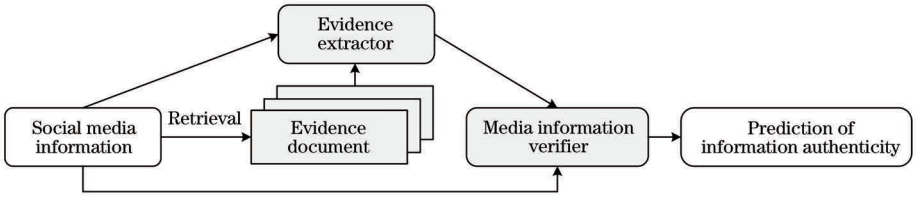

图 5 基于外部证据的虚假信息检测基本流程框架

Fig. 5 Framework for the basic process of disinformation detection based on external evidence

引入知识图谱以及证据文档是目前利用深度神经网络进行判别式事实核查的主流实现方式。基于外部知识的虚假信息检测方法相较于利用文本语义或社交上下文特征的检测方法，具有检测结果可解释性程度高、受机器人账号恶意传播影响小等优点。但是，无论是将知识图谱还是证据文档作为外部事实依据，这类方法的难点在于当可信赖的外部信息有限时，如果无法获得检测所需的反证，模型的检测

计算机工程 www.ecice06.com

能力将受到很大挑战[110]。

4.2.2 大语言模型生成式事实核查方法

大语言模型内部知识丰富、可调用工具等能力展现出了进行自动事实核查的潜力。文献[111]是大语言模型事实核查的早期工作，利用 ChatGPT 的参数化世界知识进行事实核查，验证了大语言模型进行事实核查的应用潜力。除了通过内部知识进行事实核查，文献[112]探索了大语言模型调用搜索工具，基于外部知识进行事实核查的能力，大语言模型可以根据已知信息决定是否返回最终答案，或用不同的查询内容继续检索证据。

以文献[111-112]为代表的方法直接利用大语言模型对整段信息进行事实核查，文献[88，113-114]在此基础上，进一步提出通过生成与检测任务相关的具体问题进行事实核查。文献[113]首先利用大语言模型的语境学习能力，将每条待检测信息转化为由谓词组成的一阶逻辑子句；之后，利用大语言模型将一阶逻辑子句转化为事实核查的具体问题，并调用外部工具检索更为精确的知识信息，生成对应答案；最后，整合事实核查子任务的答案，完成虚假信息检测并给出相应的解释信息。与文献[113]的思路类似，文献[88]将虚假信息检测分解为源文本拆分、分步验证和最终预测这 3 个阶段，生成检测对象的相关事实核查问题，并基于外部知识进行虚假信息检测。文献[114]在上述方法的基础上，将文本内容的事实核查扩展到大模型生成的代码、数学问题以及科学文献总结上，按“事件抽取-生成请求-工具调用-证据整理-真实性认证”的流程进行统一验证。

如图 6 所示，完整的基于大语言模型的自动事实核查方法实现步骤可以总结为 4 步，即事实声明提取、证据检索与收集、真实性判断和解释信息生成。对于长文本信息，要求从中提取出事实性相关信息，并排除常识性信息，此外还需要进行核查价值判断，优先核查具有重要影响力、对检测样本整体真实性有决定意义的声明信息。在检索收集证据信息阶段，常见的方法包括直接检索关键词和生成事实核查问题。最后进行真实性类别判断和解释信息生成，可以利用大语言模型的推理能力联合实现，将推理核查过程转化为解释信息，也可以利用提示工程等方法根据真实性判断结果生成解释信息。大语言模型的出现为构建自动事实核查系统提供了新的实现方式，目前，基于大语言模型实现的自动事实核查在一定程度上可以对专业事实核查人员起到辅助作用，缓解社交媒体信息事实核查的压力。

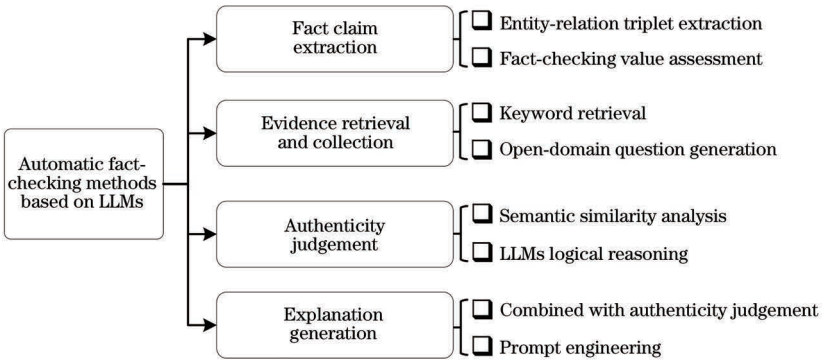

图 6 基于大语言模型的自动事实核查方法整体流程

Fig.6 The overall process of automatic fact-checking methods based on large language models

如表 3 所示，对本文梳理的 10 类虚假信息检测方法进行总结分析，其中包含各类检测方法的检测原理、优势、局限性和适用场景。

5 检测范式对比论述

5.1 虚假信息检测基本范式

在深度学习技术框架下，虚假信息检测的基本范式可以分为两类，即基于深度神经网络的虚假信息检测方法和基于大语言模型的虚假信息检测方法。

基于深度神经网络的虚假信息检测基本范式如图 7 所示，主要包含信息预处理、特征提取和特征分类 3 个部分。根据实际虚假信息检测方法的需要，对输入信息进行不同的嵌入式预处理，特征提取器利用深度神经网络提取各类输入信息的特征，训练特征分类器实现虚假信息检测。

基于大语言模型的虚假信息检测基本范式如图 8 所示，主要包含提示信息构建、模型推理和响应输出 3 个部分。得益于大语言模型强大的语义理解、逻辑推理以及调用外部知识和工具的能力，可以根据不同的检测方法设计相应的提示模板和推理分析流程，最终根据要求输出大语言模型对输入信息真实性的判断结果及解释信息。

表 3 各类虚假信息检测方法分析

Table 3 Analysis of various disinformation detection methods

Algorithm category	Category subdivision	Detection principle	Advantages	Limitations	Applicable scenarios
Semantic feature representation	Text semantic feature encoding	Learn the mapping relationship between text semantic features and labels	The detection framework is simple to train and requires minimal additional input	Relying on shallow connections between semantic features and labels results in poor accuracy and generalization ability	Detection of disinformation in domain-specific texts
	Multimodal semantic feature encoding	Learn the mapping relationship between multimodal fusion features and labels	Multimodal features can provide a comprehensive representation of contextual semantics	Modal feature interaction has high computational complexity, leading to potential information loss	Detection of disinformation in domain-specific multimodal information
Auxiliary task design	Social context feature extraction	Extract social context features such as comment information to aid the model in detection	Utilizing social context offers background knowledge and extra clues to improve detection performance	Gathering social context requires time, which makes early detection challenging	High accuracy retrospective detection of domain-specific text and multimodal disinformation
	Multidomain learning	Learn shared or domain-specific detection knowledge across domains to generalize detection performance	The ability to transfer detection capabilities to unseen event achieves high accuracy across multiple domains	It involves complex training paradigms, and requires the domain classification of detection data	High accuracy multi-domain, multimodal disinformation detection
	Multimodal deception pattern perception	Extract deceptive pattern features such as image tampering, to detect multimodal disinformation	It improves multimodal disinformation detection performance and provides some interpretability	It demands high completeness in data modality content and is easily affected by inter-modal feature bias	High accuracy multimodal disinformation detection in specific domains
Inference of internal knowledge in LLM	LLM prompt engineering	Design prompting methods to guide LLMs in fully utilizing internal knowledge for disinformation detection	Implement a simple framework that generates detection results with explanation, without parameter training	Detection results rely on input quality, cannot handle large data volumes, and require substantial time and cost	Zero-shot or few-shot detection of textual disinformation and explanation generation
	LLM instruction fine-tuning	Develop instruction data and fine-tune LLMs to master the reasoning and criteria for detecting disinformation	Improve the detection performance of open-source LLMs and format the output content as specified	Building instruction data, fine-tuning instructions, and model deployment require substantial computing resources	Multidomain textual disinformation detection and explanation generation
	Analysis of authenticity features based on LLM	Employ LLMs as authenticity feature extractors to deliver disinformation detection cues for discriminative models	Utilize small models to efficiently learn disinformation detection with attribution capabilities	Detection results depend on the output quality of LLMs, and analysis with LLMs demands significant computational resources	Multidomain textual disinformation detection
External knowledge reasoning verification	Discriminative fact-checking	Introduce external knowledge and use deep neural networks to extract representations, enabling verification through feature comparison or knowledge supplementation	Introducing external knowledge can enhance the model's detection performance and domain generalization	Detection performance depends on the accuracy of external knowledge, while the decision model lacks interpretability of the verification results	Multidomain textual disinformation fact-checking
	Generative fact-checking	Utility LLMs to generate fact-checking search entries, retrieve evidence, and produce results through reasoning and judgment	The accuracy of question-generated evidence retrieval is higher and enables complete fact checking process	Detection performance depends on the accuracy of external knowledge, judgment on the credibility of evidence is lacking	General-domain textual disinformation fact-checking

计算机工程 www.ecice06.com

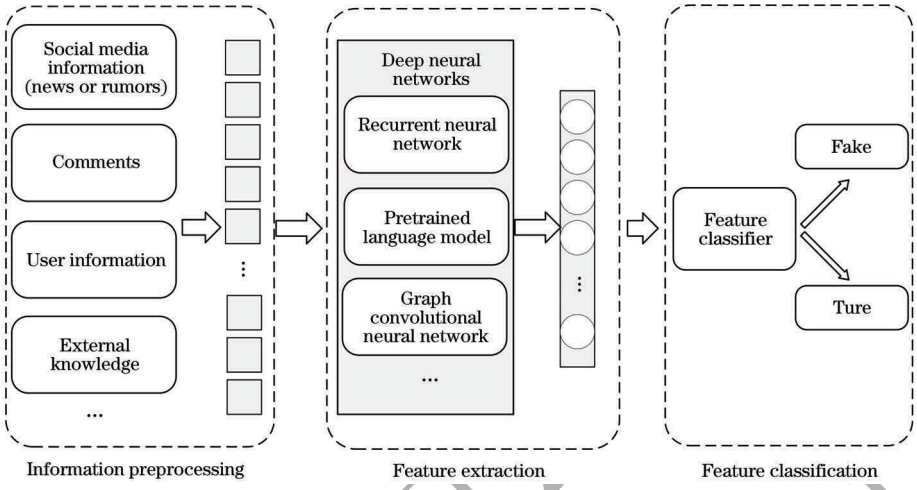

图 7 基于深度神经网络的虚假信息检测基本范式

Fig. 7 Deep neural network based disinformation detection paradigm

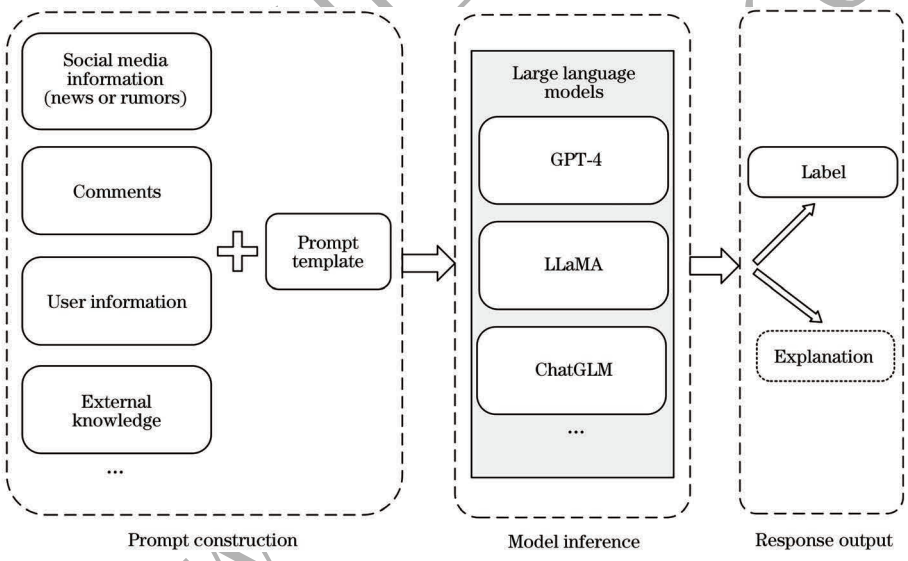

图 8 基于大语言模型的虚假信息检测基本范式

Fig. 8 Large language model based disinformation detection paradigm

5.2 检测范式性能对比分析

自 2022 年底 OpenAI 发布 ChatGPT 以来，如何利用大语言模型实现虚假信息检测的研究热度迅速上升，由此也引出了一个关键问题，即基于大语言模型的检测范式与基于深度神经网络的检测范式相比，存在什么优势和局限性？如表 4 所示，分别对两类检测范式中具有先进性能的检测方法进行了汇总，采用准确率（Acc）以及 F1 值（F1）作为评价指标。表中用于评价模型检测性能的 7 个数据集包含了大量的测试样本，在内容主题、数据结构等方面具有良好的多样性，能够较为全面地审视两类检测范式的性能差异。在虚假信息检测准确率方面，基于大语言模型的检测方法在 LIAR[15]、RAWFC[17] 和 COVID-Scientific[23] 这 3 个数据集中表现出了更优的性能，对应的检测方法包括虚假信息检测任务提示工程、指令微调以及自动事实核查。对于科学性信息验证之外的数据集，通过提示工程进行语义分析，并最终依靠大语言模型分析结果进行分类的零样本或少样本检测方法，其检测性能普遍弱于此前的有监督方法。

通过以上对比分析，可以得到基于大语言模型的虚假信息检测方法具有以下优势：

1）检测性能方面。大语言模型凭借其语义理解能力、外部工具调用能力和丰富的内部知识，对虚假信息检测任务具备一定的先验知识，在通过检索增强实现自动事实核查以及进行科学性事实信息验证方面存在一定优势。

2）检测范式方面。基于大语言模型的虚假信息检测方法在整体框架上往往更为简洁，提示信息能够直观展现检测方法的主要实现思路，并且大语言模型在多数应用场景下可以实现零样本虚假信息检测，省去了对深度神经网络进行参数训练的步骤。

表 4 常用数据集检测范式性能对比

Table 4 Comparison of performance across common dataset detection paradigms

Dataset	Detection methods based on large language model				Detection methods based on deep neural network			
	Detection method	Model	Acc/%	F1/%	Detection method	Model	Acc/%	F1/%
LIAR(Six-class)[15]	Hiss[74]	GPT-3.5	—	37.50	CofCED[17]	BiLSTM	—	28.93
LIAR(Six-class)[15]	FactLLaMA[82]	LLaMA-7B	—	30.44				
FakeNewsNet[22]	TELLER[83]	LLaMA2-7B	77.25	71.70	ExFake[115]	BERT	80.80	85.50
RAWFC[17]	HiSS[74]	GPT-3.5	—	53.90	CofCED[17]	BiLSTM	—	51.07
Twitter 15&16[14]	RA prompt[73]	ChatGPT	67.40	65.80	SBAG[116]	GNN	94.20	—
Weibo[12]	RA prompt[73]	ChatGPT	77.60	77.60	MFAN[68]	CNN,GAT	88.95	88.83
COVID-Social[23]	Standard prompt[68]	ChatGPT	73.30	—	PPL[23]	BERT	77.74	49.15
COVID-Scientific[23]	Standard prompt[68]	ChatGPT	92.00	—	PPL[23]	GPT-2	74.73	73.84

3)可解释性方面。大语言模型作为生成模型，能自然地在输出分类结果时进行决策解释，具有更优的检测结果可解释性，辅助用户进行判断。

除了上述优势，基于大语言模型的虚假信息检测方法也存在其局限性：

1)检测性能方面。大语言模型虽然能够对文本信息进行多视角的语义分析，但是欠缺特征整合能力，导致仅通过分析语义内容实现的零样本以及少样本虚假信息检测准确率未达到预期效果。

2)数据处理方面。由于大语言模型的输入和输出均为文本信息，且受到最大输入字符数的限制，难以处理以图结构出现的社交媒体信息，例如，RA 提示[73]舍弃了 Twitter 15&16 数据集[14]中的评论信息，导致检测性能显著低于此前的监督方法。此外，对于多模态数据，利用视觉大语言模型进行多模态虚假信息检测的研究目前依然较为欠缺。

3)检测成本方面。现有基于大语言模型实现的虚假信息检测方法大多依赖于模型整体性能更强的闭源大语言模型，如 ChatGPT、GPT-4 等，经济成本较高，并且即便依赖开源大语言模型构建虚假信息检测，其运行和微调所需的硬件资源以及检测所需的时间成本也高于以往的深度神经网络。

6 发展方向及风险挑战

目前基于大语言模型的虚假信息检测方法已经得到了研究人员的广泛关注，可以预见，大语言模型的进一步应用发展将会给虚假信息检测领域带来深远影响。因此，本节将主要围绕大语言模型的特性和应用，探讨虚假信息检测领域未来可能的发展方向以及存在的风险挑战。

6.1 发展方向

限制当前虚假信息检测系统以及自动事实核查系统进行实际应用的主要因素有系统领域泛化性差、决策过程可信度低等。大语言模型展现出的上下文学习、语义生成等能力为上述问题提供了新的解决方案，同时也为模拟和干预虚假信息的传播提供了新的发展方向。

1)自动事实核查框架。事实核查作为最直接的虚假信息检测方法，目前主要依赖人工实现，自动事实核查的难点在于从复杂文本中抽取出事实性陈述信息，并准确地从知识库中检索出反事实证据进行对比核查。利用大语言模型驱动多智能体框架，并结合检索增强技术进行自动事实核查，其逻辑推理、调用外部工具的能力在提高事实核查效率和准确性中起到了重要作用。

2)可信多语言通用领域虚假信息检测系统。事实核查系统难以应对缺乏外部证据的虚假信息早期检测、非公众事件虚假信息检测等应用场景，通过语义分析和逻辑推理实现虚假信息检测能够有效弥补这一缺陷。现有基于大语言模型的虚假信息检测方法已经能够通过语义分析和逻辑推理实现多领域事件的零样本虚假信息检测，并且通过对分类结果进行解释，使其具有较好的可信度。未来进一步强化大语言模型对于小语种信息的语义理解能力，提高其内部世界知识的准确性、及时性，设计推理判断流程，为实现可信多语言通用领域虚假信息检测提供可行解决方案。

3)干预信息生成。将虚假信息检测结果的解释信息与事实核查相结合，依靠大语言模型的自然语言生成能力，能够以贴近人类写作风格的方式生成防止虚假信息进一步传播的干预信息，提示用户当前信息可能存在的真实性风险。

4)动态社交网络环境模拟。收集多个社交媒体用户基本信息，将其作为大语言模型驱动的智能体画像，规定社交媒体中常见的用户行为，如评论、转发等，并学习各个用户的历史行为，从而利用多个智

计算机工程 www.ccjc06.com

能体实现社交网络环境模拟，将动态变化的社交网络环境作为虚假信息检测的实验背景，能够更好地评估虚假信息的传播特点、可能造成的影响以及虚假信息检测方法的实际效果。

6.2 风险挑战

尽管大语言模型能够对虚假信息检测相关的研究领域带来积极影响，但是大语言模型存在的模型偏见、安全性对齐、响应幻觉等问题容易加剧高迷惑性虚假信息的产生和传播，给虚假信息检测领域带来新的风险挑战。

1)恶意生成虚假信息。目前对于恶意利用大语言模型的防范机制还不够完善，通过设计提示工程方法，能够利用大语言模型恶意生成极具迷惑性的虚假信息，并且除文本信息外，多模态大模型已经具备生成高质量图像、音频甚至视频的能力，将其与社交媒体机器人技术相结合，能够快速产生并传播大量虚假信息。文献[87,90,95]利用提示工程要求大语言模型生成虚假信息，发现已有检测方法并不能进行有效检测，对虚假信息检测方法提出了新的挑战。

2)大模型幻觉干扰。幻觉问题是目前大语言模型所面临的一个主要问题，对于虚假信息检测任务，幻觉问题可能会影响大语言模型对检测信息的真实性判断，同时社交媒体用户也可能因为大语言模型幻觉问题，无意间生成并发布错误信息。

7 结束语

社交媒体虚假信息检测是信息安全领域的重要课题。本文对虚假信息检测技术进行了详细阐述，将信息内容特征与社交网络特征作为间接检测依据，将事实验证和证据推理作为直接检测依据，论述了不同技术实现路径如何获取上述依据，并分析了各类检测实现路径的特点和适用场景。此外，本文首次对大语言模型在虚假信息检测领域的应用发展进行了总结梳理，对比论述了基于大语言模型的检测方法相比基于深度神经网络的检测方法存在的优势和局限性。在未来发展中，可以进一步研究大语言模型在虚假信息检测领域更加深入的应用方向，提高检测方法的时效性、泛化性和可解释性，同时也需要防范大语言模型可能给该领域带来的风险挑战。

参考文献

[1] Weibo reports third quarter 2023 unaudited financial results [EB/OL]. [2024-07-05]. http://ir.weibo.com/static-fil es/a3170224-9394-4b70-a0d1-ef89d2bfcb75.

[2] 张志勇，荆军昌，李斐，等. 人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述[J]. 计算机学报，2021，44(11)：2261-2282.
ZHANG Z Y，JING J C，LI F，et al. Survey on fake information detection，propagation and control in online social networks from the perspective of artificial intelligence[J]. Chinese Journal of Computers，2021，44(11)：2261-2282.（in Chinese）

[3] 聂大成，汪明达，刘世钰，等. 在线社会网络虚假信息检测关键技术研究综述[J]. 通信技术，2023，56(4)：391-399.
NIE D C，WANG M D，LIU S Y，et al. Survey on key technologies of fake information detection in online social networks[J]. Communications Technology，2023，56(4)：391-399.（in Chinese）

[4] ALAM F，CRESCI S，CHAKRABORTY T，et al. A survey on multimodal disinformation detection[C]// Proceedings of the 29th International Conference on Computational Linguistics. New York，USA：ACM Press，2022：6625-6643.

[5] GHARAIBEH M，OBEIDAT R，ABDULLAH M，et al. Datasets and approaches of COVID-19 misinformation detection：a survey[C]// Proceedings of the 13th International Conference on Information and Communication Systems（ICICS）. Washington D. C.，USA：IEEE Press，2022：337-345.

[6] HANGLOO S，ARORA B. Fake news detection tools and methods—a review[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2112.11185.

[7] 毛震东，赵博文，白嘉萌，等. 基于传播意图特征的虚假新闻检测方法综述[J]. 信号处理，2022，38(6)：1155-1169.
MAO Z D，ZHAO B W，BAI J M，et al. Summary of false news detection methods based on communication intention characteristics[J]. Journal of Signal Processing，2022，38(6)：1155-1169.（in Chinese）

[8] GONG S Z，SINNOTT R O，QI J Z，et al. Fake news detection through graph-based neural networks：a survey[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2307.

[9] HAMELEERS M. Disinformation as a context-bound phenomenon：toward a conceptual clarification integrating actors，intentions and techniques of creation and dissemination[J]. Communication Theory，2023，33(1)：1-10.

[10] United Nations General Assembly. Countering disinformation for the promotion and protection of human rights and fundamental freedoms[EB/OL]. [2024-07-05]. https://undocs.org/en/A/77/287.

[11] 陈燕方，李志宇，梁循，等. 在线社会网络谣言检测综述[J]. 计算机学报，2018，41(7)：1648-1677.
CHEN Y F，LI Z Y，LIANG X，et al. Summary of online social network rumor detection[J]. Chinese Journal of Computers，2018，41(7)：1648-1677.（in Chinese）

[12] MA J，GAO W，MITRA P，et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto，USA：AAAI Press，2016：3818-3824.

[13] 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.

[14] MA J，GAO W，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. Stroudsburg，PA，USA：ACL，2017：708-717.

[15] WANG W Y. “Liar，liar pants on fire”：a new benchmark dataset for fake news detection[C]//Proceedings of the 55th

计算机工程 www.ecicos.com

Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2017: 422-426.
[16] ALHINDI T, PETRIDIS S, MURESAN S. Where is your evidence: improving fact-checking by justification modeling[C]// Proceedings of the 1st Workshop on Fact Extraction and VERification (FEVER). Stroudsburg, PA, USA: ACL, 2018: 85-90.
[17] YANG Z W, MA J, CHEN H C, et al. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2209.14642.
[18] SANTIA G, WILLIAMS J. BuzzFace: a news veracity dataset with Facebook user commentary and egos[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2018, 12(1): 531-540.
[19] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. FEVER: a large-scale dataset for fact extraction and verification[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2018: 809-819.
[20] AHMED H, TRAORE I, SAAD S. Detecting opinion spams and fake news using text classification[J]. Security and Privacy, 2018, 1(1): e9.
[21] ABU SALEM F K, AL FEEL R, ELBASSUONI S, et al. FA-KES: a fake news dataset around the Syrian war[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2019, 13(2): 573-582.
[22] SHU K, MAHUDESWARAN D, WANG S H, et al. FakeNewsNet: a data repository with news content, social context, and spatio-temporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171-188.
[23] LEE N, BANG Y J, MADOTTO A, et al. Towards few-shot fact-checking via perplexity[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 1971-1981.
[24] LIN H Z, MA J, CHEN L L, et al. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 2543-2556.
[25] KHAN K, WANG R Z, POUPART P. WatClaimCheck: a new dataset for claim entailment and inference[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 1293-1304.
[26] HU X M, GUO Z J, WU G Y, et al. CHEF: a pilot Chinese dataset for evidence-based fact-checking[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 3362-3376.
[27] DEMENTIEVA D, KUIMOV M, PANCHENKO A. Multiverse: multilingual evidence for fake news detection[J]. Journal of Imaging, 2023, 9(4): 77.
[28] BARBARO F, SKUMANICH A. Addressing socially destructive disinformation on the web with advanced AI tools: Russia as a case study[C]//Proceedings of the ACM Web Conference 2023. New York, USA: ACM Press, 2023: 204-207.
[29] SALEHINEJAD H, SANKAR S, BARFETT J, et al. Recent advances in recurrent neural networks[EB/OL]. [2024-07-05]. https://arxiv.org/abs/1801.01078.
[30] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[31] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]//Proceedings of the NIPS 2014 Workshop on Deep Learning. New York, USA: PMLR Press, 2014: 15-22.
[32] CHEN Y X, SUI J, HU L, et al. Attention-residual network with CNN for rumor detection[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2019: 1121-1130.
[33] VAIBHAV V, MANDYAM R, HOVY E. Do sentence interactions matter? leveraging sentence level representations for fake news classification[C]//Proceedings of the 13th Workshop on Graph-Based Methods for Natural Language Processing. Stroudsburg, PA, USA: ACL, 2019: 134-139.
[34] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2024-07-05]. https://arxiv.org/abs/1710.10903.
[35] WU L W, RAO Y, JIN H L, et al. Different absorption from the same sharing: sifted multi-task learning for fake news detection[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL, 2019: 4643-4652.
[36] LEE N, LI B Z, WANG S N, et al. On unifying misinformation detection[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 5479-5485.
[37] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2019: 4171-4186.
[38] KALIYAR R K, GOSWAMI A, NARANG P. FakeBERT: fake news detection in social media with a BERT-based deep learning approach[J]. Multimedia Tools and Applications, 2021, 80(8): 11765-11788.
[39] PELRINE K, DANOVITCH J, RABBANY R. The surprising performance of simple baselines for misinformation detection[C]// Proceedings of the Web Conference 2021. New York, USA: ACM Press, 2021: 3432-3441.
[40] LIN H Z, YI P Y, MA J, et al. Zero-shot rumor detection with propagation structure via prompt learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(4): 5213-5221.
[41] QIAO Y P, WIECHMANN D, KERZ E. A language-based approach to fake news detection through interpretable features and BRNN[C]// Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media. New York, USA: ACM Press, 2020: 14-31.
[42] STRöBEL M, KERZ E, WIECHMANN D, et al. CoCoGen-complexity contour generator: automatic assessment of linguistic complexity using a sliding-window technique[C]//Proceedings of Workshop on Computational Linguistics for Linguistic Complexity. Washington D. C., USA: IEEE Press, 2016: 23-31.
[43] AJAO O, BHOWMIK D, ZARGARI S. Sentiment aware fake news detection on online social networks[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2019: 2507-2511.
[44] SHENG Q, CAO J, ZHANG X Y, et al. Zoom out and observe: news environment perception for fake news

计算机工程 www.ecice06.com

detection[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 4543-4556.
[45] WU X Q, HUANG K H, FUNG Y, et al. Cross-document misinformation detection based on event graph reasoning[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2022: 543-558.
[46] LAI T, JI H, BUI T, et al. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution [C] // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 3491-3499.
[47] 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. New York, USA: ACM Press, 2017: 795-816.
[48] SINGHAL S, SHAH R R, CHAKRABORTY T, et al. SpotFake: a multi-modal framework for fake news detection[C]//Proceedings of the IEEE 5th International Conference on Multimedia Big Data (BigMM). Washington D.C., USA: IEEE Press, 2019: 39-47.
[49] 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.
[50] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]// Proceedings of the Neural Information Processing Systems. Washington D.C., USA: IEEE Press, 2019: 5754-5764.
[51] WU Y, ZHAN P W, ZHANG Y J, et al. Multimodal fusion with co-attention networks for fake news detection [C]// Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 2560-2569.
[52] QIAN S S, WANG J G, 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. New York, USA: ACM Press, 2021: 153-162.
[53] ZHOU Y M, YANG Y Z, YING Q C, et al. Multimodal fake news detection via CLIP-guided learning [C] // Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Washington D.C., USA: IEEE Press, 2023: 2825-2830.
[54] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [EB/OL]. [2024-07-05]. https://arxiv.org/abs/2103.00020.
[55] 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, USA: ACM Press, 2019: 2915-2921.
[56] WEI Z M, PAN H Y, QIAO L B, et al. Cross-modal knowledge distillation in multi-modal fake news detection[C]// Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D.C., USA: IEEE Press, 2022: 4733-4737.
[57] YING Q C, HU X X, ZHOU Y M, et al. Bootstrapping multi-view representations for fake news detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(4): 5384-5392.
[58] 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 AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018: 354-361.
[59] MA J, GAO W, WONG K F. Rumor detection on Twitter with tree-structured recursive neural networks [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2018: 1980-1989.
[60] BIAN T, XIAO X, XU T Y, 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.
[61] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2024-07-05]. https://arxiv.org/abs/1609.02907.
[62] LIN H Z, MA J, CHENG M F, et al. Rumor detection on Twitter with claim-guided hierarchical graph attention networks [C] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 10035-10047.
[63] KHOO L M S, CHIEU H L, QIAN Z, et al. Interpretable rumor detection in microblogs by attending to user interactions[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8783-8790.
[64] CUI J, KIM K, NA S H, et al. Meta-path-based fake news detection leveraging multi-level social context information[C]// Proceedings of the 30th International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2021: 325-334.
[65] MIN E, RONG Y, BIAN Y, et al. Divide-and-conquer: post-user interaction network for fake news detection on social media [C] // Proceedings of ACM Web Conference 2022. New York, USA: ACM Press, 2022: 1148-1158.
[66] 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 AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 5248-5256.
[67] WU J, HOOI B. DECOR: degree-corrected social graph refinement for fake news detection[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2023: 2582-2593.
[68] ZHENG J Q, ZHANG X, GUO S C, et al. MFAN: multi-modal feature-enhanced attention networks for rumor detection[C]//Proceedings of the 21st International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2022: 2413-2419.
[69] ZHANG L T, ZHANG X M, LI C Z, et al. Mitigating social hazards: early detection of fake news via diffusion-guided propagation path generation[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York, USA: ACM Press, 2024: 2842-2851.
[70] SILVA A, LUO L, KARUNASEKERA S, et al. Embracing domain differences in fake news: cross-domain fake news detection using multi-modal data[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 557-565.
[71] WANG Y Q, MA F L, JIN Z W, 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, USA: ACM Press, 2018: 849-857.
[72] NAN Q, CAO J, ZHU Y C, et al. MDFEND: multi-domain fake news detection[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York, USA: ACM Press, 2021: 3343-3347.

计算机工程 www.ecice06.com

[73] ZHU Y C, SHENG Q, CAO J, et al. Memory-guided multi-view multi-domain fake news detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 7178-7191.
[74] WU D K, TAN Z H, ZHAO H R, et al. LIMFA: label-irrelevant multi-domain feature alignment-based fake news detection for unseen domain[J]. Neural Computing and Applications, 2024, 36(10): 5197-5215.
[75] RAN H Y, JIA C Y. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(11): 13510-13518.
[76] ZHOU X Y, WU J D, ZAFARANI R. SAFE: similarity-aware multi-modal fake news detection[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2003.04981.
[77] QI P, CAO J, LI X R, 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, USA: ACM Press, 2021: 1212-1220.
[78] WU L W, LIU P S, ZHANG Y N. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(11): 13736-13744.
[79] CHEN Y X, LI D S, ZHANG P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference 2022. New York, USA: ACM Press, 2022: 2897-2905.
[80] MA Z H, LUO M N, GUO H, et al. Event-radar: event-driven multi-view learning for multimodal fake news detection[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2024: 5809-5821.
[81] XUE J X, WANG Y B, TIAN Y C, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5): 102610.
[82] DONG Y Q, HE D X, WANG X B, et al. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(8): 8354-8362.
[83] SHAO R, WU T X, WU J L, et al. Detecting and grounding multi-modal media manipulation and beyond[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5556-5574.
[84] WANG B, WANG S S, LI C C, et al. Harmfully manipulated images matter in multimodal misinformation detection[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2407.19192.
[85] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[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. Stroudsburg, PA, USA: ACL, 2023: 675-718.
[86] XIAO Y, ALAM F. Nexus at ArAIEval shared task: fine-tuning arabic language models for propaganda and disinformation detection[C]//Proceedings of Arabic Natural Language Processing Conference 2023. Washington D. C., USA: IEEE Press, 2023: 576-582.
[87] CARAMANCION K M. Harnessing the power of ChatGPT to decimate mis/disinformation: using ChatGPT for fake news detection[C]//Proceedings of the IEEE World AI IoT Congress (AIIoT). Washington D. C., USA: IEEE Press, 2023: 42-46.
[88] 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. Stroudsburg, PA, USA: ACL, 2023: 996-1011.
[89] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2024: 15-22.
[90] JIANG B H, TAN Z, NIRMAL A, et al. Disinformation detection: an evolving challenge in the age of LLMs[EB/OL]. [2024-07-05]. https://observatory.informationdemocracy.org/resources/disinformation-detection-an-evolving-challenge-in-the-age-of-llms/.
[91] PELRINE K, IMOUZA A, THIBAULT C, et al. Towards reliable misinformation mitigation: generalization, uncertainty, and GPT-4[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023: 6399-6429.
[92] VERGHO T, GODBOUT J F, RABBANY R, et al. Comparing GPT-4 and open-source language models in misinformation mitigation[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2401.06920.
[93] TOUVRON H, MARTIN L, STONE K, et al. LLaMA2: open foundation and fine-tuned chat models[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2307.09288.
[94] TUNSTALL L, BEECHING E, LAMBERT N, et al. Zephyr: direct distillation of LM alignment[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2310.16944.
[95] CHEN C, SHU K. Can LLM-generated misinformation be detected[C]//Proceedings of the 12th International Conference on Learning Representations. Washington D. C., USA: IEEE Press, 2024: 12-23.
[96] PAVLYSHENKO B M. Analysis of disinformation and fake news detection using fine-tuned large language model[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2309.04704.
[97] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2106.09685.
[98] CHEUNG T H, MAN K. FactLLaMA: optimizing instruction-following language models with external knowledge for automated fact-checking[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2309.00240.
[99] HU B Z, SHENG Q, CAO J, 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.
[100] LIU H, WANG W Y, LI H R, et al. TELLER: a trustworthy framework for explainable, generalizable and controllable fake news detection[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg, PA, USA: ACL, 2024: 15556-15583.
[101] LEITE J A, RAZUVAYEVSKAYA O, BONTCHEVA K, et al. Weakly supervised veracity classification with LLM-predicted credibility signals[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2309.07601.
[102] CIAMPAGLIA G L, SHIRALKAR P, ROCHA L M, et al. Computational fact checking from knowledge networks[J]. PLoS One, 2015, 10(6): e0128193.
[103] HU L, YANG T, ZHANG L, et al. Compare to the knowledge: graph neural fake news detection with external

计算机工程 www.ecice06.com

knowledge[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 754-763.
[104] DUN Y Q, TU K F, CHEN C, et al. KAN: knowledge-aware attention network for fake news detection[EB/OL]. [2024-07-05]. https://cdn.aaai.org/ojs/16080/16080-13-19574-1-2-20210518.pdf.
[105] VEDULA N, PARTHASARATHY S. FACE-KEG: FACT checking explained using KnowledgE Graphs[EB/OL]. [2024-07-05]. https://dl.acm.org/doi/10.1145/3437963.3441828.
[106] WU K, YUAN X, NING Y. Incorporating relational knowledge in explainable fake news detection[EB/OL]. [2024-07-05]. https://link.springer.com/chapter/10.1007/978-3-030-75768-7_32.
[107] CHEN J F, SRIRAM A, CHOI E, et al. Generating literal and implied subquestions to fact-check complex claims[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 3495-3516.
[108] MA J, GAO W, JOTY S, et al. Sentence-level evidence embedding for claim verification with hierarchical attention networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2019: 2561-2571.
[109] XU W Z, WU J F, LIU Q, et al. Evidence-aware fake news detection with graph neural networks[C]//Proceedings of the ACM Web Conference 2022. New York, USA: ACM Press, 2022: 2501-2510.
[110] GLOCKNER M, HOU Y F, GUREVYCH I. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2022: 5916-5936.
[111] HOES E, ALTAY S, BERMEO J. Leveraging ChatGPT for efficient fact-checking[EB/OL]. [2024-07-05]. https://europepmc.org/article/PPR/PPR640906.
[112] QUELLE D, BOVET A. The perils and promises of fact-checking with large language models[J]. Frontiers in Artificial Intelligence, 2024, 7: 1341697.
[113] WANG H R, SHU K. Explainable claim verification via knowledge-grounded reasoning with large language models[C]// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA, USA: ACL, 2023: 6288-6304.
[114] CHERN I C, CHERN S, CHEN S Q, et al. FacTool: factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios[EB/OL]. [2024-07-05]. https://arxiv.org/abs/2307.13528.
[115] AMRI S, BOLEILANGA H M, AIMEUR E. ExFake: towards an explainable fake news detection based on content and social context information[C]// Proceedings of the Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE). Washington D. C., USA: IEEE Press, 2024: 1-8.
[116] HUANG Z, LV Z, HAN X, et al. Social bot-aware graph neural network for early rumor detection[C]//Proceedings of the 29th International Conference on Computational Linguistics. New York, USA: ACM Press, 2022: 6680-6690.

文字编辑 吴云芳
栏目编辑 赖玉玲

计算机工程 www.ecice06.com