·计算机技术与应用·

基于在线社交网络的虚假新闻检测方法综述

杨思佳[1], 李显勇[1,2*], 杜亚军[1]
(1. 西华大学计算机与软件工程学院, 四川 成都 610039; 2. 宜宾维特瑞安科技有限公司, 四川 宜宾 644600)

摘　要：在线社交网络承载着人们日益增长的社会交流需求, 是现代重要的信息共享途径。随着社交网络的普及, 在线社交网络也逐渐成为了虚假新闻滋生的温床。虚假新闻的危害程度之深、范围之广, 使得虚假新闻检测成为自然语言处理领域的热点研究问题之一。文章首先介绍了虚假新闻的产生背景、检测意义、相关概念及问题描述, 其次, 在数据收集和特征处理过程的基础上, 依据特征与新闻内容之间的关系, 围绕基于内部特征和基于外部特征的虚假新闻检测方法进行综述梳理, 最后, 阐述了虚假新闻检测任务仍然面临的一系列研究挑战, 并对未来的研究方向进行了系统性的展望。

关键词：在线社交网络; 虚假新闻检测; 深度学习; 特征提取
中图分类号：TP393.09　　**文献标志码**：A　　**文章编号**：1673–159X(2025)00 − 0001 − 11
doi:10.12198/j.issn.1673 − 159X.5276

A Survey on Fake News Detection Based on Online Social Networks

YANG Sijia[1], LI Xianyong[1,2*], DU Yajun[1]
(1. *School of Computer and Software Engineering*, *Xihua University*, *Chengdu* 610039 *China*;
2. *Yibin Weite Ruian Technology Co.*, *Ltd.*, *Yibin* 644600 *China*)

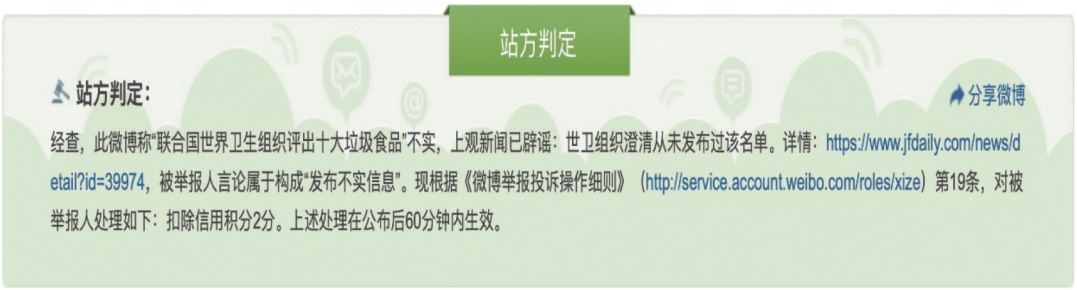
Abstract: Online social networks have become increasingly popular as a means of social communication and sharing information in the digital era. The growing demand for social interaction has resulted in the spread of fake news on social media platforms. The depth and wide scope of the harm of fake news have made its detection one of the hot research topics in natural language processing. This paper first introduces the background, significance, related terms, and problem description of fake news detection. It then explains the process of data collection and feature extraction. Considering the relationship between features and news content, this paper reviews and categorizes fake news detection methods based on internal and external features. Finally, this paper highlights the research challenges that still exist in fake news detection and presents a systematic outlook on future research directions.

Keywords: online social networks; fake news detection; deep learning; feature extraction

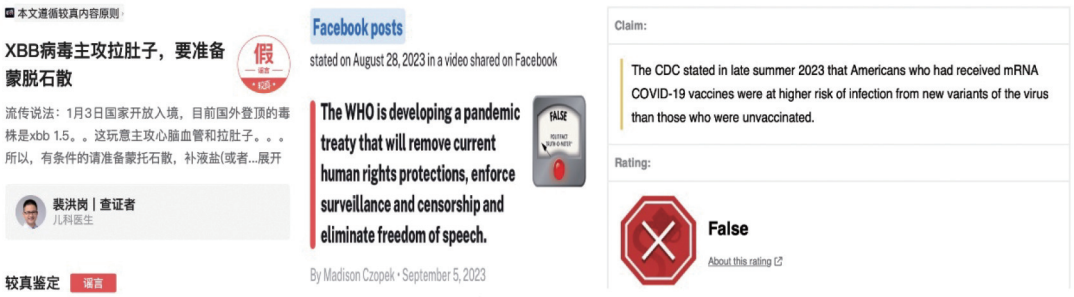
收稿日期:2024 − 04 − 20
基金项目:国家自然科学基金(61802316); 四川省科技计划资助(2022YFG0378, 2023YFS0424, 2023YFH0058, 2023YFQ0044); 宜宾市科技计划重点研发项目 (2023SF004)。
*** 通信作者**: 李显勇 (1984—), 男, 教授, 博士, 硕士生导师, 主要研究方向为人工智能, 社交网络分析, 网络舆情演化与引导等。
ORCID: 0000 − 0003 − 0097 − 1643　　E-mail: lixy@mail.xhu.edu.cn

引用格式:杨思佳, 李显勇, 杜亚军. 基于在线社交网络的虚假新闻检测方法综述[J]. 西华大学学报(自然科学版), 2025, 44(X): 1 − 11.
YANG Sijia, LI Xianyong, DU Yajun. A Survey on Fake News Detection Based on Online Social Networks[J]. Journal of Xihua University(Natural Science Edition), 2025, 44(X): 1 − 11.

随着互联网技术的蓬勃发展,在线社交网络(online social network, OSN)广泛普及,已逐渐成为人们接收、分享和交流信息不可或缺的平台。然而,广泛传播的新闻资讯质量良莠不齐,这不仅对于新闻媒体公信力存在威胁,同时也增加了公众对于信息可信度的判断难度。虚假新闻的传播与扩散对不同领域都造成了重大的不良影响[1-3]。虚假新闻一直是世界各地持续关注的热点问题之一。为了即时制止虚假新闻对个人、组织和社会形成的危害进一步蔓延,各国媒体平台积极成立辟谣中心,通过各种形式接收举报信息,收集判定证据。图 1 展示了 4 个国内外辟谣平台样例。然而这些辟谣平台依靠着很多人力成本,虽然可信度很高,但当面对铺天盖地的不实信息,检测人员往往力不从心。因此,对自动虚假新闻检测技术投入研究是十分必要的,有助于节省收集证据和判别的成本,也有利于尽早缩小传播范围以减少更多物质与精神上的损失。



(a) 微博举报处理大厅



(b) 腾讯较真辟谣平台 (c) PolitiFact (d) Snopes

图 1 国内外辟谣平台

Fig. 1 Domestic and foreign rumor refuting platform

本文将从虚假新闻的相关概念及问题描述、数据收集及特征处理、虚假新闻检测方法概述和研究挑战及未来展望等 4 部分对虚假新闻检测任务进行综述梳理。

1 相关概念及问题描述

1.1 相关概念

在线社交网络中的信息真假混杂,研究者对真假信息及更细粒度的分类进行了深入研究和探讨。虽然目前尚未有统一的定义,本文结合目前已有文献对相关术语的定义[4-7],将关系结构梳理如图 2 所示,并给出如下定义。

定义 1(虚假新闻) 虚假新闻(fake news)是指主观故意编造并已被证实为错误的新闻文章。

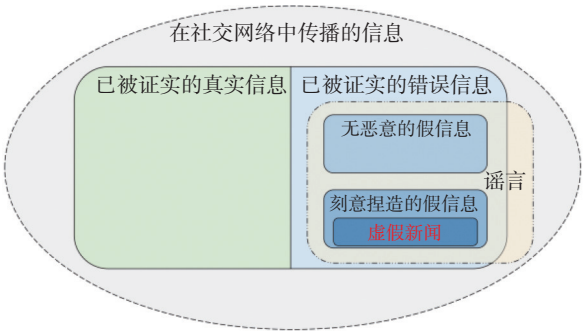


图 2 相关术语关系示意图

Fig. 2 A Schematic representation of the relationship of the related terms

1.2 问题描述

基于虚假新闻定义,虚假新闻检测问题可定义为有监督的二分类任务。

定义 2(虚假新闻检测) 给定在线社交网络

中的一则新闻帖 p 及标签集合 $T=\{0(\text{真实}),1(\text{虚假})\}$，通过结合运用从新闻帖 p 中抽取到的内部特征集合 $\mathcal{X}_{\text{in}}=\{\mathcal{X}_{\text{in}}^{(1)},\mathcal{X}_{\text{in}}^{(2)},\cdots,\mathcal{X}_{\text{in}}^{(m)}\}$ 和收集到的外部特征集合 $\mathcal{X}_{\text{ex}}=\{\mathcal{X}_{\text{ex}}^{(1)},\mathcal{X}_{\text{ex}}^{(2)},\cdots,\mathcal{X}_{\text{ex}}^{(n)}\}$，学习到一个分类器 F，使得 $(p,\mathcal{X}_{\text{in}},\mathcal{X}_{\text{ex}})\xrightarrow{F}T$。

2　数据收集及特征处理

2.1　数据收集

虚假新闻数据的主要获取途径[7]有：社交媒体平台 API、网络爬虫和公开数据集。

社交媒体平台 API 是官方提供的数据接口，授权之后可以获取具体的用户信息，例如新浪微博 API[8] 提供相关用户的粉丝服务、评论转发和地理位置等信息接口。虽然官方平台提供的 API 接口快捷可靠，但有时可能无法满足个性化的科研需求，运用网络爬虫技术可以拓展研究者们的数据挖掘范围。结合 API 接口和爬虫技术，现有研究工作整理开源了公开数据集，如表 1 所示。

表 1　虚假新闻检测数据集

Tab. 1　Data sets for fake news detection

名称	语言	内部特征	外部特征	新闻总数	假新闻数
Weibo-16[9]	中文	文本	用户信息、传播网络	4664	2313
Weibo-20[10]	中文	文本	用户信息、转发回复、传播网络	6362	3161
Weibo-21[11]	中文	文本、视觉	转发回复	9128	4488
FakeSV[12]	中文	文本、视觉、音频	用户信息、转发回复、传播网络	3654	1827
Twitter-16[13]	英文	文本	用户信息、转发回复	818	205
FakeNewsNet[14]	英文	文本、视觉	用户信息、转发回复、传播网络	23196	5755
PHEME[15]	中文、英文	文本、视觉	用户信息、转发回复	5802	1972
MM-COVID[16]	中文、英文	文本、视觉	用户信息、转发回复、传播网络	11565	3981

2.2　特征处理

虚假新闻自动检测需要从收集到的数据中筛选出有效的特征，并转化生成为计算机可以理解的特征表示。根据特征与新闻内容之间的关系，本文将用于虚假新闻检测任务中的特征划分为内部特征和外部特征。

1）内部特征：指描述新闻主体内容的特征，包括文本信息和视觉信息等。

2）外部特征：指从外部环境中收集到的与给定新闻有关的特征，包括社交上下文属性（用户信息、转发、回复和传播网络）和外部知识等。

3　虚假新闻检测方法概述

虚假新闻检测任务主要旨在寻找自动化的方法以检验待测新闻样本的真实性。相似任务包括“谣言检测”“事实核查”“信息可信度检验”等。这些任务在研究方法上存在共通之处，本文不做严格区分。总体来看，已有检测方法可以大致分为两类：基于内部特征的虚假新闻检测方法和基于外部特征的虚假新闻检测方法。

3.1　基于内部特征的虚假新闻检测方法

在早期新闻传播阶段，可以利用的外部社交信息较少，因此对新闻内容本身的核查成为了检测新闻真实性的研究重点。新闻内容是新闻样本的主要组成部分，从新闻主体内容中抽取的文本、视觉信息等内部特征可以为虚假新闻检测提供多种角度的信息。

3.1.1　文本信息

1）语言风格。为了获取流量和关注，虚假新闻在语言风格上往往与真实新闻存在差异。例如，Yang 等[17]根据写作指南抽取了 8 种高级语言特征，包括新闻的可读程度、可靠程度、交互程度、感性程度、逻辑程度、正式程度、有趣程度和完整程度等，并结合“新闻流行度”指标对人民日报、央视新闻、新华网和新华视点的新闻进行了实验分析，证实了不同的写作风格对新闻质量存在很大的影响。此外，在新闻的不同政治倾向方面，Potthast

等[18]通过实验验证了针对语言风格的分析可以将主流媒体新体与具有极端党派倾向和具有讽刺意味的新闻区分开来。基于新闻的不同写作风格，Przybyla 等[19]构建了风格型虚假新闻检测模型并验证了该类模型可以捕捉到虚假新闻使用的独特语言特征，包括耸人听闻的(sensational)词语和具有情感表达的(affective)词语等。

2)情感信号。已有很多研究工作证实，对于检测新闻真实性，情感信号是一种重要的判别模式。例如，Giachanou 等[20]分别采用了基于词典、基于情感强度和基于神经网络的 3 种方法从新闻源帖中生成情感信号表征，生成的情感信号表征可以提升新闻真实性检测效果。Ajao 等[21]指出虚假新闻文本与其传达的情感之间存在独特关联关系，并证明了将情感信号作为辅助检测特征的可能性和有效性。Zhang 等[10]提出了“双重情感”(dual emotion)概念，包括新闻内容本身包含的发布者情感(publisher emotion)和新闻评论中包含的社会情感(social emotion)，并将其与两者“情感差”(emotion gap)作为 3 种新的情感特征进一步提升了虚假新闻检测效果。

3.1.2 多媒体信息

虽然文本信息在新闻内容中占有很大比重，但是多媒体信息也是不容忽视的重要特征。图片、音频和视频等多媒体信息往往会更加吸引眼球，帮助扩大新闻的影响力和传播范围。

在数据集方面，以往都集中在文本或图片特征上，很少有研究者关注音频、视频形式的虚假新闻，且很难同时提供有关新闻内容、用户评论和用户资料的信息。Qi 等[12]从抖音和快手平台收集并构建了目前最大的中文短视频数据集 FakeSV，其中包括覆盖文本、音频和视频形式的新闻内容信息(视频、封面图像、标题、出版时间)、用户评论信息(点赞/收藏/评论数量、前 100 条评论)和发布者简介信息(身份验证、个人简介、当前 IP 位置、粉丝/订阅/喜欢/视频数量、前 100 个发布的视频覆盖)等。

在模型选择方面，早期研究工作通过特征工程方式统计新闻附带的多媒体信息，并对其进行分析。例如，Jin 等[22]提出了视觉清晰度得分、连贯性得分、相似性分布直方图、多样性得分和聚类得分等 5 种视觉特征，以及从出现频次、评论转发和长宽比等不同角度考量的 7 种图像统计特征，揭示了虚假新闻在视觉和统计 2 方面有着与真实新闻不同的分布模式。然而设计精巧的特征工程方法有其局限性，难以推广。随着深度学习技术的发展，研究者探索了利用深度卷积神经网络(convolutional neural network)如 VGG19 结构[23]，以及卷积神经网络和循环神经网络(recurrent neural network)相结合[24]的方法结合视觉特征辅助虚假新闻检测。此外，音频、视频等更加丰富的多媒体信息也逐渐受到了研究者关注[12, 25－26]。为对输入新闻视频的不同特征进行建模，Qi 等[12]首先提取包括文本、音频、关键帧、视频片段、评论和用户等多种模态特征，并使用 2 个跨模态转换网络来分别模拟文本、音频和关键帧之间的相关性，以及进一步动态融合新闻内容和社交上下文等多模态特征以检测虚假新闻，并在 FakeSV 数据集验证了有效性。

3.1.3 多模态融合

有时 OSN 中的新闻会同时携带文字和图片等多种模态的信息。相较于真实新闻而言，虚假新闻存在“模态语义不一致”问题[27－28]，即新闻声明中的文字信息可能是根据图片恶意杜撰的，导致两者描述的内容有偏差甚至冲突。多模态混合方法主要集中于研究多模态互补、多模态一致性和多模态增强等 3 种不同策略[29]。多模态互补指将不同模态的特征表示进行拼接融合，将视觉信息看作文本特征的补充。多模态一致性指核验文本与视觉所描述的信息内容是否相关且一致，通过计算两者的相似度进行度量。多模态增强指通过对比观察文本和视觉两者的对齐(alignment)部分双向增强。3 种策略虽然侧重点不同，但是仍然可以结合使用。当视觉信息与文本信息描述一致时，可以考虑采用多模态互补策略，检查是否存在可以利用的补充信息；当视觉信息与文本信息描述不一致时，可以采用多模态一致性策略，检查 2 种模态存在的冲突。多模态增强策略从两者充分交互的角度出发，寻找文本和视觉信息存在的共性特征。

3.2 基于外部特征的虚假新闻检测方法

除了新闻内容本身外,还可以从外部环境中收集与给定新闻相关的社交上下文属性和外部知识等外部特征,这些辅助信息可以对虚假新闻检测起到重要的补充作用。

3.2.1　社交上下文属性

随着虚假新闻在社交媒体中广泛传播,其周围产生的社交上下文呈现出与真实新闻不同的特征趋势[29－31]。

1)特征选择。从社交媒体平台上用户的社交参与行为中可以提取出额外的社交上下文特征。这些特征也反映着随着时间推移的新闻扩散过程,整合了其中的社交参与行为动态,为判断新闻真实性提供了有价值的辅助信息[5]。目前研究者利用到的社交上下文特征主要可以分为基于用户档案[32－35]、人群反馈[36－39] 和网际关系[40－47] 等。

基于用户档案的社交上下文特征主要包括用户名称、自我描述、是否验证、地理位置、创建时间、注册年龄、关注数量、粉丝数量、互关数量、发帖数量等个体级别特征[32－33],以及与新闻贴相关所有用户的平均画像等群体级别特征[5]。

基于人群反馈的社交上下文特征主要包括评论情感极性、人群立场分析和时序序列关系等。情感极性[10] 是社交媒体帖子中的一种重要特征,可以用来分析用户对新闻的态度和情绪。例如,评论中带有的高兴、愤怒、疑惑和否认等情感都有助于分析人群对新闻的看法和态度。人群立场分析[36] 则通过聚合多个用户在新闻事件中的支持、中立和反对等立场来得到更全面的评估结果。时序序列关系特征[39] 考虑了新闻贴的时间变化。从随时间推移捕获的新闻贴特征变化,可以帮助理解新闻在社交媒体上的传播过程和人群反应的变化趋势,为虚假新闻检测提供更丰富的上下文信息。

基于网际关系的社交上下文特征主要包括网络组成和网络互动两方面。根据群体关系的不同,网络组成可以分为立场网络、共现网络、好友网络和扩散网络等[5]。立场网络是由与新闻相关的所有推文节点以及立场相似性权重边来构建的。共现网络刻画了用户是否编写了与新闻原帖相关的新闻贴或评论,共同参与指定新闻事件的讨论。好友网络指与新闻原帖相关的用户关注或被关注的网络结构,反应了群体间的社交关系。扩散网络可以跟踪新闻传播轨迹,其节点表示用户,边表示它们之间的信息扩散路径和传播过程。构建网络后,分析新闻贴、用户和评论之间词级别、句级别、帖子级别和事件级别[40－47] 的网络互动关系有助于捕捉不同粒度的与待测新闻贴有关的社交信息,用以作为辅助虚假新闻检测的依据。

这些社交上下文特征在新闻传播过程中自然产生,为检测虚假新闻及了解用户偏好提供了重要支持。除了用于虚假新闻的检测任务,社交上下文特征也可以被转化于缓解虚假新闻产生的不利影响。因此,一个具有前景的研究方向是设计一个将虚假新闻检测及传播抑制任务相结合的统一框架。Wang 等[48] 通过构建个性化新闻推荐策略证实了构建统一框架想法的可能性。

2)模型构建。利用社交上下文特征的虚假新闻检测模型通常围绕面向用户立场和传播模式展开[35]。

面向用户立场的虚假新闻检测模型是利用用户对新闻的观点或意见来推断新闻的真实性[31]。这些观点可以是显式的,如直接表达的情绪或意见,也可以是隐式的,需要从社交媒体帖子中自动提取[5]。一般的方法是结合立场检测和主题模型方法,从事件主题中学习潜在立场,再根据相关帖子的立场值,推断新闻的真实性。Ma 等[36] 进一步证实采用统一的神经多任务学习框架将立场检测和新闻真实性检测等多任务进行联合优化,有利于多任务之间的信息共享和特征表示强化。

面向传播模式的虚假新闻检测模型主要分为基于序列结构、基于树形结构和基于图结构三种。基于序列结构的方法将新闻传播过程中一系列用户评论建模为序列结构,并利用循环神经网络[9]、卷积神经网络[39] 和 Transformer 模型[41] 来捕获序列上的高级特征交互。为了更好地理解和捕捉新闻传播相关的结构信息和层级结构,研究者试图将新闻传播过程建模为树形结构。例如,Ma 等[49] 发现针对新闻传播路径方向进行建模有助于挖掘潜在的社交上下文特征,并提出了一个基于自上而下和自下而上路径的树形递归神经网络,实现了新闻内容和传播信息之间的有效连接。然而,上述方

法对于学习传播结构特征的效率较低,也忽略了谣言分散的全局结构特征。与之相比,基于图结构的图神经网络模型能够更好地捕获全局结构特征。Bian 等[50] 引入双向图卷积网络,进一步揭示了自上而下路径会携带基于深度的传播信息,而自下而上路径会携带基于广度的扩散信息。然而,当面对蓄意捏造的错误信息和刻意编排的评论对话时,上述传统方法呈现脆弱性。为了解决这类问题,研究者探索了对抗学习和对比学习在虚假新闻检测任务的潜力[51 − 52]。

3.2.2 外部知识

除了收集与新闻有关的社交上下文属性外,引入额外的事实证据作为外部知识也是检测新闻真实性的有效方案。

1)知识收集。虚假新闻检测中可以使用的知识来源主要包括搜索引擎、知识图谱和大语言模型等。

搜索引擎作为获取事件相关描述的信息检索工具,具有实效性和全面性等特点。在虚假新闻检测任务中,研究者使用搜索引擎为待测新闻文本收集了大量的相关文章,并探索了新闻原帖、新闻原帖来源、相关文章、相关文章来源等在检测新闻真实性中的作用[53 − 58]。

此外,还可以利用或者构建知识图谱作为虚假新闻检测的外部知识来源。知识图谱主要由节点(真实世界中的实体)和边(实体间的关系)组成,可以从语义知识层面在实体间建立联系。例如,Pan 等[59] 通过从新闻中构建知识图谱,并应用 TransE[60] 及其变体模型学习到更有效的知识表示以检测新闻真实性。值得注意的是,使用知识图谱的检测方法在很大程度上受限于图谱的构建质量。

随着算力和数据的提升,大语言模型也成为了可以带来额外补充知识的工具。爆炸式增长的语料数据量和模型参数量,使得通用大语言模型学习到了海量的知识,这已经足够在广泛任务上取得优异效果。虽然针对具体的虚假新闻检测任务,大语言模型尚不能独立超过之前最佳模型,其掌握的通用知识仍然可以为虚假新闻检测带来额外知识,提供有建设意义的视角补充。

2)知识融合。虚假新闻检测中的知识融合方法主要分为基于注意力机制和基于图神经网络的方法。

由于具有并行化程度高、擅长处理序列数据等优点,注意力机制已经被广泛用于自然语言处理领域,如文本匹配、文本分类、机器翻译和问答任务等。在虚假新闻检测领域中,注意力机制可以被用来建模新闻原帖和相关文章等文本之间的关系[53]。在此基础上,Vo 等[57] 提出了一个基于词级和文章级的分层多头注意力网络,能够区分重要的词级和文章级信息。此外,注意力机制还可以被用来建模新闻原帖和知识图谱中实体及其上下文之间的关系。例如,Dun 等[61] 提出知识感知注意网络,在其中引入了“新闻−实体”和“新闻−实体−实体上下文”2 种注意力机制,既更合理有效地将知识整合到新闻中,又有助于确定实体和实体上下文的相对重要性。

除了注意力机制,图神经网络也可以有效地建模文本实体间的关系。全连通图可以在大型文档中保留长期依赖关系,这种优势可以使图神经网络有效捕捉句子之间的交互关系。基于此,Vaibhav 等[62] 将虚假新闻检测任务重新定义为图分类任务,并使用 2 种图神经网络,即图卷积神经网络(graph convolutional network, GCN)和图注意力网络(graph attention network, GAT),验证了句间交互在不同类型的文章中存在差异。此外,由于图结构中节点类型的丰富性,Li 等[63] 使用异构图神经网络对树型汇话图和星型证据图 2 种拓扑结构进行节点状态更新。每个节点通过不断聚合其邻居的隐藏状态,并将其与自己的状态相结合,并生成新的隐藏状态。这个过程使节点能够获得更加丰富的信息,从而实现信息的增量和更新。

3)知识对比。虽然知识融合方法能够将外部知识整合到新闻中,从而提高模型的新闻理解能力,但是虚假新闻中的内容经常与外部知识相矛盾。因此,除了知识融合,还需要更加关注将外部知识与原新闻文本进行对比,以便更准确地识别和评估新闻的真实性。为了缩小错误信息的指定范围,Wu 等[55] 提出了一个可解释的证据推断模型,旨在动态捕捉核心语义冲突。Nie 等[56] 构建了具有三阶段(文档检索阶段、句子选择阶段、文本验证

阶段)的同质神经语义匹配网络,并验证了页面浏览频率(pageview frequency)和知识本体词网(ontological wordNet)是 2 种有效的外部知识特征。Hu 等[64] 提出了一个端到端的异构图神经网络,将新闻与知识库中的实体进行直接比较,由于主题信息和虚假新闻检测任务高度相关[4,65],该模型还将学习目标扩展到生成带有主题信息增强的新闻文章表示。

4)其他辅助策略。为了扩大验证的检索规模以及方便用户使用,Botnevik 等[58] 开发了一个浏览器插件,可以在线实时使用搜索引擎验证文本真实性,为在线实时检测提供了便利。此外,结合不同视角也可以为检测虚假新闻提供增益,例如 Sheng 等[66] 融合新闻风格视角和事实描述视角,Wu 等[67] 融合传播特征视角和事实特征视角,对新闻事件进行更加综合全面的分析。在此基础上,Hu 等[68] 发现使用大语言模型可以为虚假新闻检测提供视角提示和决策引导,帮助已有精调模型进一步提升性能。

4　研究挑战及未来展望

4.1　研究挑战

1)即时检测问题。为了尽可能缩小虚假新闻危害的影响范围,在虚假新闻检测任务中即时性是一个重要考量指标。如何在较短时间内挖掘出隐藏在有限信息背后的丰富内涵和关联内容是设计出一个兼顾实效性与准确性的虚假新闻检测模型的重难点。

2)数据获取问题。为了保护用户隐私,部分用户画像和偏好等个人信息的获取是有难度的。在数据利用方面,捕捉并分析有利于模型检测的不侵害用户权益的社交属性特征,并以此构建出较为全面的事件信息与用户群体之间的关系网络仍在存在挑战。

3)类不平衡问题。为了让检测模型拥有更好的训练效果,在人为构建的数据集中会尽量避免真假新闻数量相差过大。在真实场景中,虚假新闻的数量远低于真实新闻,例如约 1∶300[69]。因此,这样的类不平衡性导致离线测试与在线测试的准确率之间可能存在差距。

4)旧谣新传问题。在广泛传播的虚假新闻中,有一部分是针对过往已经被判定为虚假的新闻进行了改造与翻新。但由于存在描述方式和语义逻辑等方面的变化,模型针对同一事件可能会给出不同的判定结果。因此,旧谣新传对于模型鲁棒性而言是一种考验。

4.2　未来展望

1)面向突发事件的即时虚假新闻检测方法。在突发事件发生时,准确、及时的信息对于公众的安全和决策至关重要,而虚假新闻可能会引发恐慌、误导公众或干扰应急响应。因此,面向突发事件的即时假新闻检测方法具有极高的实用价值。然而突发事件一般缺少可用的标注数据,且可能涉及到的领域和事件十分广泛,难以提前预知。而领域间及事件间存在的相似模式可以成为虚假新闻检测的突破线索,结合跨领域、跨事件的迁移学习、少样本学习和大语言模型辅助决策的方法将有助于即时检测突发事件中的虚假新闻。

2)多模态和可解释的虚假新闻检测方法。利用包括如文本、图像、音频和视频等多种模态混合的综合性信息对虚假新闻进行检测,并从中提取具有可解释性的证据对检测结果进一步阐述,可以增强检测方法的透明度和可信度。跨模态语义对齐、模态间内容冲突、自适应模态融合等问题的解决将有助于模型更好地理解和运用多模态信息提升检测性能。此外,提升检测模型的可解释性也是进一步值得探索的研究方向。比如通过结合语义相似性方法,查找出与已有检测案例相似的真实和虚假新闻案例进行对比,帮助用户理解模型的判断依据。

3)多语言和跨文化的虚假新闻检测方法。同一个虚假新闻事件有时会跨越语言界限,传播到不同的国家和地区。然而由于语言使用频率和范围不同,模型的训练语料通常集中在英语、中文等,提升模型在不同语言上的泛化性具有实际应用价值。同时,不同文化背景也会对虚假新闻检测存在影响。如相同的表情、手势和符号在不同文化中可能会有不同的含义,如果不能正确理解,将会影响模型对其情感极性的判断。因此有必要重视不同文化背景下的语言习惯、表达方式和价值观等差异因素,从多语言和跨文化角度出发,更准确地

识别虚假新闻。

4)多任务联合的虚假新闻检测与干预方法。虚假新闻治理是一个综合的复杂研究课题,检测是其中一个环节。其他相关任务包括:事件抽取、情感分析、事实核查、传播预测和缓解干预等。设计从预防到检测再到治理的多任务联合虚假新闻治理框架是值得探索的研究方向。例如,未来可以考虑将虚假新闻检测和缓解干预等任务融合进统一框架中,结合个性化的新闻推荐系统,在检测到可疑的虚假新闻时,即时推送具有针对性的相关主题下的真实新闻或科普知识,唤起用户的警惕意识,制定针对特定群体或个人的终端定制虚假新闻防治系统。

5 总结

本文从虚假新闻检测任务的背景和定义出发,介绍了虚假新闻检测的意义、内涵和目的,还围绕新闻传播过程中产生和收集得到的内部和外部特征,从不同角度对近年来现有虚假新闻检测方法进行归纳梳理,最后总结了虚假新闻检测领域中仍然存在的研究挑战,并对未来研究方向进行了展望。

参 考 文 献

[1] PETRATOS P N. Misinformation, disinformation, and fake news: Cyber risks to business[J]. Business Horizons, 2021, 64(6): 763 − 774.

[2] GRINBERG N, JOSEPH K, FRIEDLAND L, et al. Fake news on Twitter during the 2016 U. S. presidential election[J]. Science, 2019, 363(6425): 374 − 378.

[3] ROCHA Y M, DE MOURA G A, DESIDÉRIO G A, et al. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review[J]. Journal of Public Health, 2023, 31(7): 1007 − 1016.

[4] LAZER D M J, BAUM M A, BENKLER Y, et al. The science of fake news[J]. Science, 2018, 359(6380): 1094 − 1096.

[5] SHU K, SLIVA A, WANG S H, et al. Fake news detection on social media[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22− 36.

[6] GUO B, DING Y S, YAO L N, et al. The future of false information detection on social media: new perspectives and trends[J]. ACM Computing Surveys, 53(4): 68.

[7] 高玉君, 梁刚, 蒋方婷, 等. 社会网络谣言检测综述[J]. 电子学报, 2020, 48(7): 1421 − 1435.

GAO Y J, LIANG G, JIANG F T, et al. Social network rumor detection: a survey[J]. Acta Electronica Sinica, 2020, 48(7): 1421 − 1435.

[8] 微博开放平台 API 文档[EB/OL]. (2020-12-28)[2023-10-12].https://open.weibo.com/wiki/API.

Weibo open platform API document[EB/OL]. (2020-12-28) [2023-10-12].https://open.weibo.com/wiki/API.

[9] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 3818 – 3824.

[10] ZHANG X Y, CAO J, LI X R, et al. Mining dual emotion for fake news detection[C]//Proceedings of the Web Conference 2021. Ljubljana, Slovenia: ACM, 2021: 3465 – 3476.

[11] 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. Virtual Event: ACM, 2021: 3343 – 3347.

[12] QI P, BU Y Y, CAO J, et al. FakeSV: a multimodal benchmark with rich social context for fake news detection on short video platforms[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(12): 14444 − 14452.

[13] 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.Vancouver: Association for Computational Linguistics, 2017: 708 − 717.

[14] SHU K, MAHUDESWARAN D, WANG S H, et al. FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171 − 188.

[15] ZUBIAGA A, LIAKATA M, PROCTER R. Exploiting context for rumour detection in social media[C]//CIAMPAGLIA G, MASHHADI A, YASSERI T. International Conference on Social Informatics. Cham: Springer, 2017: 109 − 123.

[16] LI Y C, JIANG B H, SHU K, et al. MM-COVID: a multilingual and multimodal data repository for combat-

ing COVID-19 disinformation[EB/OL]. 2020: arXiv: 2011.04088. http://arxiv.org/abs/2011.04088.

[17] YANG Y T, CAO J, LU M Y, et al. How to write high-quality news on social network? predicting news quality by mining writing style[EB/OL]. 2019: arXiv: 1902.00750. http://arxiv.org/abs/1902.00750.

[18] POTTHAST M, KIESEL J, REINARTZ K, et al. A stylometric inquiry into hyperpartisan and fake news[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018, 1: 231 − 240.

[19] PRZYBYLA P. Capturing the style of fake news[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 490 − 497.

[20] GIACHANOU A, ROSSO P, CRESTANI F. Leveraging emotional signals for credibility detection[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019: 877 – 880.

[21] AJAO O, BHOWMIK D, ZARGARI S. Sentiment aware fake news detection on online social networks[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 2507 − 2511.

[22] JIN Z W, CAO J, ZHANG Y D, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598 − 608.

[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//International Conference on Learning Representations. [S. l.]: Computational and Biological Learning Society, 2015.

[24] QI P, CAO J, YANG T Y, et al. Exploiting multi-domain visual information for fake news detection[C]// 2019 IEEE International Conference on Data Mining (ICDM). Beijing: IEEE, 2019: 518 − 527.

[25] QI P, ZHAO Y Y, SHEN Y F, et al. Two heads are better than one: improving fake news video detection by correlating with neighbors[C]//Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics, 2023: 11947 – 11959.

[26] YANG Z, PANG Y C, LI X H, et al. Topic audiolization: a model for rumor detection inspired by lie detection technology[J]. Information Processing & Management, 2024, 61(1): 103563.

[27] 金志威, 曹娟, 王博, 等. 融合多模态特征的社会多媒体谣言检测技术研究[J]. 南京信息工程大学学报(自然科学版), 2017, 9(6): 583 − 592.

JIN Z W, CAO J, WANG B, et al. Rumor detection on social media with multimodal feature fusion[J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2017, 9(6): 583 − 592.

[28] 刘华玲, 陈尚辉, 曹世杰, 等. 基于多模态学习的虚假新闻检测研究[J]. 计算机科学与探索, 2023, 17(9): 2015 − 2029.

LIU H L, CHEN S H, CAO S J, et al. Survey of fake news detection with multi-model learning[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(9): 2015 − 2029.

[29] HU L M, WEI S Q, ZHAO Z W, et al. Deep learning for fake news detection: a comprehensive survey[J]. AI Open, 2022, 3: 133 − 155.

[30] GLENSKI M, WENINGER T, VOLKOVA S. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly?[J]. IEEE Transactions on Computational Social Systems, 2018, 5(4): 1071 − 1082.

[31] JIN Z W, CAO J, ZHANG Y D, et al. News verification by exploiting conflicting social viewpoints in microblogs[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix: ACM, 2016: 2972 – 2978.

[32] CHEN X Q, ZHOU F, TRAJCEVSKI G, et al. Multi-view learning with distinguishable feature fusion for rumor detection[J]. Knowledge-Based Systems, 2022, 240: 108085.

[33] 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 31st ACM International Conference on Information & Knowledge Management. Atlanta: ACM, 2022: 325 – 334.

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

[35] SHU K, WANG S H, LIU H. Beyond news contents: the role of social context for fake news detection[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. Melbourne: ACM, 2019: 312 – 320.

[36] MA J, GAO W, WONG K F. Detect rumor and stance jointly by neural multi-task learning[C]//Proceedings of the The Web Conference 2018. [S. l.]: ACM, 2018: 585 – 593.

[37] SHU K, CUI L M, WANG S H, et al. dEFEND: Explainable fake news detection[C]//Proceedings of the 25th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining. Anchorage: Association for Computing Machinery, 2019: 395 − 405.

[38] 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]. 2022: arXiv: 2209.14642. http://arxiv.org/abs/2209.14642.

[39] LIU Y, WU Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 354 − 364.

[40] GUO H, CAO J, ZHANG Y Z, et al. Rumor detection with hierarchical social attention network[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino: ACM, 2018: 943 – 951.

[41] 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.

[42] LIU B, SUN X G, MENG Q, et al. Nowhere to hide: online rumor detection based on retweeting graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 4887 − 4898.

[43] 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. Stroudsburg: Association for Computational Linguistics, 2020: 505 − 514.

[44] NGUYEN V H, SUGIYAMA K, NAKOV P, et al. FANG: leveraging social context for fake news detection using graph representation[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Virtual Event Ireland: ACM, 2020: 1165 − 1174.

[45] REN Y X, ZHANG J W. Fake news detection on news-oriented heterogeneous information networks through hierarchical graph attention[C]//2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen: IEEE, 2021: 1 − 8.

[46] RUCHANSKY N, SEO S, LIU Y. CSI: a hybrid deep model for fake news detection[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017: 797 – 806.

[47] TU K F, CHEN C, HOU C Y, et al. Rumor2vec: a rumor detection framework with joint text and propagation structure representation learning[J]. Information Sciences, 2021, 560: 137 − 151.

[48] WANG S J, XU X F, ZHANG X Z, et al. Veracity-aware and event-driven personalized news recommendation for fake news mitigation[C]//Proceedings of the ACM Web Conference. Virtual Event: ACM, 2022: 3673 – 3684.

[49] 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. Melbourne: Association for Computational Linguistics, 2018: 1980 − 1989.

[50] 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.

[51] SUN T N, QIAN Z, DONG S J, et al. Rumor detection on social media with graph adversarial contrastive learning[C]//Proceedings of the ACM Web Conference 2022. Virtual Event: ACM, 2022: 2789 – 2797.

[52] LIN H Z, MA J, CHEN L L, et al. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning[C]//Findings of the Association for Computational Linguistics: NAACL 2022. Seattle: Association for Computational Linguistics, 2022: 2543 – 2556.

[53] POPAT K, MUKHERJEE S, YATES A, et al. DeClarE: Debunking fake news and false claims using evidence-aware deep learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 22 – 32.

[54] 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. Florence: Association for Computational Linguistics, 2019: 2561 – 2571.

[55] WU L W, RAO Y, SUN L, et al. Evidence infer-

ence networks for interpretable claim verification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14058 − 14066.

[56] NIE Y X, CHEN H N, BANSAL M. Combining fact extraction and verification with neural semantic matching networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6859 − 6866.

[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: Stroudsburg: Association for Computational Linguistics, 2021: 965 – 975.

[58] BOTNEVIK B, SAKARIASSEN E, SETTY V.BRENDA: browser extension for fake news detection[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event : ACM, 2020: 2117 – 2120.

[59] PAN J Z, PAVLOVA S, LI C X, et al. Content based fake news detection using knowledge graphs[C]//International Semantic Web Conference. Cham: Springer, 2018: 669 – 683.

[60] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. Lake Tahoe: ACM, 2013: 2787 – 2795.

[61] DUN Y Q, TU K F, CHEN C, et al. KAN: knowledge-aware attention network for fake news detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 81 − 89.

[62] VAIBHAV V, MANDYAM R, HOVY E. Do sentence interactions matter? leveraging sentence level representations for fake news classification[C]//Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13). Hong Kong: Association for Computational Linguistics, 2019: 134 – 139.

[63] LI J W, NI S W, KAO H Y. Meet the truth: leverage objective facts and subjective views for interpretable rumor detection[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 705 – 715.

[64] HU L M, YANG T C, ZHANG L H, 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. Stroudsburg: Association for Computational Linguistics, 2021: 754 –763.

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

[66] SHENG Q, ZHANG X Y, CAO J, et al. Integrating pattern- and fact-based fake news detection via model preference learning[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Virtual Event: ACM, 2021: 1640–1650.

[67] WU Y, SUN J H, YUAN X, et al. Dual-channel early rumor detection based on factual evidence[J]. Expert Systems with Applications, 2024, 238: 121928.

[68] HU B Z, SHENG Q, CAO J, et al. Bad actor, good advisor: exploring the role of large language models in fake news detection[EB/OL]. 2023: arXiv: 2309.12247. http://arxiv.org/abs/2309.12247.

[69] 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.

（责任编辑：饶莉）