

网络信息生态系统中的虚假信息:检测、缓解与挑战

Amrita Bhattacharjee¹ 舒 凯² 高 旻³ 刘 欢¹

¹(亚利桑那州立大学计算机科学与工程系 美国亚利桑那州 坦佩 85281)

²(伊利诺伊理工大学计算机科学系 美国伊利诺伊州 芝加哥 60616)

³(重庆大学大数据与软件学院 重庆 400044)

(abhattach43@asu.edu)

Disinformation in the Online Information Ecosystem: Detection, Mitigation and Challenges *

Amrita Bhattacharjee¹, Shu Kai², Gao Min³, and Liu Huan¹

¹(Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, USA 85281)

²(Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA 60616)

³(School of Big Data and Software Engineering, Chongqing University, Chongqing 400044)

Abstract With the rapid increase in access to the internet and the subsequent growth in the population of social media users, the quality of information posted, disseminated, and consumed via these platforms is an issue of growing concern. A large fraction of the common public turn to social media platforms and, in general, the internet for news and even information regarding highly concerning issues such as COVID-19 symptoms and treatments. Given that the online information ecosystem is extremely noisy, fraught with misinformation and disinformation, and often contaminated by malicious agents spreading propaganda, identifying genuine and good quality information from disinformation is a challenging task for humans. In this regard, there is a significant amount of ongoing research in the directions of disinformation detection and mitigation. In this survey, we discuss the online disinformation problem, focusing on the recent "infodemic" in the wake of the coronavirus pandemic. We then proceed to discuss the inherent challenges in disinformation research, including data collection, early detection and effective mitigation, fact-checking based approaches, multi-modality approaches, and policy issues and fairness, and elaborate on the interdisciplinary approaches towards the detection and mitigation of disinformation, after a short overview of the various directions explored in computational detection and mitigation efforts.

Key words disinformation; disinformation mitigation; disinformation detection; social media; interdisciplinary approaches

摘 要 随着互联网的迅速发展及网络社交媒体中用户的增加,通过社交媒体发布和传播信息的真实性和质量受到日益广泛的关注.目前大部分公众已习惯从社交媒体平台与互联网获取新闻,甚至是获取受到高度关注的话题(如新冠病毒感染症状)的信息.鉴于网络信息生态系统非常嘈杂,充斥着错误和虚假信息并经常受到恶意媒介的污染,从中识别真实的信息成为一项艰巨任务.对此,研究者们已开始

收稿日期:2020-11-25;修回日期:2020-12-18

通信作者:高旻(gaomin@cqu.edu.cn)

* The English version of this article is at <https://arxiv.org/abs/2010.09113>

致力于虚假信息检测和减缓虚假信息传播影响方面的工作.讨论了网络信息生态系统中的虚假信息问题,特别是随着新冠病毒大爆发而来的“信息疫情”.随后,简述了虚假信息检测方法,分析了减缓虚假信息影响的方法,并探讨了虚假信息研究中的固有挑战.最后从跨学科角度阐述了检测和减缓虚假信息影响的方法和未来研究展望.

关键词 虚假信息;缓解虚假信息;虚假信息检测;社交媒体;跨学科方法

中图法分类号 TP391

在过去的几十年里,随着计算硬件的发展,人工智能研究突飞猛进,智能方法被应用到医学、环境技术、农业等不同领域.利用人工智能解决社会经济、政治甚至文化问题具有巨大潜力,因此学术界和工业界的专家聚集在一起,形成谷歌 AI^①、微软^②、国际电信联盟(ITU)^③等组织机构并启动人工智能项目.

在人工智能应用中,信息素质(information literacy)领域受到日益重视,尤其是在社交媒体检测和缓解虚假信息^[1].根据国际电信联盟的报告^[2],2019 年互联网用户已达 41.3 亿,占全球人口总数的 50%以上;在社交媒体平台中,脸书(Facebook)是使用最广泛的平台之一,2018 年用户已达 22 亿.除 Facebook 外,还有其他如 Twitter,Reddit 等可以发表个人观点的社交媒体平台,这些平台是不良用户散布虚假信息的重要场所^[3].大规模的网络虚假信息可能会主导舆论,使人们两极分化,甚至可能威胁到公共卫生、民主和国际关系^[4-5].

社交媒体上虚假或欺骗性的内容以不同的形式存在,并具有不同名称,如图 1 所示:

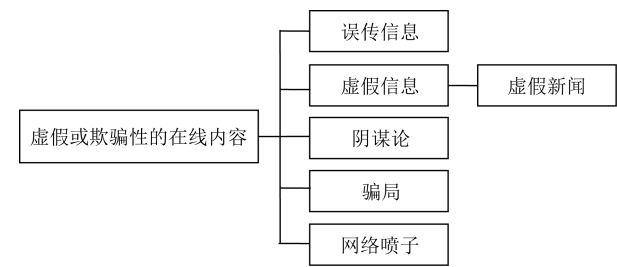


Fig. 1 Types of false or deceptive online content

图 1 虚假或欺骗性的网络内容的类型

其中,误传信息(misinformation)^[6]通常指无意传播的错误或虚假信息,被视为“诚实的错误”.虚假信息(disinformation)是指信息因别有用心

而故意传播,无论是政治上的、经济上的,或者仅仅以煽动人们为目的.这些信息通常是捏造的,还可能在虚假的上下文中加入一些真实信息以误导读者.虚假新闻(fake news)^[7]通常被视为虚假信息的子集,指以新闻方式发表或宣传的虚假内容.阴谋论(conspiracy theories)^[8]被定义为不可告人的阴谋诡计,它们怀疑恶意行为者秘密聚集在一起欺骗公众并实现某些目标.伴随阴谋论的常用词是骗局(hoax),意为看似真实但意图欺骗的言论.网络喷子(troll)^[9]试图通过传播煽动性和令人反感的言论来分化人们.

本文将综述以人工智能和跨学科思想解决虚假信息领域特定问题的相关研究、挑战和发展方向.

1 虚假信息:新冠肺炎期间的案例研究

虚假信息问题在过去几十年一直存在,近年来由于社会媒体的普及以及在线制造和传播消息的低成本,使得虚假信息更加泛滥,严重影响了公众舆论和社会稳定.

最近,新冠疫情在社交媒体上引发了一波新的误传信息和虚假信息.疫情相关的真假信息混杂造成的信息过载势不可挡,以至于世界卫生组织将其称为“信息疫情”^[10].据文献^[11]报道,新冠疫情已经引发了 2 000 个谣言和阴谋论的传播.

更令人担忧的是,目前公众已经习惯在网上寻找信息,这些虚假信息会严重地误导人们对事实的判断.2020 年 5 月,名为《Plandemic》的新冠肺炎相关阴谋电影在 Facebook 和 Youtube 等社交媒体平台上疯狂传播^[12],散布危险言论.此类信息的传播尤其具有威胁性,它们可能使人们失去对政府的信任,并可能严重破坏对该病毒防控的协调组织,从而对控制或消除病毒的进程造成严重影响.还有不良

① <https://ai.google/social-good/>
② <https://www.microsoft.com/en-us/ai/ai-for-good>
③ <https://aiforgood.itu.int/>

用户为蹭疫情热度增加流量,批量发表疫情相关的多条雷同虚假言论^[13]。

总体上,与新冠病毒有关的影响严重的虚假信息主要分为3类:健康相关的虚假信息、阴谋论、骗局和欺诈,本节将分别展开叙述。

1.1 健康相关的虚假信息

新冠病毒的高度传染性及其有害影响在全世界引起了恐慌,很多人迫切地寻找治疗方法或有希望的预防措施。不法分子利用消费者的恐慌和轻信心理,声称发现了能够“治愈”或“增强免疫力”的产品。目前已经有一些公司受到查处,相关部门对这些欺诈活动和可能对消费者造成巨大伤害的宣传发出了警告^[14]。

此类与健康有关的虚假信息可能产生严重影响,例如,某国家食品和药物管理局曾对使用氯喹和羟基氯喹治疗新冠病毒住院患者进行授权,而有证据表明这种治疗的风险大于益处^[15]。互联网上还流传着其他几个与药物有关的说法^[16-18],也都被核查为不实言论。但这些说法已经产生了严重后果,影响了被误导者的身体健康甚至导致其死亡^[19]。

1.2 阴谋论

关于新冠病毒的起源和传播,也出现了越来越多的阴谋论。皮尤研究中心(Pew Research Center)^[20]最近的一项研究显示,某发达国家71%的人曾接触过新冠病毒阴谋论,约25%的人相信其中某种阴谋^[21]。

一些阴谋论者声称戴口罩会导致二氧化碳中毒,而有些则认为这是对公民自由的侵犯^[22-23],还有一些宣称5G网络有助于病毒的传播^[22],甚至说这种病毒是在实验室制造的人造生物武器^[23]。另一些人则声称,有企业家资助了病毒的产生^[24],打算让世界人口减少,并在人身上植入微芯片^[24]。也有一部分阴谋论者认为病毒只是一个骗局,是出于某种原因而被捏造并夸大的^[25]。

1.3 骗局和欺诈

新冠病毒的流行对全球经济产生了巨大的影响,各国政府已经安排了救济/激励计划、失业救济金等,以减轻人们在困难时期的压力。不幸的是,有不法分子试图在困难时期趁火打劫^[26-27]。

在新冠疫情期间,有恶意组织和个人利用人们对病毒的恐惧,促使其仓促做出决定。许多这样的骗局都是以诈骗电话、与经济激励计划相关的电子邮件、慈善诈骗等形式出现的^[26]。

根据文献^[27],在向美国联邦贸易委员会举报

新冠疫情相关付款的204 000多起投诉电话中,68%是举报欺诈或身份盗窃。有的诈骗网站早在2020年6月就声称提供新冠病毒疫苗,或者提供免费的抗体测试^[28]。欺诈者甚至假装是接触追踪者,获取到受骗者的个人信息,然后用这些信息进行诈骗^[29]。据文献^[27]报道,虚假信息的渠道还包括提供贷款和资金帮助的诈骗电话,甚至是带有恶意软件链接的电子邮件。一些政府和民间机构已经向大众发出警告,提醒大家目前欺诈和骗局越来越多,尤其是与新冠病毒相关的骗局。

2 虚假信息检测与减缓影响的方法

近5年来,人们逐渐意识到虚假信息的影响力^[30-31],越来越多的研究者开始探索检测社交媒体虚假新闻的方法。网络信息生态系统的不可预测性和嘈杂性,以及内容的创造和传播速度快导致的恶劣影响巨大,使得减缓虚假内容的影响成为另一项重要任务,受到研究者和业界的广泛关注。如国内多家互联网企业设立辟谣平台^[32-33]帮助大众甄别谣言,强化疫情时期信息的准确性。美国Twitter和Facebook等平台也致力于管理和删除违反社区标准的内容,包括与新冠病毒相关的误导性帖子^[27]。本节将简单总结虚假信息主要的检测方法和缓解虚假信息影响的方法。

2.1 虚假信息检测方法

虚假信息检测任务是在给定文章作为输入的情况下输出“true”或“false”标签。在检测时,除了文本内容外,通常还会用到社交关系与评论等辅助信息。此外,还有基于事实核查的方法及可解释性和跨域检测方法。

1) 结合内容和社交信息的方法。大多数早期的虚假新闻或欺诈检测方法使用人工特征,通常侧重于新闻内容的文本或语言特征。语言特征通常包括词汇和语法特征等^[34-35],其他特征是依赖于平台的,例如,Twitter上的转发和点赞数量。后来,研究人员逐渐意识到,社交网络拥有许多有用信息,可以用来提高检测性能^[36]。如通过用户个人资料^[37]、用户参与度^[38],以及新闻文章、读者和发布者之间的关系来检测虚假新闻和虚假信息^[39]。

2) 结合用户评论的方法。社交媒体上的用户评论是判断一篇文章真实性的丰富信息来源。对于新出现的话题和事件,通常无法获得基本事实。在这种情况下,用户评论可以为新闻的真实性提供一些

信息,有助于虚假新闻的早期发现^[40].相关技术还可以解决标记数据有限的挑战.在早期检测的情况下,由于新闻文章刚刚发布到网上,用户的回复或评论可能不可用.对此,可以利用模型学习用户如何对新闻文章做出反应,然后生成用户反馈,以辅助检测过程^[41].与此相似的工作还有文献^[42-44].

3) 基于事实核查的方法.基于事实核查^[45]的方法有2种类型-人工或自动的方法.现阶段人工事实核查的检测方法有2种:依靠领域专家对言论进行事实核查,或者从用户处获取众包知识进行核查.自动事实核查的检测方法使用结构化的知识库(如知识图谱)检测言论和新闻文章的真实性,首先从新闻文章中提取知识(以主语、谓语、宾语三元组的形式),然后与预先存在的知识图谱进行比较实现检测.Ciampaglia等人^[46]在加权知识图谱中使用传递闭包来推导语义接近度的值.Karadzhov等人^[47]提出了一种简单的基于Web搜索的自动事实核查方法.自动事实核查中最重要的挑战之一是缺乏大规模的支持知识库或知识图谱^[48].

4) 可解释性和跨域检测的方法.可解释性检测和跨域检测是虚假新闻检测研究中的新领域.随着公平性成为人工智能研究的一个重要话题,解释一篇新闻文章为什么被归类为“虚假”是值得关注的.相关工作如文献^[49-51]通常将文章内容与用户评论作为线索,进行可解释性的研究.理解一篇文章为什么被贴上假标签可以帮助设计更好的检测算法.跨域检测是标记数据稀缺问题的潜在解决方案,但简单的虚假新闻检测器在跨领域的表现并不好^[52],因此这方面还有很广阔的研究空间.

2.2 缓解虚假信息的方法

有研究表明,因为虚假信息会使读者产生更多的极端反应,使读者对内容的参与度更高,所以虚假新闻比真实新闻传播得更快、更广^[53-54].此外,由于社交媒体排名算法的设计目的是促进帖子和内容获得更多用户参与,用户活跃度高又会使得虚假信息帖子具有更高的可见性.因此,及时发现此类虚假信息也是有效缓解社交媒体上的虚假信息传播所亟需的.本节我们将讨论一些减缓虚假信息传播影响的计算方法.

2.2.1 网络干预

社交媒体平台可以表示为一个网络,其中的节点是用户(即该平台上的帐户),边是用户之间的联系(例如“朋友关系”).社交媒体上的虚假信息以级联形式传播,其中“感染”由创建或发布虚假内容的几个用户节点引发,然后由受这些“感染源”影响的

其他用户节点传播.如果虚假信息能够有效地影响阅读它的用户,这种传播会呈指数级增长,因此,及早发现和减缓虚假信息传播影响至关重要.

减缓虚假信息级联传播的一种方法是发起反向级联,例如,包含虚假信息文章的事实核查版本的反级联.该问题通常被称为影响限制或影响最小化问题,目标是在网络中找到一个(或最小的)节点集合,在该集合中可以反级联,使得原始级联的影响最小化.

尽管为反级联传播寻找最优种子节点集已被证明是NP难的,但是研究者已经对问题的一些变体开发了近似算法^[55-57].在缓解虚假信息传播的情况下,该问题的目标是尽快“免疫”或接种尽可能多的节点,以尽量减少因虚假信息传播所造成的损害.Budak等人^[55]将一个影响限制问题的变体描述为一种贪心方法,使用精心设计的启发式方法得到问题的近似解.与文献^[57]中使用的连续时间模型相反,该算法使用了离散时间模型.尽管大多数工作只考虑了2个级联,但文献^[58]解决了多个级联的问题,并引入了“级联优先级”的概念,也即网络中的节点受每个级联影响的可能性各不相同.最近的一项工作^[59]中,考虑了跨平台的误导信息传播问题,并提出了一种基于生存理论的贪心算法解决这个问题.

2.2.2 内容标记

发布用户自创内容的社交媒体平台通常让用户选择“标记”或“报告”某些冒犯性的、有害的或虚假的信息.在这些平台上,因为用户发布的内容众多而且揭穿虚假内容需要的领域知识量巨大,所以用一个集中式系统判断这些内容是极具挑战性的.因此,多数社交媒体平台通过用户来在线维护内容,如果某些内容违反该平台的社区指南,其他用户可以标记该内容为虚假内容.

在将内容标记为“虚假信息”的情况下,这些平台通常使用第三方事实核查程序判断其真实性并采取必要的行动.如果内容被证明是虚假或有害的,平台可以删除该帖子以减少其传播范围.一些平台还告知用户平台已经执行了事实核查,并指出帖子内容的分享或交互中所存在的问题^[60-61].

虽然一些组织一直要求社交媒体平台采取积极措施控制虚假信息的传播^[62],但也有一些组织对这些平台删除虚假信息和虚假言论的行为进行了谴责,理由是发布这种虚假言论和观点不是“非法的”^[63].因此,虚假信息的标记和删除是具有挑战性的,有待探索更好的解决方案.

2.2.3 易感用户分析与识别

减缓虚假信息的另一个研究方向是分析用户行为并识别易受影响或易受骗的用户,即最容易接受和传播虚假信息用户^[64-65]。Rajabi 等人^[64]试图在 Twitter 上识别易受虚假新闻影响的用户群体,他们使用一种无监督的方法,基于用户、内容和网络特征将用户分组为 3 类:正常用户、恶意用户、脆弱或天真的用户。这种对脆弱群体的识别可能有助于进行针对性事实核查,或在接触虚假新闻后,使用特定的、合乎道德的方式向该群体提供真实信息。

3 虚假信息研究面临的挑战

在虚假信息领域的研究是具有挑战性的,尽管在某些方面研究者们已经进行了大量有意义的工作,但仍有待进一步探索。阻碍研究进程的主要问题点和该领域与其他领域问题的差别主要有 3 点:

1) 广泛性。从商品宣传、体育新闻到财经信息,几乎都存在着虚假信息。

2) 动态性。不断变化的研究领域。新的问题、挑战和威胁每天都在出现,如人工智能生成的新闻文章和极端的类人语言生成能力。

3) 跨学科性。虚假新闻和虚假信息在很大程度上是跨学科的问题,涉及计算机科学、政治科学、战略学、决策学、伦理学,甚至心理学和认知科学。

无论是 Facebook, Twitter 这样的社交媒体网站,还是维基百科这样的众包知识库,几乎都存在着虚假/捏造的内容。因为虚假信息表现形式多种多样,如文本、假视频、伪造的图片、断章取义的声明和图像等,甚至是人也很难判断一条信息是否为虚假信息。此外,先前信念、认知偏差以及回声室效应^[66]也可能影响人们对社交媒体上的声明或帖子的真实性做出判断。因此,理解和识别虚假信息这一任务对人来说已经相当具有挑战性的,实现自动化检测算法正确、有效、高效地识别虚假信息这一目标将会更加困难。

同时,缓解虚假信息传播也是一项更艰巨的任务,因为缓解的第 1 步是检测出虚假信息,这本身是一项困难的任务。而且,缓解虚假信息传播伴随着一系列挑战。我们将在下文中讨论这些挑战以及解决方法,包括:数据方面的挑战、早期检测和有效缓解方面的挑战、基于事实核查方法面临的挑战、多模态方法带来的挑战、政策问题与公平性方面的挑战。

3.1 数据挑战

缓解虚假信息传播影响的第 1 步是检测虚假信息并给予应对措施,因此虚假信息检测的准确性和实时性都是缓解虚假信息影响的重要因素。而虚假信息几乎存在于各个领域,包括政治、健康、名人绯闻等,不同领域的新闻有不同的写作风格,甚至使用完全不同的词汇,用一个领域的训练数据得到的自动化检测模型在检测另一个领域的虚假信息时往往表现不佳。然而为每个领域开发单独的分类器也是不可行的,因为这会消耗大量时间和资源,而且带有注释文本的数据集可能不适用于所有领域。获取这些数据是一个昂贵的过程,有时还需要领域知识专家。更值得探讨的是,这些数据往往会有一些关于特定事件或主题的关键字和短语,在一个数据集上训练的分类器在事件或主题不明确的领域中表现得不会很好。同时,一些新闻文章可能是纯粹的讽刺某些人或事,没有任何恶意。如果虚假信息检测器没有使用这种数据进行训练,可能无法正确地判断一篇文章是虚假的还是只为了讽刺。

文献^[67]的作者指出,另一个挑战在于训练数据集的收集,常见的收集方法是从信誉高的新闻网站提取真实的新闻文章并从可疑网站提取虚假的新闻文章。但这样的数据可能难以满足训练目的,因为即使是有信誉的新闻来源也可能会发布错误信息,而可疑的网站上并非所有的新闻都是虚假的。因此,相关研究亟需更好的注释方式,如利用事实核查网站来获取新闻文章的真实性标签,而不是依赖新闻来源的信誉^[67]。

这种虚假信息特定领域的标签问题可以利用基于迁移学习的方法训练自动虚假信息检测器。这个想法虽然还没有被广泛研究,但是最近已经开始有研究者利用迁移学习解决某些领域缺少标记数据的问题^[68-69]。针对领域或事件具有特定特征这个问题,可以在训练虚假信息分类器过程中采用对抗学习技术,防止分类器学习到领域或事件的特定特征,使其能很好地推广到其他领域或事件中。基于这一想法,Wang 等人^[70]提出了一种事件对抗神经网络,该网络试图学习与事件无关的、能共享的特征,而不是事件特定的特征,从而使该模型在新出现的虚假信息上表现良好。

3.2 早期检测和有效缓解

社交媒体平台上的不良用户为了影响读者并获得快速反应,常常用耸人听闻的标题、夸张的图片和带有情绪的语言吸引用户的注意,并让用户参与到

话题中,这些内容可能会通过读者的分享和再次分享快速蔓延.例如,“Plandemic”的虚假视频^[21]在被删除之前有 800 万观众,有大约 250 万用户通过点赞、评论和分享的形式参与了传播^[71].视频中有很多关于新冠病毒大流行的虚假说法,许多人分享并赞同它,其中包括有影响力的政治家和运动员^[71].这类虚假信息可能会造成灾难性的影响,例如降低公众对卫生保健系统的信任和协同应对病毒的积极性,进而影响公共卫生.在这样的高风险情况下,虚假信息早期检测对减少危害是极其重要的.

然而,无论是人工检测还是自动检测虚假信息,都是一项极具挑战的任务.1)对于新出现事件的說法利用人工事实核查进行早期检测具有挑战性,因为缺乏相关知识,需要领域专家进行核实.同时,第三方核查人员进行事实核查需要一定时间.2)自动检测模型没有使用新出现的内容进行训练,因此检测效果并不好.如果在训练模型的领域和使用模型的领域之间存在不一致性,性能将会更差.

此外,在虚假信息传播早期缺乏用户评论和反馈,而这些内容是虚假信息检测的辅助信息来源,能帮助人或机器甄别信息的真伪.用户评论描述了他们对信息的观点,可以辅助评判此报道是不属实的.因此,信息的缺乏是虚假信息的早期检测的重要挑战.

虚假信息的早期检测对有效缓解虚假信息的影响也至关重要.虚假信息越早被检测,就能更快地启动反向级联的应对措施,也即从受虚假信息影响的用户开始进行反向传播真实信息.虽然这是一个相对缺乏探索的领域,但一些研究人员已经提出了框架和模型来解决这个问题.Liu 和 Wu^[72]提出了一种基于用户在传播路径上的特征检测虚假新闻的方法,该方法只需获取前 5 min 的传播数据就能获得较好的检测精度.Zhou 等人^[73]提出了一种基于新闻内容提取特征的检测框架,提取的特征具有不同的层次,如语法层次、词汇层次、语义层次等,以检测已经在新闻网站上发布但尚未在社会媒体上传播的虚假新闻.Qian 等人^[41]开发了一个模型学习生成基于文章的用户反馈,然后与来自实际文章的单词和句子级信息一起用于分类过程,以解决在早期检测时缺乏用户评论作为辅助信息来源的问题.Vicario 等人的工作^[74]可以预测未来可能出现在误传信息帖子中的主题,从而提供早期预警.

3.3 基于事实核查的缓解方法面临的挑战

新闻领域的研究人员和从业者一直在使用基于事实核查的方法验证言论,并提出缓解虚假言论影

响的方法.但人工事实核查过程非常缓慢,需要领域专家检查和研究新闻文章并确定其准确性.而且每天网上都会生成和传播大量信息,人工事实核查难以及时应对,给缓解虚假信息的影响带来较大的时延.而在人们接触过虚假文章之后,再对其进行反驳和曝光,在减缓虚假信息传播的影响方面并没有像我们预期的那样有效.Ecker^[75]从心理学的角度解释了在撤回错误言论背景下人类如何反应的现象.我们将在 4.1 节中详细介绍该主题.

虽然自动事实核查^[46]比手动事实核查及时性更强,可以缓解验证言论和文章真实性的延迟问题.然而,这种方法也面临一系列挑战和障碍.大多数自动事实核查方法需要知识库(通常以知识图谱的形式),并且需要定期更新以确保知识库是最新的.但是这很难应对突然出现的话题和知识,例如大流行早期阶段的新冠肺炎症状.而且,知识库在创建的时候可能有噪音或偏差,会对事实核查性能造成影响.此外,在某些语料库上训练的自然语言处理模型可能存在语义偏差.虽然可以通过使用结构化数据(主语、谓语、宾语)来缓解这种情况,但是获得结构化数据也是很困难的.

3.4 多模态信息带来的挑战

社交媒体上的新闻文章和帖子通常伴随有与文本相关或支持文本内容的图像或视频.即使图像是伪造的,在文字旁边配上相关图片或视频也会提高新闻报道或声明的可信度.那些旨在引起舆论轰动的帖子通常会利用精心伪造的图像、视频甚至采用 Deepfakes 进行 AI 换脸,利用人眼几乎是无法区分的.真实的图像和视频也可能被裁剪,以表达一些额外的含义.由于内容的伪造和欺骗以多种不同的形式出现,让大众更容易接受,很难设计缓解方法,特别是普适性的缓解方法进行应对.

目前相关工作主要集中在多模态的虚假信息上,如 Wang 等人^[70]使用文本和视觉特征提取器,并联合 2 类特征训练虚假新闻检测器.然而,这种方法不能捕捉图像和文本之间的相关性.如果帖子中的图像和文本各自都是真实的,但一起查看可能代表完全不同的含义并且有误导读者的作用,因此,学习文本和图像的联合表示很重要.Khattar 等人^[76]使用可变自编码器学习文本和图像的共享表示,能够从该共享的潜在表示中分别重构文本和图像,从而捕获 2 种形式之间的相关性和依赖性.Zhou 等人^[77]的方法类似,先分别学习文本和图像表示,然后计算文本和视觉表示之间的相似性.识别模式

之间的“不匹配”有助于检测虚假新闻.Cui 等人^[78]在他们的模型中加入了 3 种类型的模块——文本、图像和用户个人资料,还结合了用户评论的情感判断文章/帖子是否虚假.Zhang 等人^[79]提出了一种利用外部知识图谱学习基础知识,再结合文本和视觉特征的检测方法.文献[80]在文本基础上结合使用了视频和音频,文献[81-82]也在多模态虚假新闻检测方面开展了一些工作.尽管研究者已经取得了一些检测方面的进展,但是有关缓解多模态虚假信息传播影响的研究仍十分有限,具有广阔的研究空间.

3.5 政策问题与公平性

这一挑战是在缓解虚假信息传播任务中特有的,主要表现为:当在社会媒体平台上成功检测到虚假信息后,下一步应当确定采取什么纠正措施.从社会媒体平台的角度来看,这一工作极具挑战性,因为言论自由和防止虚假信息传播之间的平衡始终是一个存在争议的问题.审查制度,尤其是社会媒体审查制度一直是国际辩论的一个问题,管理者很难采取行之有效的措施打击网上虚假信息.最近,美国宣布了一项反对社会媒体平台在线审查的行政命令^[83],理由是平台应容纳更多不同的观点.另一方面,社会媒体平台也常因其在遏制错误和虚假信息方面的糟糕表现而受到公众的指责^[84].一个有争议的帖子是否是虚假信息通常需要细致的讨论,这使得社会媒体平台很难迅速采取行动.鉴于虚假新闻和虚假信息在网络社会媒体平台上的传播速度之快,各国政府也对影响社会稳定的虚假信息和宣传感到担忧.很多国家都要求科技巨头积极主动地监管虚假信息和虚假新闻^[85-86].但在媒体、新闻以及信息的使用方面,世界上不同的国家和地区有不同的规定,这让事情变得更加复杂.

随着公平性成为人工智能研究的一个重要话题,社会媒体平台是否公平地、无偏见地监管言论也成为一个问题.自我表达和可能造成伤害的虚假信息之间存在模糊的界限,尤其是在新冠病毒应对重要事件期间.一些群体可能觉得自己比其他人更容易成为社会媒体监管的目标,因此制定缓解虚假信息政策时保护每个人的利益以达到公平是非常重要的.

4 对虚假信息的跨学科研究

正如第 3 节中提到的,虚假信息研究需要跨学科合作.利用人工智能和机器学习检测虚假信息并缓解虚假信息传播的计算方法和模型只是虚假信息

解决方案的一部分内容.这些技术需要其他领域的知识进行扩充.例如,可以从社会科学、心理学和认知科学等领域了解,理性的人类为什么会成为虚假新闻和虚假信息的受害者,如何成为受害者的,哪些心理过程触发了对虚假信息的反应和参与等.尽管研究人员已经开始在此方向展开探索,但还需要更深入的理解和学习来自其他学科的知识,以便设计更完善的解决方案.

此外,虚假信息的问题不仅仅是技术问题.需要组织令人信服的信息和活动的公众推广计划,帮助普通用户对新闻和信息的质量做出更准确的判断.

4.1 从跨学科的角度理解和检测虚假信息

为了从底层解决网络虚假信息和欺骗的问题,一些研究尝试了解个体与虚假新闻互动并相信该新闻时认知和心理的作用过程,并探究通常是什么原因让人们信任了虚假信息.

正如文献[87]中所描述的,当面对太多的信息和资源时,人类的大脑自然倾向于使用认知捷径或“启发式”来做决定,而不是批判性地分析每一条信息.在网络信息生态系统中,人们淹没在新闻、评价和观点之中,而这些内容往往是极其矛盾的,这使得评估内容真实性的任务变得越来越困难.

在可信度评估的背景下,认知捷径和启发式的想法已经在文献[88]中提到,作者解释了这种启发式的使用如何导致人们做出不正确的判断和决定.其中,声誉启发式(reputation heuristic)是指用户通过新闻来源的声誉或认可程度判断一篇新文章的可信度;支持启发式(endorsement heuristic)是指如果一个可信任的熟人共享并认可了一则新闻,用户倾向于相信新闻文章是可信的;自我确认启发式(self-confirmation heuristic)或确认偏差用户(confirmation bias users)倾向于信任与他们既有信念一致的新闻和信息源.

双重过程理论是人类易受虚假信息影响的另一个理论支撑^[89],根据此理论,人类的大脑有 2 种思维和处理信息的方式——第 1 种是一个自动(无意识的)过程,仅需要很少的能量;第 2 种是一个分析(有意识)的过程,需要大量的能量.快速自动的处理类似于启发式信息处理方式,可能误导人们让他们误判新闻的可信度.此外,动机推理,即以与先前信念一致的方式进行推理和合理化,也是人类易受信息欺骗的另一个原因^[90-91].

如今,社会媒体平台经常使用第三方事实核查人员来核实平台上分享的新闻文章和声明的真实

性.如果虚假信息被核查属实,平台通常会显示一个标签或免责声明表示事实核查结果^[60].然而,研究表明这样的处理是无效的^[75],甚至可能会促使虚假信息的读者加强他们对原始虚假叙述的信念.一些研究者^[92-93]解释了错误信息的反驳和撤回如何通过“适得其反”的作用甚至增强了人们对错误信息的信任度.由于暴露于原始错误信息与撤回之间存在时滞,这种纠正性撤回可能对读者没有任何影响,因为那时读者已经建立了心理模型来理解这一条错误信息^[94].更为复杂的是,实验表明:虚假新闻标题上的“争议”标志或事实核查似乎会增加未经过事实核查而没有标志的新闻的真实性^[95].因此,缓解虚假信息问题是一项极具挑战性的任务,需要来自各个学科和领域的知识和概念.

虚假信息检测中也用到认知科学的知识,如使用认知科学中的信息一致性及连贯性、信息源的可信度和普遍可接受性等内容^[96].

4.2 缓解虚假信息方面的跨学科努力

虚假新闻和虚假信息是一个全球现象,随着在线内容的生成和传播越来越便捷,让人们知道虚假新闻盛行并形成有效防范假新闻的意识变得日益重要.全球大约 45.7 亿名活跃互联网用户^[97](截至 2020 年 7 月),需要接受虚假信息和欺骗一直存在的事实,并了解如何识别假新闻.

此外,一些研究和调查表明,人们极其容易受到虚假新闻的影响,甚至是声称对假新闻免疫的那部分人^[98-99].斯坦福大学历史教育小组^[100]在 2019 年进行的一项研究表明,部分学生难以评估某些网络资源的可信度,容易受到错误信息的影响.令人震惊的是,90% 的学生在被要求完成的 6 项任务中有 4 项不及格.

许多组织、学者和非营利组织为这方面的研究做出了贡献,如美国媒体素养教育协会^[101]、媒体素养中心^[102-103]媒体教育实验室^[104]等,新闻业也制定了相应的标准防止产生误导性的或虚假的信息.此外,一个由约 120 家新闻机构组成的联盟“信任项目”(Trust Project, <https://thetrustproject.org/>)制定了 8 个“信任指标”,旨在建立新闻来源的可信赖性(trustworthiness)、透明度(transparency)和问责性(accountability).他们的目标是通过这些信任和问责的指标重建公众对高质量新闻来源的信任.许多国家开展了提高公众对虚假新闻流行程度的认识活动,并指导人们如何发现假新闻^[12,105].学校课程中也纳入了类似的活动,教育学生如何与网络媒

体^[106-109]互动.另一个媒体素养倡议是 Drog^[110]:除了教育项目和活动之外,他们还设计了一款游戏,通过了解虚假信息的创建和传播过程,帮助玩家建立对虚假信息的心理抵抗^[111].

5 总结及未来工作

本文简要概述了网络社会媒体平台上的虚假信息研究现状,阐述了检测虚假信息和减缓虚假信息传播的计算方法,探讨了虚假信息检测与减缓其影响中的各种挑战,以及应对这些挑战的一些方法,最后从跨学科的角度讨论了检测和减缓虚假信息传播的一些努力.

目前,许多有关虚假信息研究的领域还没有得到充分的探索,还有很大的研究空间.随着最近在大规模、高表现力语言模型和自动文本生成能力方面的进步,大规模的虚假信息攻击逐渐成为研究人员的潜在威胁.OpenAI 最近发表的 GPT-2 和 GPT-3 模型性能强大,恶意代理模型可以微调这些模型自动生成虚假或误导性的新闻文章^[112-113].读者通常是感性的,很可能会相信机器生成的文章是合法的新闻.因此,防御这种攻击是很有前景的新研究领域,值得重点关注.

此外,在缓解虚假信息方面进行更多的跨学科研究是十分有必要的.目前已经有在人类的轻信性和对虚假信息的敏感性方面的研究,可以用于计算机科学领域以设计更好的推荐或排序算法,对社会媒体平台上的新闻进行排名,但是,目前社会媒体平台上的推荐算法仍倾向于有选择地向用户展示与他们历史兴趣相一致的内容,从而加剧了回声室的问题,加剧了人群的对立,最终使得减缓虚假信息传播影响的任務更加具有挑战性.因此,研究者们有必要设计更好的算法,以确保用户在理想的新闻文章和具有相反观点的文章之间保持良好的平衡.

参 考 文 献

[1] Shu Kai, Wang Suhang, Lee D, et al. Disinformation, Misinformation, and Fake News in Social Media [M]. Berlin: Springer, 2020: 1-19

[2] Clement J. Number of internet users worldwide from 2005 to 2019 [EB/OL]. [2020-09-20]. <https://www.statista.com/statistics/273018/number-of-internet-users-worldwide>

[3] Liu Bo, Li Yang, Meng Qing, et al. Evaluation of content credibility in social media [J]. Journal of Computer Research and Development, 2019, 56(9): 1939-1952 (in Chinese)

- (刘波, 李洋, 孟青, 等. 社交媒体内容可信性分析与评价[J]. 计算机研究与发展, 2019, 56(9): 1939-1952)
- [4] Gangware C, Nemr W. Weapons of mass distraction: Foreign state-sponsored disinformation in the digital age [EB/OL]. [2020-09-21]. <https://www.state.gov/weapons-of-mass-distraction-foreign-state-sponsored-disinformation-in-the-digital-age/>
 - [5] Jackson D. Issue brief: How disinformation impacts politics and publics [EB/OL]. [2020-09-21]. <https://www.ned.org/issue-brief-how-disinformation-impacts-politics-and-publics/>
 - [6] Wu Liang, Morstatter F, Carley K M, et al. Misinformation in social media: Definition, manipulation, and detection [J]. ACM SIGKDD Explorations Newsletter, 2019, 21(2): 80-90
 - [7] Stroud F. Fake news definition [EB/OL]. [2020-09-21]. <https://www.webopedia.com/TERM/F/fake-news.html>
 - [8] Bale J M. Political paranoia v. political realism: On distinguishing between bogus conspiracy theories and genuine conspiratorial politics [J]. Patterns of Prejudice, 2007, 41(1): 45-60
 - [9] Wikipedia. Internet troll [EB/OL]. [2020-09-22]. https://en.wikipedia.org/wiki/Internet_troll
 - [10] World health organization. Novel coronavirus (2019-ncov): Situation report-13 [EB/OL]. [2020-08-31]. <https://www.who.int/docs/default-source/coronaviruse/situationreports/20200220-sitrep-13-ncov-v3.pdf>
 - [11] Rettner R. COVID-19 has fueled more than 2000 rumors and conspiracy theories [EB/OL]. [2020-09-20]. <https://www.livescience.com/covid-19-rumors-conspiracy-theories-infodemic.html>
 - [12] WebWise. Be media smart. New public awareness campaign launched [EB/OL]. [2020-09-17]. <https://www.webwise.ie/trending/be-media-smart-new-public-awareness-campaign-launched/>
 - [13] Baidu Baijia (capricious field). Freelance journalists should not make false news and exaggerate facts to get a wave to ride [EB/OL]. [2020-10-08]. <https://baijiahao.baidu.com/s?id=1661592437045003415&wfr=spider&for=pc> (in Chinese)
(百度百家号(任性领域)). @自媒体从业者, 为了蹭热度, 不应该制造假新闻和夸大事实[EB/OL]. [2020-10-08]. <https://baijiahao.baidu.com/s?id=1661592437045003415&wfr=spider&for=pc>
 - [14] James L. Attorney general James orders Alex Jones to stop selling fake coronavirus treatments [EB/OL]. [2020-07-25]. <https://ag.ny.gov/press-release/2020/attorney-general-james-orders-alex-jones-stop-selling-fake-coronavirus-treatments>
 - [15] Food and Drug Administration. Coronavirus (COVID-19) update: FDA revokes emergency use authorization for chloroquine and hydroxychloroquine [EB/OL]. [2020-09-22]. <https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-revokes-emergency-use-authorization-chloroquine-and>
 - [16] Smith D. Coronavirus: Medical experts denounce trump's theory of 'disinfectant injection' [EB/OL]. [2020-09-20]. <https://www.theguardian.com/world/2020/apr/23/trump-coronavirus-treatment-disinfectant>
 - [17] Pilkington E. Bleach touted as 'miracle cure' for COVID being sold on amazon [EB/OL]. [2020-09-20]. <https://www.theguardian.com/world/2020/sep/19/bleach-miracle-cure-amazon-covid>
 - [18] Emily P. Drinking bleach, hydroxychloroquine and other COVID-19 myths [EB/OL]. [2020-09-20]. <https://www.ems1.com/coronavirus-covid-19/articles/drinking-bleach-hydroxychloroquine-and-other-covid-19-myths-4cpZ1a0uImeCph70/>
 - [19] Waldrop T, Dave A. Fearing coronavirus, Arizona man dies after taking a form of chloroquine used to treat aquariums [EB/OL]. [2020-09-20]. <https://www.cnn.com/2020/03/23/health/arizona-coronavirus-chloroquine-death/index.html>
 - [20] Schaeffer K. A look at the Americans who believe there is some truth to the conspiracy theory that covid-19 was planned [EB/OL]. [2020-07-29]. <https://www.pewresearch.org/fact-tank/2020/07/24/a-look-at-the-americans-who-believe-there-is-some-truth-to-the-conspiracy-theory-that-covid-19-was-planned/>
 - [21] Newton C. How the 'plandemic' video hoax went viral [EB/OL]. [2020-09-15]. <https://www.theverge.com/2020/5/12/21254184/how-plandemic-went-viral-facebook-youtube>
 - [22] Rettner R. 14 coronavirus myths busted by science [EB/OL]. [2020-09-20]. <https://www.livescience.com/coronavirus-myths.html>
 - [23] Citroner G. No, face masks can't cause CO2 poisoning [EB/OL]. [2020-09-20]. <https://www.healthline.com/health-news/no-face-masks-cant-cause-co2-poisoning>
 - [24] Wakefield J. How Bill Gates became the voodoo doll of COVID conspiracies [EB/OL]. [2020-09-20]. <https://www.bbc.com/news/technology-52833706>
 - [25] Newton C. Coronavirus deniers and hoaxers persist despite dire warnings, claiming 'it's mass hysteria' [EB/OL]. [2020-09-20]. https://www.washingtonpost.com/national/coronavirus-deniers-outbreak-hoax/2020/03/19/46bc5e46-6872-11ea-b313-df458622c2cc_story.html
 - [26] Federal Communications Commission. Coronavirus scams-consumer resources [EB/OL]. [2020-10-11]. <https://www.fcc.gov/covid-scams>
 - [27] Waggoner J, Markowitz A. Beware of robocalls, texts and emails promising COVID-19 cures or stimulus payments [EB/OL]. [2020-09-21]. <https://www.aarp.org/money/scams-fraud/info-2020/coronavirus.html>
 - [28] Federal Bureau of Investigation. FBI warns of potential fraud in antibody testing for COVID-19 [EB/OL]. [2020-09-21]. <https://www.fbi.gov/news/pressrel/press-releases/fbi-warns-of-potential-fraud-in-antibody-testing-for-covid-19>
 - [29] American Association of Retired Persons. Health insurance scams [EB/OL]. [2020-09-21]. <https://www.aarp.org/money/scams-fraud/info-2019/health-insurance.html>

- [30] Allcott H, Gentzkow M. Social media and fake news in the 2016 election [J]. *Journal of Economic Perspectives*, 2017, 31(2): 211–36
- [31] Shu Kai, Liu Huan. *Detecting Fake News on Social Media* [M]. San Mateo, CA: Morgan & Claypool, 2019: 3
- [32] Tencent. A serious rumor-refuting platform [EB/OL]. [2020-10-09]. <https://xw.qq.com/z/20200124008392/HLH2020012400839200> (in Chinese)
(腾讯. 较真辟谣 [EB/OL]. [2020-10-09]. <https://xw.qq.com/z/20200124008392/HLH2020012400839200>)
- [33] Liu Junhuan. News flies high in the sky, with rumors floating aside. Under the epidemic, how can domestic and foreign social platforms accurately refute rumors? [EB/OL]. [2020-10-12]. https://mp.weixin.qq.com/s/c1W8_Gg4yNt9alBYrrIi7A (in Chinese)
(刘俊寰. 新闻和谣言齐飞, 疫情之下, 国内外社交平台如何精准辟谣? [EB/OL]. [2020-10-12]. https://mp.weixin.qq.com/s/c1W8_Gg4yNt9alBYrrIi7A)
- [34] Ott M, Choi Y, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination [J]. *arXiv preprint*, arXiv:1107.4557, 2011
- [35] Song Feng, Banerjee R, Choi Y. Syntactic stylometry for deception detection [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2012: 171–175
- [36] Shu Kai, Wang Suhang, Liu Huan. Beyond news contents: The role of social context for fake news detection [C] //Proc of the 12th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2019: 312–320
- [37] Shu Kai, Wang Suhang, Liu Huan. Understanding user profiles on social media for fake news detection [C] //Proc of the 2018 IEEE Conf on Multimedia Information Processing and Retrieval (MIPR). Piscataway, NJ: IEEE, 2018: 430–435
- [38] Shu Kai, Mahudeswaran D, Wang Suhang, et al. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media [J]. *Big Data*, 2020, 8(3): 171–188
- [39] Della Vedova M L, Tacchini E, Moret S, et al. Automatic online fake news detection combining content and social signals [C] //Proc of the 22nd Conf of Open Innovations Association (FRUCT). Piscataway, NJ: IEEE, 2018: 272–279
- [40] Shu Kai, Zheng Guoqing, Li Yichuan, et al. Leveraging multi-source weak social supervision for early detection of fake news [J]. *arXiv preprint*, arXiv:2004.01732, 2020
- [41] Qian Feng, Gong Chengyue, Sharma K, et al. Neural user response generator: Fake news detection with collective user intelligence [C] //Proc of IJCAI. Palo Alto, CA: AAAI, 2018, 18: 3834–3840
- [42] Kim J, Tabibian B, Oh A, et al. Leveraging the crowd to detect and reduce the spread of fake news and misinformation [C] //Proc of the 17th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 324–332
- [43] Tschachtschek S, Singla A, Gomez Rodriguez M, et al. Fake news detection in social networks via crowd signals [C] //Proc of the Web Conf 2018. Republic and Canton of Geneva, Switzerland: Int World Wide Web Conferences Steering Committee, 2018: 517–524
- [44] Tchakounte F, Faissal A, Atemkeng M, et al. A reliable weighting scheme for the aggregation of crowd intelligence to detect fake news [J]. *Information*, 2020, 11(6): 1–25
- [45] Vlachos A, Riedel S. Fact checking: Task definition and dataset construction [C] //Proc of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, MD: Association for Computational Linguistics, 2014: 18–22
- [46] Ciampaglia G L, Shiralkar P, Rocha L M, et al. Computational fact checking from knowledge networks [J]. *PloS One*, 2015, 10(6): No.e0128193
- [47] Karadzhov G, Nakov P, Márquez L, et al. Fully automated fact checking using external sources [J]. *arXiv preprint*, arXiv:1710.00341, 2017
- [48] Graves D. Understanding the promise and limits of automated fact-checking [EB/OL]. [2020-08-31]. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226FINAL.pdf
- [49] Shu Kai, Cui Limeng, Wang Suhang, et al. defend: Explainable fake news detection [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 395–405
- [50] Yang Fan, Penttala S K, Mohseni S, et al. Xfake: Explainable fake news detector with visualizations [C] //Proc of the 28th World Wide Web Conf. New York: ACM, 2019: 3600–3604
- [51] Reis J C S, Correia A, Murai F, et al. Explainable machine learning for fake news detection [C] //Proc of the 10th ACM Conf on Web Science. New York: ACM, 2019: 17–26
- [52] Janicka M, Pszozna M, Wawer A. Cross-domain failures of fake news detection [J]. *Computación y Sistemas*, 2019, 23(3): 1089–1097
- [53] Vosoughi S, Roy D, Aral S. The spread of true and false news online [J]. *Science*, 2018, 359(6380): 1146–1151
- [54] Fox M. Fake news: Lies spread faster on social media than truth does [EB/OL]. [2020-09-15]. <https://www.nbcnews.com/health/health-news/fake-news-lies-spread-faster-social-media-truth-does-n854896>
- [55] Budak C, Agrawal D, Abbadi A E. Limiting the spread of misinformation in social networks [C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 665–674
- [56] Nguyen N P, Yan G, Thai M T, et al. Containment of misinformation spread in online social networks [C] //Proc of the 4th Annual ACM Web Science Conf. New York: ACM, 2012: 213–222

- [57] Luo Chuan, Cui Kainan, Zheng Xiaolong, et al. Time critical disinformation influence minimization in online social networks [C] //Proc of the 2014 IEEE Joint Intelligence and Security Informatics Conf. Piscataway, NJ: IEEE, 2014: 68-74
- [58] Tong A, Du Dingzhu, Wu Weili. On misinformation containment in online social networks [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 341-351
- [59] Hosni A I E, Li K, Ahmad S. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors [J]. Information Sciences, 2020, 512: 1458-1480
- [60] Facebook Investor Relations. News feed fyi: Addressing hoaxes and fake news [EB/OL]. [2020-08-31]. <https://www.facebook.com/facebookinvestorrelations/posts/news-feed-fyi-addressing-hoaxes-and-fake-news/1030894543703515/>
- [61] Facebook Newsroom. News feed fyi: Replacing disputed flags with related articles [EB/OL]. [2020-08-31]. <https://twitter.com/fbnewsroom/status/943647232485507073>
- [62] Lomas N. Tech giants pressured to auto-flag “illegal” content in Europe [EB/OL]. [2020-08-31]. <https://techcrunch.com/2017/09/28/tech-giants-pressured-to-auto-flag-illegal-content-in-europe/>
- [63] Ohlheiser A. Trump responds to Twitter’s fact-check by targeting social-media protections [EB/OL]. [2020-08-31]. <https://www.technologyreview.com/2020/05/28/1002357/trump-signs-social-media-order-after-twitter-fact-check/>
- [64] Rajabi Z, Amarda S, Hemant P. User behavior modelling for fake information mitigation on social Web [C] //Proc of the 12th Int Conf on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Berlin: Springer, 2019: 234-244
- [65] Sameki M, Zhang Tianyi, Ding Linli, et al. Crowd-o-meter: Predicting if a person is vulnerable to believe political claims [C] //Proc of HCOMP. Palo Alto, CA: AAAI, 2017: 157-166
- [66] Martin D. # Republic: Divided democracy in the age of social media [J]. Business Ethics Quarterly, 2018, 28(3): 360-363
- [67] Asr F T, Taboada M. The data challenge in misinformation detection: Source reputation vs. content veracity [C] //Proc of FEVER. Stroudsburg, PA: ACL, 2018: 10-15
- [68] Dhamani N, Azunre P, Gleason J L, et al. Using deep networks and transfer learning to address disinformation [J]. arXiv preprint arXiv, 2019: 1905.10412
- [69] Guo Meicheng, Xu Zhiwei, Liu Limin, et al. An adaptive deep transfer learning model for rumor detection without sufficient identified rumors [J]. Mathematical Problems in Engineering, 2020(2020): No.7562567
- [70] Wang Yaqing, Ma Fenglong, Jin Zhiwei, et al. Eann: Event adversarial neural networks for multi-modal fake news detection [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 849-857
- [71] Andrews T M. Facebook and other companies are removing viral ‘plandemic’ conspiracy video [EB/OL]. [2020-09-15]. <https://www.washingtonpost.com/technology/2020/05/07/plandemic-youtube-facebook-vimeo-remove/>
- [72] Liu Yang, Wu Yi-Fang Brook. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 354-361
- [73] Zhou Xinyi, Jain A, Phoha V V, et al. Fake news early detection: A theory-driven model [J]. Digital Threats: Research and Practice, 2020, 1(2): 37-61
- [74] Vicario M D, Quattrociocchi W, Scala A, et al. Polarization and fake news: Early warning of potential misinformation targets [J]. ACM Transactions on the Web, 2019, 13(2): 108-129
- [75] Ecker U K H. Why rebuttals may not work: The psychology of misinformation [J]. Media Asia, 2017, 44(2): 79-87
- [76] Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection [C] //Proc of the World Wide Web Conf. New York: Association for Computing Machinery, 2019: 2915-2921
- [77] Zhou Xinyi, Wu Jindi, Zafarani R. SAFE: Similarity-aware multi-modal fake news detection [J]. arXiv preprint, arXiv: 2003.04981, 2020
- [78] Cui Limeng, Wang Suhang, Lee D W. SAME: Sentiment-aware multi-modal embedding for detecting fake news [C] //Proc of the 2019 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. New York: ACM, 2019: 41-48
- [79] Zhang Huaiwen, Fang Quan, Qian Shengsheng, et al. Multi-modal knowledge-aware event memory network for social media rumor detection [C] //Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1942-1951
- [80] Krishnamurthy G, Majumder N, Poria S, et al. A deep learning approach for multimodal deception detection [J]. arXiv preprint, arXiv: 1803.00344, 2018
- [81] Kang S, Hwang J, Yu H. Multi-modal component embedding for fake news detection [C] //Proc of the 14th Int Conf on Ubiquitous Information Management and Communication (IMCOM). Piscataway, NJ: IEEE, 2020: 35-40
- [82] Singhal S, Shah R R, Chakraborty T, et al. Spotfake: A multi-modal framework for fake news detection [C] //Proc of the 5th Int Conf on Multimedia Big Data (BigMM). Piscataway, NJ: IEEE, 2019: 39-47

- [83] Whitehouse.gov. Executive order on preventing online censorship [EB/OL]. [2020-09-23]. <https://www.whitehouse.gov/presidential-actions/executive-order-preventing-online-censorship/>
- [84] Natasha Lomas. US legislator, David Cicilline, joins international push to interrogate platform power [EB/OL]. [2020-09-23]. <https://techcrunch.com/2019/08/15/us-legislator-david-cicilline-joins-international-push-to-interrogate-platform-power/>
- [85] Rankin J. Tech firms could face new EU regulations over fake news [EB/OL]. [2020-09-23]. <https://www.theguardian.com/media/2018/apr/24/eu-to-warn-social-media-firms-over-fake-news-and-data-mining>
- [86] The Guardian. Emmanuel macron promises ban on fake news during elections [EB/OL]. [2020-09-23]. <https://www.theguardian.com/world/2018/jan/03/emmanuel-macron-ban-fake-news-french-president>
- [87] Steinmetz K. How your brain tricks you into believing fake news [EB/OL]. [2020-09-17]. <https://time.com/5362183/the-real-fake-news-crisis/>
- [88] Metzger M J, Flanagin A J. Credibility and trust of information in online environments: The use of cognitive heuristics [J]. *Journal of Pragmatics*, 2013, 59 (1/2): 210-220
- [89] Pennycook G. A perspective on the theoretical foundation of dual process models [G] //Current issues in Thinking and Reasoning. Dual Process Theory 2.0. Oxfordshire, UK: Routledge/Taylor & Francis, 2018: 5-27
- [90] The psychology of misinformation: Why we're vulnerable [EB/OL]. [2020-09-20]. <https://firstdraftnews.org/latest/the-psychology-of-misinformation-why-were-vulnerable/>
- [91] Why do people fall for fake news? [EB/OL]. [2020-09-20]. <https://www.nytimes.com/2019/01/19/opinion/sunday/fake-news.html>
- [92] Cook J, Ecker U, Lewandowsky S. Misinformation and how to correct it [G] //Emerging Trends in the Social and Behavioral Sciences. Hoboken, NJ: John Wiley & Sons, 2015: 1-17
- [93] Nyhan B, Reifler J. When corrections fail: The persistence of political misperceptions [J]. *Political Behavior*, 2010, 32(2): 303-330
- [94] Thorson E. Belief echoes: The persistent effects of corrected misinformation [J]. *Political Communication*, 2016, 33(3): 460-480
- [95] Pennycook G, Bear A, Collins E T, et al. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings [J]. *Management Science*, 2020
- [96] Kumar K K, Geethakumari G. Detecting misinformation in online social networks using cognitive psychology [J]. *Human-centric Computing and Information Sciences*, 2014, 4(1): 250-271
- [97] Clement J. Global digital population as of July 2020 [EB/OL]. [2020-09-17]. <https://www.statista.com/statistics/617136/digital-population-worldwide>
- [98] Ng Huiwen. 4 in 5 Singaporeans confident in spotting fake news but 90 per cent wrong when put to the test: Survey [EB/OL]. [2020-09-17]. <https://www.statista.com/statistics/617136/digital-population-worldwide>
- [99] Diana G. Why are we still falling for fake news? [EB/OL]. [2020-09-17]. <https://www.psychologytoday.com/us/blog/raising-humans-in-digital-world/201905/why-are-we-still-falling-fake-news>
- [100] Breakstone J, Smith M, Wineburg S, et al. Students' civic online reasoning [EB/OL]. [2020-09-23]. <https://sheg.stanford.edu/students-civic-online-reasoning>
- [101] The National Association for Media Literacy Education (NAMLE). Our mission [EB/OL]. [2020-09-17]. <https://namle.net/about>
- [102] Media Literacy. Media literacy now [EB/OL]. [2020-09-17]. <https://medialiteracynow.org/>
- [103] Center for Media Literacy. CML's Leadership Role in Professional Development [EB/OL]. [2020-09-17]. <http://www.medialit.org/cmls-leadership-role-professional-development>
- [104] Media Education Lab. About us [EB/OL]. [2020-09-17]. <https://mediaeducationlab.com/>
- [105] Mackintosh E, Kiernan E. Finland is winning the war on fake news. what it's learned may be crucial to western democracy [EB/OL]. [2020-10-01]. <https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/>
- [106] Danielson M. "media literacy is literacy": Here's how educators and lawmakers are working to set students up for success online [EB/OL]. [2020-09-17]. <https://www.the74million.org/article/media-literacy-is-literacy/>
- [107] Foley R J. Spread of fake news prompts literacy efforts in schools [EB/OL]. [2020-09-17]. <https://www.pbs.org/newshour/education/spread-of-fake-news-prompts-literacy-efforts-in-schools>
- [108] Charlton E. How Finland is fighting fake news-in the classroom [EB/OL]. [2020-09-17]. <https://www.weforum.org/agenda/2019/05/how-finland-is-fighting-fake-news-in-the-classroom/>
- [109] Timsit A. In the age of fake news, here's how schools are teaching kids to think like fact-checkers [EB/OL]. [2020-09-17]. <https://qz.com/1533747/in-the-age-of-fake-news-heres-how-schools-are-teaching-kids-to-think-like-fact-checkers/>
- [110] Arentze M. Fight disinformation through gaming and education: The Drog media literacy initiative [EB/OL]. [2020-09-23]. <https://www.disinfo.eu/outreach/our-webinars/fight-disinformation-through-gaming-and-education-the-drog-media-literacy-initiative/>

[111] Roozenbeek J, Linden S V D. Fake news game confers psychological resistance against online misinformation [J]. Palgrave Communications, 2019, 5(1): No.65

[112] Tung L. Openai's ‘dangerous’ AI text generator is out: People find gpt-2’s words ‘convincing’ [EB/OL]. [2020-09-23]. <https://www.zdnet.com/article/openais-dangerous-ai-text-generator-is-out-people-find-gpt-2s-words-convincing/>

[113] Holmes A. A fake blog written by ai shot to the top of hacker news after people thought it was real-here’s how a college student made it [EB/OL]. [2020-09-23]. <https://www.businessinsider.com/fake-ai-generated-gpt3-blog-hacker-news-2020-8>



Amrita Bhattacharjee, born in 1997. PhD candidate. Her main research interests include data mining, machine learning, and social computing.

Amrita Bhattacharjee, 1997 年生.博士研究生.主要研究方向为数据挖掘、机器学习、社会计算.



Shu Kai, born in 1990. PhD, assistant professor. His main research interests include machine learning, data mining, and social computing. (kshu@iit.edu)

舒 凯,1990 年生.博士,助理教授.主要研究方向为机器学习、数据挖掘、社会计算.



Gao Min, born in 1980. PhD, associate professor. Member of CCF. Her main research interests include recommender systems, anomaly detection, and data mining.

高 旻,1980 年生.博士,副教授.CCF 会员.主要研究方向为推荐系统、异常检测、数据挖掘.



Liu Huan, PhD, professor. Fellow of ACM, AAAI, AAAS, and IEEE. His main research interests include AI, machine learning, data mining, and social computing. (huanliu@asu.edu)

刘 欢,博士,教授.ACM/AAAI/AAAS/IEEE 会士.主要研究方向为人工智能、机器学习、数据挖掘、社会计算.