社交媒体优化综述

王晓诗[1),2)] 景少玲[2)] 孙 飞[2)] 尹芷仪[2)] 沈华伟[1),2)] 程学旗[1),2)]

[1)](中国科学院大学 北京 101408)

[2)](智能算法安全全国重点实验室 中国科学院计算技术研究所 北京 100190)

摘 要 随着移动互联网的崛起,社交媒体在信息传播中的重要性日益凸显。同时,大量证据表明,社交媒体优化被广泛应用于商业营销、政治宣传或传播观点,尤其以 ChatGPT 为代表的大语言模型技术正被当作工具进一步降低社交媒体优化的门槛,优化技术及其带来的潜在风险引起越来越多研究者的关注,但现有针对社交媒体优化议题的研究主要集中在证实恶意优化现象的存在,或揭示量化特定事件中的优化行为,缺乏系统全面的介绍。本文从宏观和微观两个视角系统性地综述了该研究领域问题,首先,总结归纳了社交媒体优化相关案例和研究进展,首次明确定义了社交媒体优化的概念,提炼了优化的三个核心要素:优化内容、受控账号和优化活动;其次,以三要素为研究对象,结合大语言模型等新兴技术的发展,深入剖析了各要素的定义、分类、获取方法、识别技术;此外,从量化和解决优化问题的角度,讨论了社交媒体优化的度量方法、对抗社交媒体恶意优化的技术措施以及社会治理手段。最后,总结了该领域面临的挑战,提出了未来的研究建议,为深入分析社交媒体优化行为和解决社交媒体恶意优化问题提供重要参考和借鉴。

关键词 社交媒体优化;虚假信息;社交媒体机器人;计算宣传;大语言模型

中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2025.02181

Social Media Optimization: A Survey

WANG Xiao-Shi[1),2)] JING Shao-Ling[2)] SUN Fei[2)] YIN Zhi-Yi[2)] SHEN Hua-Wei[1),2)]
CHENG Xue-Qi[1),2)]

[1)](University of Chinese Academy of Sciences, Beijing 101408)

[2)](State Key Lab of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract With the rise of mobile internet usage, the importance of social media in information dissemination has become increasingly prominent. At the same time, substantial evidence shows that social media optimization has been widely applied in fields such as commercial marketing, political propaganda, and opinion dissemination. Notably, large language model technologies, represented by ChatGPT, are being utilized as tools to further lower the threshold for social media optimization. Researchers' focus on these optimization techniques and their potential risks continues to grow. For instance, in 2023, Meta published four landmark papers in *Science* and *Nature*, exploring the impact of social media on stances, attitudes, and behaviors during the 2020 U.S. election. However, existing research on social media optimization, conducted from three perspectives—revealing and quantifying phenomena, analyzing implementation processes, and discussing application impacts—has mainly focused on confirming the existence of malicious optimization. It

收稿日期:2024-03-18;在线发布日期:2025-04-01。本课题得到国家自然科学基金项目(U21B2046)资助。**王晓诗**,博士研究生,工程师,主要研究领域为算法安全、可信人工智能、算法治理等。E-mail:wangxiaoshi@ict.ac.cn。**景少玲**,博士,工程师,主要研究领域为自然语言处理、大模型测评等。**孙 飞**,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为推荐算法、自然语言处理等。**尹芷仪**,博士,高级工程师,主要研究领域为网络安全、社会计算等。**沈华伟(通信作者)**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为网络数据挖掘、社交网络分析、图神经网络。E-mail:shenhuawei@ict.ac.cn。**程学旗**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为数据科学与大数据分析系统、网络科学与社会计算、Web 搜索与挖掘。

has also concentrated on revealing and quantifying such behaviors in specific events. This has led to a lack of systematic and comprehensive summarization and organization, offering limited guidance for understanding research in this field or addressing the risks and harms posed by optimization. To address this gap, this paper systematically reviews research on social media optimization by synthesizing insights from computer science, media studies, law, and industry, examining the issue from both macro and micro perspectives. At the macro level, this paper organizes relevant cases and research on social media optimization, providing a clear definition of the concept for the first time. Social media optimization refers to the organized and deliberate dissemination or suppression of specific information or viewpoints through social media platforms, with the aim of shaping or damaging the image or arguments of specific interests, thereby influencing public opinion. Furthermore, this paper identifies three core elements of social media optimization: optimized content, controllable accounts, and optimization activities, and summarizes the relationships among them. The combination of these three elements enables the implementation of a social media optimization campaign. At the micro level, this paper first examines each of the three core elements—optimized content, controllable accounts, and optimization activities—in light of the development of emerging technologies such as large language models, delving into their definitions, classifications, acquisition methods, and identification techniques. Next, the paper discusses methods for measuring social media optimization from four perspectives: the scale of exposed optimization behaviors, the degree of impact on real-world events, the extent of influence on group cognition, and the platform's ability to counteract optimization, summarizing the current progress in quantitative research on social media optimization. Additionally, from both technical and social governance perspectives, the paper analyzes measures to address malicious social media optimization, including the use of artificial intelligence to identify optimization behaviors, fact-checking to verify information authenticity, evaluating and enhancing algorithmic reliability, government policy-making and regulation, platform norms and information transparency, and public oversight. Finally, the paper concludes by summarizing social media optimization, highlighting the challenges in the field, and proposing directions for future research. This provides valuable references and insights for the in-depth analysis of social media optimization behaviors and addressing the issues arising from malicious optimization.

Keywords social media optimization; disinformation; social media bots; computational propaganda; lardge language model

1 引 言

随着移动互联网的崛起,信息传播范式发生了根本性变化,从传统媒体主导的集中式传播模式过渡到基于社交媒体的、由大众网民驱动的开放性分布式信息传播模式。据 SmartInsight 组织公布数据显示,2023 年 4 月,全球 60%的人口使用社交媒体,平均每日使用时长为 2 小时 24 分钟①。社交媒体平台已成为公众获取信息和消费新闻的主要媒介,也被广泛用于社会事件动员中,利用其交互属性对用户群体进行告知、吸引和动员。然而,社交媒体平台内容的生产和传播依靠公众广泛参与而存在不可控性,传播者通过一系列策略和技巧来优化社交媒体平台上的内容和互动,甚至被恶意优化来误导或控制公众舆论。随着“阿拉伯之春”“英国脱欧”“2016 年美国大选”“冲击美国国会山”“俄乌战争”等事件在社交媒体上的舆论发酵和塑造,社交媒体优化问题已引起民众的警觉和广泛关注。同时,随着 ChatGPT 为代表的大语言模型等新兴技术的兴起,这类技术的应用大大降低了社交媒体优化的门

① https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/。

槛,因此社会各界对优化的技术手段及其带来的潜在问题表示关切和忧虑。

社交媒体优化常被应用于商业营销场景,部分学者探讨其作为一种营销技术在提高商业影响力方面的作用,但更多的研究集中在社会议题及优化带来的潜在风险上,如 2006 年,牛津大学互联网研究所的 Philip 首次对利用社交媒体影响公众舆论和传播错误信息的可能表示担忧[1]。2010 年,首次出现优化活动在政治活动中出现的证实研究,Ratkiewicz 等人[2]发现社交机器人在美国中期选举期间被用于支持或抹黑部分候选人。2012 年,《Science》期刊发布了首例分析恶意优化行为实施方式的文章,总结了"网络垃圾邮件""Twitter 炸弹""预制推文工厂""购买在线搜索广告""使用嘲笑对手的照片和视频"共 5 种方式[3]。2013 年,世界经济论坛将大量数字错误信息与网络攻击、恐怖主义一起列为社会面临的核心技术风险[4]。2016 年美国总统大选作为社交媒体分析的最重要案例被广泛研究[5-8],揭示了大量利用社交媒体优化技术影响选举的内幕。2023 年 7 月 28 日,"Meta 的算法影响 2020 年美国选举"登上了《Science》封面,Meta 同时在《Science》和《Nature》发表了 4 篇具有里程碑意义的论文[9-12],探讨了美国 2020 年选举中社交媒体对立场、态度和行为的影响,引起越来越多研究者对社交媒体优化的关注。目前,对社交媒体优化的研究从侧重点不同可以分为案例、方法和影响三个角度:

(1)揭示量化现象。具体表现为通过案例研究(Case Study)的方式分析单个社交媒体优化事件,以论证其存在。由于在商业领域的优化实践不胜枚举,本文汇编了从 2010 年开始至今共 22 个发生在政治议题上的社交媒体事件及分析,具体列表如表 1 所示,相关研究主要针对某一具体社会事件的一个或两个在社交媒体上的优化角度如假新闻、虚假信息、社交机器人开展分析。

表 1　以 Case Study 形式开展的政治领域的社交媒体优化研究

时间	研究对象	文献
2010	美国中期选举	Ratkiewicz 等[2]
2012	韩国总统选举	Keller 等[26]
2014	日本众议院议员总选举	Schafer 等[21]
2016	美国黑人遭警察暴力执法	Arif 等[27]
2016	美国总统大选	Ourney 等[5]、Grinberg 等[6]、Bovet 等[7]、Eady 等[8]、Allcott 等[28]、FGuess 等[29]、Shao 等[30]、Howard 等[31]、Hindman 等[32]、Shao 等[33]、Bessi 等[34]、Abu 等[35]、Faris 等[36]
2016	英国脱欧公投	Howard 等[37]、Bastos 等[38]、Bastos 等[39]
2016	意大利宪法公投	Del 等[40]、Del 等[41]、Beyond 等[42]
2017	法国总统选举	Ferrara 等[43]
2017	德国联邦议院选举	Brachten 等[44]、Morstatter 等[45]
2017	海湾危机	Jones 等[46]
2018	美国中期选举	Cardoso 等[47]
2018	意大利大举	Giglietto 等[48]
2018	巴西总统选举	Resende 等[49]
2019	欧洲议会选举	Pierri 等[50]
2019	北非移民救助	Caldarelli 等[51]
2019	印度大选	Narayanan 等[52]
2020	英国正式脱欧	Assenmacher 等[53]
2020	美国总统大选	Barbera 等[9]、Guess 等[10]、Guess 等[11]、Gonzalez 等[12]、Ferrara 等[54]、Chang 等[55]、Rogers 等[56]
2020	新冠疫情	Bridgman 等[15]、Chang 等[55]、Ferrara 等[57]、Al 等[58]、Ferrara 等[59]
2021	美国国会山骚乱	Prabhu 等[60]
2022	俄乌战争	Shen 等[61]、Geissler 等[62]
2022	法国大选	Abdine 等[63]

(2)分析实施环节。不同研究者从组织架构、信息内容、技术工具等角度分析实施社交媒体优化行为的各环节。组织架构方面,探讨众包组织[13]、大型科技公司、情报机构以及各国网络部队[14]在社交媒体优化中的角色和表现;信息内容方面,讨论虚假信息[15]、错误信息[16]、假新闻[17]、谣言[18]的定义、创造、生产、分发的方法或检测手段;技术工具方面,分析机器人的购买渠道[19]、第三方机器人服务[20]以及使用情况[21]等。

(3)探讨应用影响。研究分析社交媒体优化的应用领域及产生的社会、经济、文化影响。①从选举等现实事件的角度出发,探讨社交媒体优化对现实

事件的意义和影响[3,22]。②从媒体和传播的角度出发,分析媒体宣传的原理方法、宣传内容和渠道工具[23]。③从经济利益驱动的角度出发,分析品牌营销、流量吸引、关注购买[24]、股市控制[25]、编撰假新闻获利等获利手段。

社交媒体优化是一个跨学科的研究领域,研究者来自不同学科背景,对问题定义、关注角度、分析对象和预期目标各不相同。根据上述总结,现有研究仅针对社交媒体优化的某一具体事件现象、优化过程的某一个环节或单一应用场景的影响展开分析,尚未形成对该领域的系统性总结,对于理解该领域研究和解决优化带来风险及危害的指导作用有限。本文通过归纳总结社交媒体优化问题的相关研究,综合计算机科学、传媒、法律和工业界对该议题不同维度的讨论,明确了社交媒体优化的概念和构成,剖析了优化的实现方式并总结了当前识别、度量和对抗恶意优化的技术及手段,旨在从更全面的视角为研究或治理社交媒体恶意优化提供参考。

为了加强社交媒体优化领域的系统性研究,本文对涉及社交媒体优化的研究点进行了系统的梳理,如图 1 所示,之后依照该脉络分章节讨论。第 2 节从概念层面明确社交媒体优化的定义、构成的三大核心要素及社交媒体信息表示;第 3 至 5 节根据优化行为拆解的优化内容、受控账号和优化活动三要素,分别讨论其定义、分类、实现方法及具体的识别技术;第 6 节从优化行为曝光规模、现实事件影响程度、社会认知影响程度、平台对抗能力四个维度分析社交媒体优化的度量方法;第 7 节从技术措施和社会治理两个角度讨论对抗社交媒体恶意优化的手段;最后总结全文,论述了社交媒体优化的挑战并给出未来研究建议。此外,本文部分研究点,如恶意优化内容识别、机器人识别、优化活动识别、优化对抗的技术措施等需要一篇或多篇综述文章才能详细论述,鉴于篇幅限制,本文选择部分经典方法代表性总结。

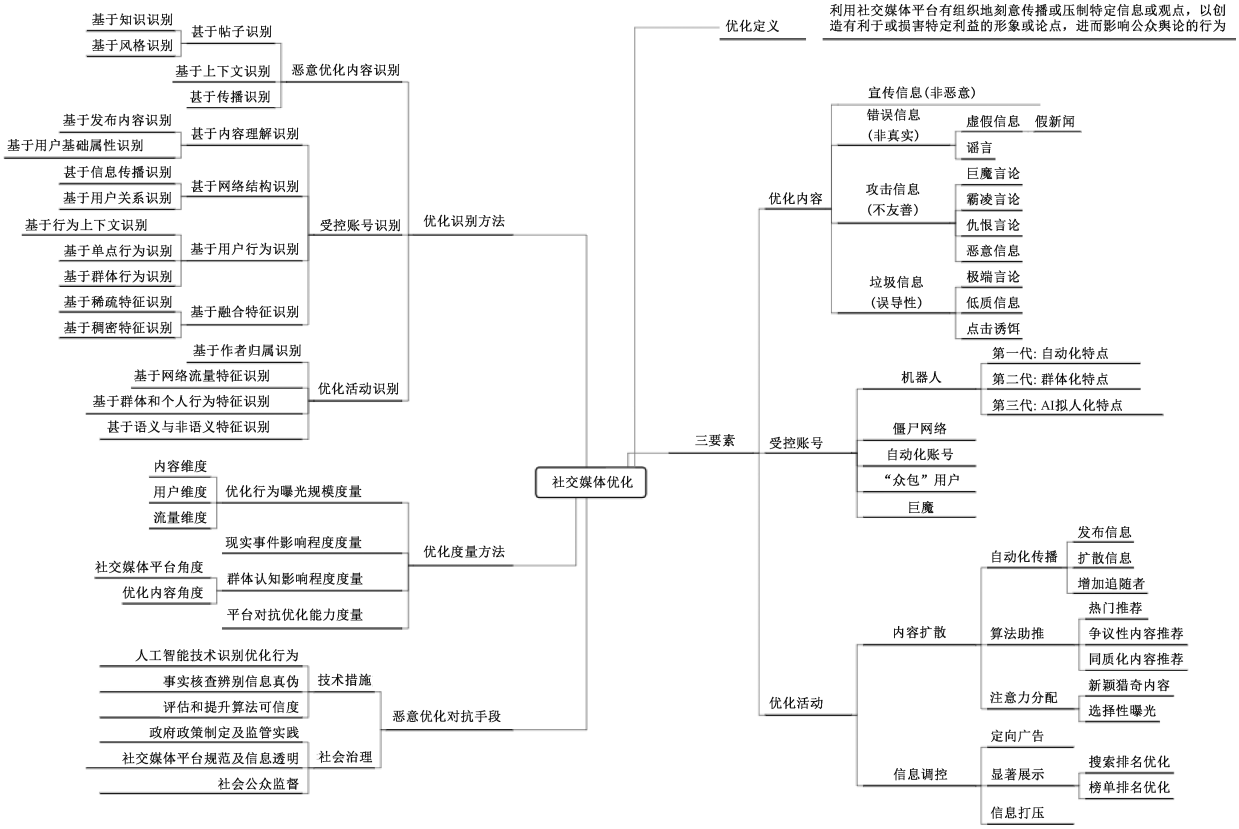

图 1 社交媒体优化研究概览图

2 定义、构成及表示

2.1 社交媒体优化的定义

由于社交媒体优化的跨学科属性,目前为止学术界尚未形成"社交媒体优化"概念的统一定义[64-65],社交媒体优化不仅是提升品牌或个人在社交平台上的曝光度和参与度的过程,还涉及通过优化内容、互动、数据分析和平台算法来增强社交媒体的整体效果。在这一过程中,社交媒体优化与控制相交织,通

过操控算法、排名、榜单和用户行为等方式来控制公众讨论或塑造特定的舆论,因此,社交媒体优化的定义应考虑其潜在的操控性,有的研究者将其定义为“利用互联网作为协作、交流和共同生产的平台,某些团体针对新闻媒体生态系统中的漏洞,提高其信息的可见度和受众范围的行为”[23]。本文综合以上概念及其他研究者的论述[66-68],认为社交媒体优化不局限于商业和品牌营销领域,将其定义为:

社交媒体优化:是指利用社交媒体平台有组织地刻意传播或压制特定信息或观点,以创造有利于或损害特定利益的形象或论点,进而影响公众舆论的行为。

2.2　社交媒体优化三要素

社交媒体优化的主要目标是创造和传播特定观念,引导公众注意力,具体行为包括创建、分发、扩散、压制、删除和控制社交媒体平台上的信息,这些信息的传播媒介是社交平台账号,传播的具体动作称为优化行为,一系列优化行为共同组成优化活动。本文将社交媒体优化涉及的要素提炼为优化内容、受控账号和优化活动,其相关关系如图 2 所示,三个要素组合起来共同开展一次社交媒体优化,部分优化活动如信息打压、个性化广告等,不需要受控账号的参与也可直接开展。通过针对各要素开展具体操作,社交媒体优化分为三个环节:(1)优化内容:优化实施者首先明确想要优化的议题,并对目标受众的特征进行分析,根据分析结果精心定制吸引人的优化内容。(2)获取受控账号:优化实施者通过自动化构建、账号购买、第三方服务等方式,组建由社交媒体机器人、网络巨魔、“众包”用户等构成的“账号军团”。(3)开展优化活动:具体包含内容扩散和信息调控两种优化手段。内容扩散指利用受控账号、社交平台的分发算法、用户注意力机制发布和传播互动。信息调控指利用定向广告、搜索引擎优化、平台过滤机制等,指向性精细化控制信息,确保目标受众能够尽可能多地接触到优化实施者希望传达的信息,同时屏蔽不希望受众接触到的信息。

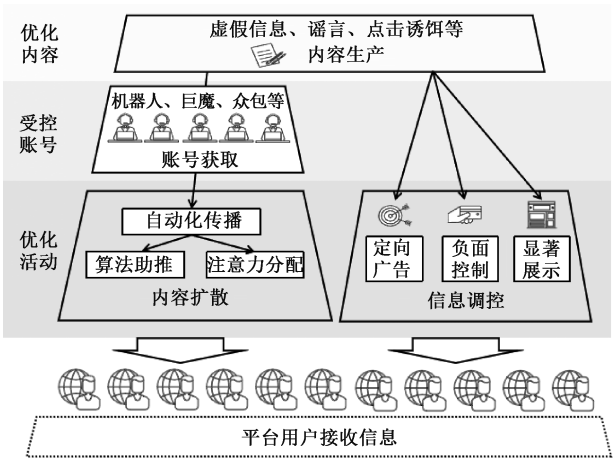

图 2　社交媒体优化三要素

2.3　社交媒体信息表示

社交媒体优化发生在社交媒体平台的信息传播过程中,为方便后续讨论,我们引入符号和公式,对社交媒体平台信息统一表示。

社交媒体信息由帖子集合 $P=\{p_1,p_2,\cdots,p_n\}$、用户集合 $U=\{u_1,u_2,\cdots,u_n\}$、行为集合 $B=\{b_1,b_2,\cdots,b_n\}$ 和关系集合 $R=\{r_1,r_2,\cdots,r_n\}$ 四部分构成:

(1) p_i 代表一条社交媒体帖子,由帖子内容 c_i 和帖子上下文信息 h_i 组成,帖子内容包含标题、正文、图片等信息,上下文信息包括发布时间、地域、帖子作者 u_j、评论内容、点赞量、收藏量、浏览量、转发量等信息。

(2) u_i 代表一个社交媒体用户,可由昵称、年龄、地域、性别等用户基本属性信息。

(3) b_i 代表一个社交媒体信息交互行为,发生在一个用户和一条帖子之间,包括发布、转发、评论、点赞、收藏等。

(4) r_i 代表两个社交媒体用户之间的关系,包括关注、收藏等。

3　优化内容

为了开展社交媒体优化,优化实施者制造能够吸引读者注意力的内容,主要目的是鼓励用户参与、讨论和分享,优化内容本身不一定都是恶意的,但如果内容创作者采取不正当手段制作内容,或创造虚构的、不存在的信息,则会产生恶意的优化内容,不同的研究者使用不同的术语表示这些内容,其中,最具代表性的就是“假新闻”。“假新闻”指的是“以新闻报道为幌子传播的虚假信息,通常是耸人听闻的信息”,被柯林斯词典宣布为 2017 年年度词汇①。除了“假新闻”之外,优化实施者们还创作“垃圾新闻”[31]“误导性言论”[69]“恶意信息[70]”等。在本研究中,本文将社交媒体优化中被创造或控制的各类信息统一称为“优化内容”。

① 2016 年美国大选期间有一条教皇弗朗西斯支持唐纳德 特朗普担任总统的“假新闻”故事震撼世界[28]。

3.1 优化内容分类

优化内容包含了一系列相关概念，表 2 总结了这些概念的具体定义及代表性文献，这些概念常常在同一语境中出现，在含义上也存在相互交叉和重叠，极易混淆，关系见图 1 中优化内容部分，它们之间的差异点为：(1)“优化内容”是所有信息的总称，重点强调的是出于社交媒体优化目的而制造的内容，包含“宣传信息”“错误信息”“攻击信息”和“垃圾信息”，除了“宣传信息”，其他 3 类信息都是恶意的。(2)“宣传信息”是为了建立用户关系、提高用户满意度而发布的健康真实内容，如引人入胜的帖子、互动性强的问答、投票等；(3)“错误信息”强调的是信息的虚假或不准确性，包括“虚假信息”“谣言”和“假新闻”。(4)“攻击信息”强调信息的不友善，包括“巨魔言论”“霸凌言论”“仇恨言论”和“恶意信息”。(5)“垃圾信息”不强调信息的真实性，而是强调信息的无关和误导性，包括“极端言论”“低质信息”“点击诱饵”等。

表 2 社交媒体中优化内容的类型及定义

名称	英文	定义
宣传信息	Promotional Information	又可称广告信息，在社交媒体中推广、宣传或推广某一产品、服务、品牌、理念或观点的内容
错误信息[16,30,49,56,59,71]	Misinformation	在社交媒体中传播的故意制造并有意或无 意传播的所有虚假或不准确信息
虚假信息[15,36,41-40,48,50,72]	Disinformation	故意传播误导人们的不准确信息
假新闻[5-7,28-29,46,73]	Fake News	故意以新闻形式传播的错误信息
谣言[18,74-75]	Rumor	未经证实或故意编造的传播广泛的信息
攻击信息	Attack Information	伤害、恐吓或损害目标对象的恶意信息
巨魔言论[39]	Trolling	旨在对特定人群造成破坏和争论，制造争端、煽动情绪的信息
霸凌言论[76]	Cyberbullying	对他人进行恶意攻击或骚扰的网络暴力言论
仇恨言论[77]	Hate speech	针对特定人群的辱骂性内容、表达偏见和威胁
恶意信息[70]	Malinformation	为骚扰目的而发布的准确信息，例如人肉搜索或发布私人详细信息
垃圾信息[31,52]	Junk Information	极端主义、耸人听闻、阴谋论、蒙面评论的总称
低质信息	Low-quality Information	缺乏深度、无价值、无意义的信息
极端言论[78]	Extreme speech	偏离主流的或歧视性的激进言语
点击诱饵[24]	Clickbait	吸引注意力并鼓励访问者点击指向特定网页的链接内容

3.2 优化内容生产方法

优化内容的具体形式包括文本、图像、音频、视频、动画等，生产方式可以通过组建专业的内容生产团队开展内容创作，如特朗普的数字运营部门，该部门包括数据科学家、图形艺术家、广告撰稿人和媒体工作者等；也可以向专业从事内容生产的第三方机构购买，如韦莱斯①和其他地方团体通过创作和销售广告以及虚假信息来获利。

随着深度伪造技术和以 ChatGPT 为代表的大语言模型的兴起，生成式人工智能广泛应用于优化内容生产中。Westerlund 等人[79]总结了使用深度伪造技术生成社交媒体内容的例子，包括美国前总统奥巴马②、Facebook 首席执行官马克扎克伯格③等公众人物被制作过深度伪造视频。2024 年 8 月，由 AI 绘画软件 Midjourney④ 制作的“特朗普被捕”“特朗普入狱”“五角大楼爆炸”“泽连斯基和普京签署和平协议”等虚假图片在社交媒体上广泛传播，展示了新一代的人工智能技术在制作虚假信息上的巨大潜力。使用人工智能技术开展优化内容生成对比过去传统的篡改技术，在效果上更为逼真且门槛更低。一方面，生成的作品难以用肉眼辨别真伪；另一方面，商业性或甚至免费的伪造服务已在公开市场上出现，普通人也具备创建专业级伪造作品的能力。这两个特点使得人工智能技术成为制作优化内容的重要生产工具。

3.3 恶意优化内容识别

恶意优化内容是发布者对信息的扭曲偏见，因此恶意优化内容识别本质上是失真偏差检测，可以被建模为一个二元分类问题。这看起来与普通的文本分类任务具有相同的设置，但在普通的文本分类任务中内容是有机的，例如体育与财经新闻在表述和形式上不相同，但是恶意优化内容会刻意让其看起来真实和准确[17]因此，不能仅仅关注文本特征，作者、热度、时间等上下文特征和内容传播网络也需

① 欧洲国家马其顿的一座小城韦莱斯以年轻人为主力专门生产假新闻，通过社交媒体平台吸引受众并从中牟利。

② https://www.cbsnews.com/news/what-are-deepfakes-how-to-tell-if-video-is-fake/。

③ https://edition.cnn.com/2019/06/11/tech/zuckerbergdeepfake/index.html。

④ https://www.midjourney.com/home。

要被考虑进来。

令 a 代表一条信息内容,a 在社交媒体上的数据可由三部分信息表示 $X_a=\{P_a,U_a,B_a\}$,社交媒体内容可能涉及多个帖子,P_a 表示内容 a 涉及帖子的集合,U_a 表示所有涉及帖子的作者集合,$B_a=\{b_{ijt}\}$ 表示内容传播过程,其中,$P_a\subseteq P$,$U_a\subseteq U$,$B_a\subseteq B$。$b_{ijt}=\{u_i,p_j,t\}$ 代表一次信息传播行为,表示发生于时间 t 用户 u_i 对帖子 p_j 的一次发布或转发,其中 $u_i\in U_a$,$p_j\in P_a$。令 F 为需要学习的检测函数,则检测目标可由以下公式表示:

$$F(X_a)=\begin{cases}1, & \text{if } a \text{ 是一个恶意优化内容}\\ 0\end{cases} \tag{1}$$

对社交媒体恶意优化内容识别的研究主要针对最常见的三类恶意信息,分别是假新闻检测[17,80]、谣言检测[18]和虚假信息检测[16]。他们的区别是不同类型信息有不同的数据集、标签和评分策略,但问题的定义和方法都非常相似,本文总结了代表性的数据集,详细见表 3 所示。恶意优化内容识别大部分采取监督学习方法,也有少量研究因为缺少数据和有效标签采用半监督学习[81-82]或无监督学习[83]方法。识别主要可以分为传统机器学习和深度学习两种技术路径,相对于机器学习,深度学习减少了对特征工程的依赖,且通常具有更好的泛化能力。

表 3　恶意优化内容识别的代表性数据集

数据集名称	样本数	标注类	平台	描述
CREDBANK[84]	62 000	2	Twitter	共 1.69 亿推文,开展了可信度标注
LIAR[85]	12 836	5	PolitiFact	1.28 万条手动标记的简短陈述
FakeNewsNet[86]	23 196	2	Twitter,PolitiFact,GossipCop	涵盖了新闻内容本身和社交上下文内容,共 2.3 万条新闻和 69 万条推文
FEVER[87]	185 445	3	维基百科	基于维基百科文章的 18.5 万条事实声明
PHEME[88]	6 425	4	Twitter	以信息传播树的方式组织的特定新闻事件相关推文
Twitter15[89]	1 490	3	Twitter	根据谣言公布网站信息发布的 Twitter 消息源,爬取相关的转发信息形成信息传播树
CHEF[90]	10 000	3	谷歌搜索	使用搜索引擎返回的文档作为原始证据,由人工标注者开展声明标注
Weibo[91]	9 528	2	微博	使用搜索引擎返回的文档作为原始证据,由人工标注者开展声明标注

根据所使用的数据类型不同,可以将恶意优化内容识别分为三类,分别是基于帖子的、基于上下文的和基于传播的。

3.3.1　基于帖子的恶意优化内容识别

基于帖子的又称做基于内容(content-based)的恶意优化内容识别,指的是使用社交媒体中发布的文本、图像和视频等来检测发布内容是否是以人为恶意影响社交媒体为目的。在使用此类方法开展识别时仅使用帖子内容,因此在表示分类任务时,公式(1)中 $X_a=C_a$,其中 $C_a=\{C_{a_1},C_{a_2},\cdots,C_{a_n}\}$,$C_{a_1}$ 涉及帖子 p_{a_i} 中的内容信息。此类方法通过收集的帖子内容和人工标注标签训练分类器,在各类型恶意优化内容的识别上均有实施,例如,Guacho 等人[82]提出利用张量表示和图中的半监督学习方法来检测误导性新闻,Yu 等人[92]基于信息熵理论检测谣言,Zhou 等人[93]提出用于假新闻检测的理论驱动模型,在内容信息有限的情况下实现对假新闻的早期检测。在检测思路上,研究者又从基于知识(knowledge-based)和基于风格(style-based)两个角度开展研究:

(1)基于知识的恶意优化内容识别。通过提取待检测内容中的关系知识,并将其与代表事实的知识库之间的"比较"来评估真实性,这个过程也被称为"事实核查"。主要的挑战是过于依赖于外部资源[17],最常使用的外部资源包括开放网络和结构化知识图谱,开放网络资源常被用作参考,可以与给定的声明进行比较,以验证其一致性[94]。利用知识图谱开展事实核查旨在核查待检测内容中的声明是否可以从知识图谱中的现有事实中推断出来[95-96]。

(2)基于风格的恶意优化内容识别。旨在捕捉和量化恶意优化内容和真实信息之间的风格差异,其假设是人造的恶意内容和真实信息在内容表达风格上存在区别[93,97],Rubin 等人[98]提出区分文本数据中真实性与欺骗性的方法,可以从内容中捕获欺骗性陈述与真实性句子之间的差异。具有误导性和欺骗性的"点击诱饵"标题常被用作识别假新闻文章的风格指标[24]。

3.3.2　基于上下文的恶意优化内容识别

基于上下文的(Context-based)恶意优化内容

识别指的是根据帖子在社交媒体中可用的上下文信息，如作者、时间、评论等来检测受恶意优化内容，在表示分类任务时，公式(1)中 $X_a = P_a$，其中 $p_a = \{p_{a_1}, p_{a_2}, \cdots, p_{a_n}\}$，$p_{a_i}$ 是 a 涉及的一条帖子，包括内容信息 c_{a_i} 和上下文信息 h_{a_i}。上下文信息通常与其他信息联合使用，或直接向量化作为附加特征使用，如 Jin 等人[99]利用其他用户在帖子内容中的观点来推断原始帖子的真实性。Zhao 等人[75]利用谣言会激起用户发推文质疑或询问其真实性这一假设，根据询问推文判断信息是否为谣言。还有研究者使用帖子的时间特征或热度特征，Kwon 等人[100]对信息的突发模式进行建模来识别错误信息，其假设是错误信息是故意传播的，因此具有不同的发布模式特点。

3.3.3 基于传播的恶意优化内容识别

基于传播的(Propagation-based)恶意优化内容识别指的是区分恶意互动和真实信息在社交媒体上的传播模式的差异来预测内容可信度，在表示分类任务时，公式(1)中 $X_a = \{P_a, U_a, B_a\}$，传播关系主要用 B_a 表示。一般使用两种方式对社交网络中的消息传播过程进行建模：

(1)通过信息级联建模。Wu 等人[101]开发了一个基于随机游走图的方法，通过计算两个跳跃式信息级联之间的相似性来识别微博平台上的谣言。Alamsyah 等人[102]使用社会网络分析(SNA)和易感感染(SI)模型对信息级联机制进行建模，通过研究传播机制发现假新闻的传播范围比真新闻更广，传播速度更快。Bian 等人[103]利用递归神经网络(RvNNs)构建基于信息级联的树状结构神经网络，用于谣言检测。

(2)使用同构网络和异构网络建模。Jin 等人[99]构建三层层次网络利用图优化框架推断事件的可信度。Gupta 等人[104]在用户－推文－事件的三层异构信息网络上设计了类似 PageRank 的算法，用于评估新闻的可信度。Shu 等人[105]利用网络捕捉新闻发布者、新闻文章和新闻传播者(用户)之间的关系，并使用线性分类器对新闻文章(假新闻或真新闻)进行分类。

基于传播的检测方法相对于基于内容和基于上下文的检测方法使用更加丰富的信息，检测效果相对稳健，但是基于传播的检测应用在恶意优化内容的早期检测时效率低下，很难在恶意优化内容大范围传播之前检测到它们。

4 受控账号

优化实施者生产的内容需要通过发布、转载、讨论等方式在社交媒体平台上传播，这些行为主要通过社交媒体账号作为媒介来实现，本文将被优化实施者操控的账号统称为“受控账号”。

4.1 受控账号分类

在参与社交媒体优化的账号中，数量最多、最引人注目且受到最广泛研究的是社交媒体机器人，但社交媒体机器人只是受控账号的一种形态。牛津大学的“民主与技术计划”组织将受控账号分为在线评论员账户、自动账户、混合账户三种[68]，龙辰一等人[106]将社交媒体优化背后使用的账号称为“技术武器”，分为社交机器人、僵尸网络、巨魔、真实事件控制、Cyborg 和黑客入侵窃取共六类。本文综合了各方研究，分析总结了最常见的 5 类受控账号类型，具体定义及相关研究如表 4 所示。不同类型的受控账号显示出不同的行为模式、自动化程度和复杂性，在目的、使用场景上也存在区别。机器人和自动化账号通常指的是自动化的软件程序，“众包”用户和巨魔则涉及真实用户的行为，僵尸网络特指被控制的计算机网络。机器人相比于自动化账号具有不同程度的复杂性和智能性，自动化账号通常设计得比较简单，只专注于执行特定任务；对比“众包”用户和巨魔，“众包”用户以团体化协作方式完成特定的工作任务，巨魔的行为更多是个体化的且基于个人意图。

表 4 社交媒体中受控账号的类型及定义

名称	英文	定义
机器人[21,33-35,37,44-45,51,57-58,61,63]	Bot	一个用于参与社交媒体的自动程序，由算法操作的自动社交媒体账号
僵尸网络[38]	Botnet	利用机器人账号模拟真实用户的行为，以达到影响舆论、诱导点击、传播垃圾信息等目的的网络
自动化账号[107]	Cyborg	使用辅助工具，自动化与手动控制集成的账号
“众包”用户[108-109]	Crowdturfing	利用人力众包平台按照悬赏者意图行动的账号
巨魔[39,110]	Troll	故意发起网络冲突或冒犯其他用户，用于分散注意力和在社交网络中制造分歧

4.2 受控账号获取

受控账号的获取分析以机器人为例展开,主要包括自主构建和购买第三方服务两种。一方面,机器人的使用者可以通过第三方 API 或自动化软件工具构建机器人,例如,Telegram 平台于 2015 年正式推出其开放机器人平台①,这使得程序员可以通过外部 API 创建自动化机器人程序。另一方面,由于验证码、电子邮件确认和 IP 黑名单等注册障碍,构建和销售机器人已经演变为商业活动,如 Fiverr② 等网站上,机器人制造商向第三方销售机器人[19],Dennis 等人通过调研机器人相关商品发现,付费的机器人服务与社交媒体平台的技术可访问性呈负相关关系[20]。优化实施者在获取机器人账号后,通过具体执行活动单元来操控账号,包括预定发布、各种形式的热度创造、按时间顺序删除[111]以及许多其他的自动化行为。

随着新技术的发展应用以及检测技术的提高,机器人策略的复杂性也相应有所提高,根据社交机器人在技术上的迭代过程,将社交机器人分为三代:(1)第一代社交机器人表现出明显的自动化行为,在结构上账号间相对独立,缺少社交链接。基于人工规则,如通过语言模板产生对话,其能力局限在简单的问答。(2)第二代社交机器人是目前社交网络中存在的主力,以大规模、群体化、社交链接稠密为主要特点。第二代机器人策略的关键不仅是模仿普通用户的表面"外观",还模仿账号在社交网络中的位置属性,因此,这代机器人账号经常成群出现并伪装成为团体或组织,形成具有复杂社交生态系统的社交媒体机器人网络[112],共同行动发挥作用,网络中每个账号都配备了不同的角色和常规性的社交行为,并与周围的环境互动[113]。(3)第三代社交机器人则更多地加入了人工智能技术,如使用生成式人工智能技术生成差异化文本和高可信图片,使用 AI 发现活动目标、生成特定对话内容提高对目标受众的精准影响等,这些技术的应用极大增强了社交机器人账号的拟人程度,使得它们更加难以识别。

4.3 受控账号识别

受控账号是指受到社交媒体操控者操控的非人账号识别,受控账号的假设是相比于真人用户,它们在个人属性和行为特征上存在差异性,因此受控账号检测也常被视为一个分类任务。

令 b 代表一个社交媒体账号,b 在社交媒体上的数据可由 $X_b=\{u_b,P_b,B_b,U_b,R_b\}$ 这 5 部分信息来表达,u_b 表示用户 b 的基本属性,P_b 表示用户 b 涉及帖子的集合,$B_b=\{b_{itk}\}$ 表示用户 b 行为的集合,$U_b=\{u_j\}$ 表示与用户 b 有社交关系的用户的集合,和 $R_b=\{r_{jl}\}$ 共同构成用户 b 的社交网络。其中,$u_b\in U$,$P_b\subseteq P$,$B_b\subseteq B$,$U_b\subseteq U$,$R_b\subseteq R$,$b_{itk}=\{p_i,t,k\}$ 表示用户 b 在时间 t 对帖子 p_i 一次类型为 k 的行为,其中 $p_i\in P_b$,k 可以是传播、点赞、评论等;$r_{jl}=\{u_j,l\}$ 表示用户 b 与用户 j 的类型为 l 的关系,其中 $u_j\in U_b$,l 可以是追随、关注、被追随、被关注等。令 F 为需要学习的检测函数,则检测目标可写为下面公式所示的二分类任务:

$$F(X_b)=\begin{cases}1, & \text{if } b \text{ 是一个受控账号}\\ 0\end{cases} \tag{2}$$

对于不同类型受控账号的识别问题,它们的定义和表述是基本一致的,使用方法的差异性取决于不同类型账号具备不同特征和不同行为模式,如巨魔检测[114]、僵尸网络检测[115]、"众包"账号检测[116]、水军检测[117-118]等。但最常见且研究最丰富的受控账号检测技术是针对社交媒体机器人[119],本章节的检测方法主要以社交媒体机器人为对象展开分析,本文总结了机器人识别的代表性数据集,详细见表 5 所示。机器人检测识别大部分采取监督学习方法,也有部分使用无监督学习方法,其假设是机器人账户是批量生成的,它们会与真实用户保持距离。

表 5　机器人识别的相关数据集

数据集名称	平台	真人账号数量	机器人数量	账号总数	博文总数
gilani-2017[120]	Twitter	1 394	1 090	2 484	0
cresci-2017[121]	Twitter	3 474	10 894	14 368	6 637 615
midterm-18[122]	Twitter	8 092	42 446	50 538	0
cresci-stock-2018[25]	Twitter	6 174	7 102	13 276	0
botometer-feedback-2019[123]	Twitter	380	138	518	0
TwiBot-22[124]	Twitter	860 057	139 943	1 000 000	88 217 457

根据所使用的数据类型,受控账号的识别方法可以分为四类:基于内容理解的方法、基于网络结构

① https://telegram.org/blog/bot-revolution。
② http://fiver.com。

的方法、基于用户行为的方法和基于融合特征的方法。

4.3.1 基于内容理解的受控账号识别

基于内容理解的受控账号识别指的是通过分析用户发文内容和用户属性特征开展识别，在表示分类任务时，公式(2)中 $X_b=\{u_b,P_b\}$。方法上与恶意优化内容识别的方法相似，但落脚点放在账号层面。例如，Mukherjee 和 Venkataraman 等[125]提出了基于内容生成模型的文本检测方法，通过使用不良文本生成模型生成文本并聚类，发现发布不良信息的账号群体。Clark 等人[126]提出了基于自然语言处理的自动推特账号识别方法，主要使用了文本内容作为特征，包括推特的 URL 数、推特对词汇的差异性等。除了使用发布内容，用户昵称、头像、个人资料、IP 等属性信息也常被用于识别受控账号，这类方法的假设是受控账号，尤其是机器人账号是批量自动化生成的，因此与真实账号保持一定的差异。例如，Webb 等人[127]尝试通过用户的个人资料信息发现高度相似的机器人账号。沈一和鲍新平提出了一种基于昵称开展水军检测的方法，通过计算昵称的 minhash 值来衡量昵称的相似性并检查是否符合水军账号的特征。Koosha 等人[128]提出了一种深度神经网络架构，通过个人资料和帖子特征来预测用户发布内容是否属于机器人生成。

4.3.2 基于网络结构的受控账号识别

基于网络结构的受控账号识别指的是利用账号关系和信息传播这两类网络信息开展识别，在表示分类任务时，公式(2)中 $X_b=\{P_b,B_b,U_b,R_b\}$，用户 b 的社交关系网络用 $\{U_b,R_b\}$ 来表示，信息传播网络用 $\{P_b,B_b\}$ 来表示。大多数研究使用账号关系网络作为数据源，例如 Maxim 等人[129]使用朋友关系图来识别机器人账号，并采用多种图算法计算图指标。Feng 等人[130]提出了一种多重异构图卷积神经网络模型，考虑不同的互动关系(如关注和被关注)，构建了多个图卷积网络，通过传播节点的网络信号来检测社交机器人。另外，一些研究者利用信息传播网络开展识别。例如，陈侃等人[131]在新浪微博的数据上建立了交互行为的信息传播模型，根据不同传播主题间的交互定义特征，开展水军传播检验。

4.3.3 基于用户行为的受控账号识别

基于用户行为的受控账号识别是指通过受控账号的行为模式开展识别，其中，根据使用不同的行为进行对比，可以将其分为三类，分别是基于行为上下文、基于单点行为和基于群体行为。在表示分类任务时，在基于行为上下文、单点行为时，仅关注用户 b 的行为序列，公式 2 中 $X_b=B_b$，在基于群体行为时，关注数据中全量用户的行为，公式 2 中 $X_b=\{u_b,B\}$。

(1)基于行为上下文的受控账号识别。是指使用账号的历史行为信息，发现与用户行为模式有明显差异的突发性行为。这类方法被称为基于行为上下文的受控账号识别。例如，Chu 等人[107]发现受控账号经常用于发布与特定主题相关的信息，其发布行为模式包含长时间的休眠和突发峰值，因此提出了一种利用时间特征发现恶意账号的方法。

(2)基于单点行为的受控账号识别。是指通过分析用户的某些非常规行为模式来识别单独的机器人账号。这类方法的假设是受控账号的行为模式存在相似性，且与正常账号存在差异。例如，Rajendran 等人[132]使用双向 LSTM 网络研究推特机器人的时间模式和行为模式，通过推文的速率和频率来区分机器人账号和正常账号。

(3)基于群体行为的受控账号识别。是指通过分析受控账号群体的一致性或交互性行为来进行识别。这类方法的有效性来源于受控账号往往在群控软件控制下协同工作。例如，Chavoshi 等人[133]认为人类不可能长时间高度同步地完成某些活动，提出了一种名为 Debot 的检测方法，基于相关性扭曲原理和滞后敏感哈希技术对机器人聚类识别。Zhao 等人[134]通过学习机器人之间的交互行为模式，提出了一种基于多属性异构的图卷积神经网络模型(BotAHGCN)，用于检测社交机器人。

4.3.4 基于融合特征的受控账号识别

近年来，研究者们往往不使用单类数据，而是使用多类特征共同开展受控账号检测，此时，与用户 b 相关联的全部数据都可以被同时使用，在表示分类任务时，公式(2)中 $X_b=\{u_b,P_b,B_b,U_b,R_b\}$。从实现角度来讲，可以分为基于稀疏特征和基于稠密特征的方法。

(1)基于稀疏特征的受控账号识别。是指对事先获取的数据进行处理后，选择一些具有代表性和判别性的特征开展识别，以达到更好的效果。例如，Twitter 机器人检测的公共工具 Botometer 提取账户的网络、用户、朋友、时间、内容、情绪共六组主要特征，使用 RF 来检测账号是否是一个机器人[135]。Alothali 等人[136]回顾了当前使用的各种机器人检测方案，分析对比了各方法的使用特征和验证分类

器的评估技术。

(2)基于稠密特征的受控账号识别。是指在不设置和提取明确特征的情况下开展受控账号检测,简化了特征选取过程,多采取深度学习的方法,Wang 等人[137]利用点击流模型来检测服务器端社交账户的真实身份,这些作者输入点击流序列,然后计算序列距离以分类社交账户。Shi 等人[138]提出了一种检测恶意社交机器人的方法,包括基于点击流序列转移概率的特征选择和半监督聚类,该方法分析了用户行为点击流的转移概率和时间信息。

5　优化活动

优化活动是在优化内容生产、受控账号获取及使用的基础上开展的一系列社交媒体优化行为,研究者对社交媒体优化活动的称呼方式并不相同,有宣传活动[139]、计算式宣传[140]或人造草皮[141]等。各大社交媒体平台也有自己的称呼方式,如 Facebook 将其称为"协调的不真实活动(CIB)"①,Twitter 将其称为"平台操纵和垃圾邮件"②。

优化活动的具体手段又分为内容扩散和信息调控。内容扩散和信息调控同属优化活动,但它们之间还存在明显的差异性,扩散强调信息传播的发散性,无明确的指向,主要通过平台上信息交流达成。信息调控则强调优化活动的指向性,涉及使用合法或非法的第三方服务进行定向的信息投放、展示或打压,目的是进一步提高优化内容的曝光,降低和控制其他对立观点,主动引导社交媒体讨论议题和公众认知。

5.1　内容扩散方法

研究者发现人类倾向于相信不同来源但反复出现的信息,可以通过重复出现来提高信息在人们心中的准确性感知[142],因此大规模的信息曝光成为优化活动的目的之一,也是衡量优化效果的重要指标。通过优化活动达到信息曝光的有效途径是内容扩散,它主要得益于三种机制,分别是自动化传播、算法助推及用户注意力分配。自动化传播指的是通过控制受控账号主动开展内容铺陈,往往发生在优化活动的初期,发展到中后期,社交媒体平台算法和用户注意力机制在内容扩散上也起到积极的推动作用。

5.1.1　自动化传播

自动化传播指优化实施者通过受控账号大量发布优化内容以达到其优化目标,大量关于社交媒体机器人检测、行为分析、信息传播规律分析的研究都围绕这一现象展开。受控账号自动化传播信息的方式可以总结为发布信息、扩散信息和增加追随者三种具体行为:

(1)发布信息。是指通过控制有影响力的社交媒体账号发布具体优化内容,它是最原始最初级的内容扩散手段。例如,研究者在分析 2016 年美国总统大选期间社交媒体信息传播规律时发现,社交媒体中存在大量人为扩大候选人或相关议题流量的现象,以制造共识性意见[35]。在分析信息传播的网络结构、动态特征和主要传播者时发现,Twitter 上的机器人通过回复和提及目标用户,负责了错误信息的早期传播[30]。

(2)扩散信息。是指通过浏览、点赞、转发、评论特定内容,增加信息热度和可见度,引发讨论和形成高频词,进一步提升优化内容的曝光度,并将其塑造成主流民意的代表。如 Bessi 等人[34]认为在美国总统大选期间,机器人被用于支持了两位候选人,但支持特朗普的机器人比希拉里更多,且呈现比整体辩论更积极的情绪。CorpusOng 等人[110]在菲律宾大选的背景下开展研究,展示了等级化的专业优化团队如何使用受控账号以"强调"的方式,个性化和灵活地开展优化活动的案例及其具体细节,证明了这种现象的普遍存在,机器人大军通过追踪和转发候选人的相关内容,帮助其观点在公众中获得公信力和更广泛的支持,制造出其拥有大量在线支持度的假象,这些假象可能引发连锁反应,包括引导线下民众对其表示实际支持。

(3)增加追随者。是指通过虚假"涨粉"增加特定用户的关注者数量,提升受关注程度,赋予其更高的"虚荣指标"和象征意义,比如 Ouya 等人[143]分析数据发现印度总理莫迪的 4590 万 Twitter 粉丝中有 60%是购买的水军账户。

5.1.2　算法助推

社交媒体平台的算法也是导致优化内容快速扩散的重要因素,主要原因可能是社交媒体平台算法优先推送流行内容而非可信内容,同时内容倾向性和用户偏好也是算法推荐的重要考量因素。

(1)热门推荐。平台算法根据计算同时段爆发

① "协调的不真实活动"定义为"为实现战略目标而控制公众辩论的协调努力,虚假账户是运营的核心,在每一种情况下,人们都相互协调并使用虚假账户误导他人关于他们是谁以及他们在做什么"。

② 平台操纵和垃圾邮件"定义为"使用多账号和协调的方式人为地夸大自己或他人的关注者或参与度"。

的相关讨论情况来判断事件等级，优化内容经过受控账号的自动化传播，其“短频快”“发酵迅速”特点，导致优化内容快速成为被平台推荐算法感知的“新热”内容，从而加大其在推荐信息流中的推送比重，同时社交媒体平台也通过功能侧提升“新热”内容的曝光，如新浪微博推荐展示区域的推荐理由、搜索框实时展现热点事件等。

(2)争议性内容推荐。除信息本身的热度外，信息内容偏向性也影响推送结果。研究表明，仇恨言论[144]、虚假信息[145]以及愤怒情绪都能使用户在社交媒体平台上停留更长的时间，停留时长这一最主要的推荐算法优化目标导致此类信息在社交媒体平台泛滥。Mozilla 的一项研究证实，YouTube 不仅展示，而且积极向用户推荐违反平台政策的视频，这些内容涉及政治和医疗相关的虚假信息、仇恨言论和不当内容等。据内部人士透露，Facebook 的新闻推送算法赋予引发愤怒情绪的内容的权重是唤起快乐情绪的内容权重的 5 倍，Facebook 前雇员 FrancesHaugen 揭露了 Facebook 内部的研究文件，并指责 Facebook 为实现盈利目标，故意使用算法放大仇恨言论。

(3)同质化内容推荐。当前社交媒体平台的个性化推荐算法持续将相似兴趣的内容推荐给用户，促成了互联网社区的“回声室”效应[146]。一方面社交平台让志同道合的个人更容易找到彼此，公民更多地接触到强化其观点的信息，另一方面推荐算法的过滤效果会产生过滤气泡[147]，让用户没有能力选择看到的内容，相近声音不断重复，用户被平台算法进一步裹挟。

5.1.3 注意力分配机制

优化内容的早期传播依赖受控账号，但在传播的中后期，真实用户大量参与到内容的转载和讨论中，通过分析优化事件在社交媒体的传播规律和真实用户消费信息的特点，猎奇心理和选择性曝光这两种用户注意力分配机制也影响着优化内容在社交媒体的扩散。

(1)新颖猎奇内容更容易吸引人们的注意力。优化内容通常由一群精通心理学的优化实施者精心制造，在无法有效辨别其真伪的情况下，这些内容天然具备传播优势。一项对 Twitter 真假新闻传播的全面调查表明，假新闻比真实新闻传播得更快、范围更广，研究者认为这主要是由于假新闻的新颖性所具有的吸引力[145]，这也是将优化内容设计得生动、引人注目、质量上乘的原因。

(2)选择性曝光是指用户倾向于坚持和强化与他们世界观一致的信息，而忽视不同的信息，优化内容的倾向性和引导性导致用户陷入被制造的“信息陷阱”中，反复消费某一类信息。例如，Michela 等人[148]在分析“英国脱欧讨论”议题时，发现关注脱欧进程的 Facebook 用户分为两个不同的群体，每个群体的注意力都限制在特定的页面上，研究者认为这种模式就是选择性曝光和确认偏差导致的。对美国大选的研究也发现了相同的结论，Andrew 等人[29]发现假新闻消费主要集中在一小群人中，几乎 60％的假新闻访问来自 10％的最保守人群，社交媒体增加了对事实可疑但态度一致的信息的消费。

5.2 信息调控策略

信息调控的策略种类繁多，一直在持续进化中，本文重点介绍三类最常见的信息调控手段：定向广告、显著展示和信息打压。

5.2.1 定向广告

数字广告是互联网平台中收费的优化内容，它们在搜索引擎中以“赞助”搜索的形式出现在搜索结果的顶部，在社交平台中以“推广内容”的形式出现在首页或推荐信息流中。与其他优化技术相比，定向广告是合法且常见的行为，可直接向社交媒体平台或广告商购买。它的优势在于通过消费用户准确定位，实现精准的内容投放，方式有以下三种：(1)使用如人口统计、兴趣、位置或行为等用户特征发现目标用户；(2)直接给定目标用户列表；(3)自动发现给定用户列表的相似群体。

广告已被证实在商业议题的优化中大量使用，剑桥分析公司被指认非法使用超过 8000 万 Facebook 用户的信息来开发微目标广告。

5.2.2 显著展示

将信息置于显著位置也是一项有效提升信息的曝光量和展示效率的策略。常见的手段包括作用于搜索结果的搜索排名优化和作用于热点榜单的榜单排名优化。

(1)搜索排名优化是一种网络营销途径，早期主要作用于搜索引擎，社交媒体中的排名优化方法与搜索引擎一脉相承，具体包括搜索优化和竞价排名，这两种都属于合法的网络营销，但也存在提升排名的作弊手段，称为“黑帽搜索优化”，指的是人为提升内容在搜索结果中位置的非法策略，包括“偷换页面”“链接工厂”“假链接”“网页劫持”等。此外，“众包”平台也提供排名优化服务，通过工作人员搜索某个关键字，然后点击指定的链接提高社交媒体内容

的排名[13]。

(2)榜单排名优化包括平台方对榜单排名进行恶意的人工干预,以及利用受控账号进行有组织攻击,迫使非自然流量话题上榜[149]的行为。最终目标都是让优化目标内容登上榜单、在榜单留存,并利用榜单的扩大效应以吸引更大的流量,如在 Twitter 趋势榜中发现的 astroturfing 攻击,通过协调和不真实的活动人为提升选定的关键字或主题,使其看起来受欢迎最终登上榜单。Elmas 等人检测到超过 108 000 个账户推动的超过 19 000 个虚假榜单话题,最终得出 20%的全球 Twitter 趋势榜话题被人为影响的结论[150],这个比例在新浪微博中是 49%[151]。

5.2.3　信息打压

虽然社交媒体优化主要通过内容的重复、扩散和强调来扩大具体优化目标的影响范围和程度,但通过对其他对立内容的打压或删除也是优化的重要手段。有些“黑灰产”服务商提供信息打压服务,它们通过“口碑营销”“舆情优化”“负面压制”“形象维护”等名义开展宣传,往往采取混合手段实施,包括反复投诉、伙同平台内部工作人员后台操作、反复联系发帖人骚扰利诱、发布正面信息“下沉”“压制”、黑客攻击等多种手法达到删帖和控制影响的效果。

当社交媒体平台作为优化实施者时,负面信息的控制变得容易。通过“灵活可控”的社区标准、内容审核以及内容过滤算法共同配合,可以轻易地封禁特定信息或账号。例如 2019 年 8 月 19 日,三大社交媒体平台 Twitter,Facebook 和 YouTube 同一天以“中国官方散布假新闻”为由,封停了近千个揭露香港暴徒行径的账号。2024 年 8 月 26 日,Meta 首席执行官马克扎克伯格声称,Facebook 在疫情期间受到拜登政府的“压力”,审查和控制与新冠疫情相关的内容。类似的信息打压现象也存在于国内社交媒体平台,新浪微博曾因为限制微博流量并禁止评论而因此受到监管部门的处罚①。

5.3　优化活动识别

社交媒体优化活动包括信息扩散和调控,往往不是独立存在的某一例优化行为,而是多个单独个体行为共同组成,可被理解为系列化的优化行为。优化活动识别使用综合性的技术,账号识别和内容识别方法都被广泛应用其中。研究者过去更多关注自动化账户,然而近期研究表明,优化活动涉及的账号并非全部都是机器人,而是包括一部分由真实人类操作的账号,因此,研究者开始意识到优化活动的识别应该从简单的账号检测转向对协调策略的识别,即识别多个账号之间的协调行为[152]。优化活动识别可以看作一个异常检测问题,目标是在社交媒体平台全量数据中发现一个具备共同目标的优化行为的集合,这些关联的异常行为集合被称作“集体异常”。这种系列化行为可以被发现识别的假设是自动化和不真实的行为与人类驱动的合法行为在模式和特征上有所不同,因此大量具有高度相似性的行为是优化活动的警示信号。

用 c 表示一个优化活动,令 F_γ 为需要学习的检测函数,将所有的社交媒体信息作为输入数据,寻找一个子集 X_C 来表示优化活动 c,X_C 具体包括与优化活动 c 相关的优化内容集合 P_C,受控账号集合 U_C 和优化行为集合 $B_C=\{b_{ijtk}\}$,$U_C\subseteq U$,$P_C\subseteq P$,$B_C\subseteq B$,其中,$b_{ijtk}=\{u_i,p_j,t,k\}$ 表示用户 u_i 在时间 t 对帖子 p_j 一次类型为 k 的行为,其中 $p_j\in P_C$,$u_i\in U_C$。优化活动检测任务表示为

$$X_C=\{\{U_C,P_C,B_C\}\mid F_\gamma(U,P,B,R)\}\quad(3)$$

对于部分有已知特点或基于已有标注数据的优化行为检测,可以使用有监督学习方法实现活动识别,此时检测函数 F_γ 中 γ 用于描述优化行为和非优化行为之间的差异性,它可以是描述优化行为的特征规则,也可以是通过学习标注数据后得到的向量表示。在更多情况下,使用无监督或半监督的优化活动检测方法,以克服训练数据集不足和有监督检测器的泛化缺陷[153],此时检测函数 F_γ 中 γ 表示离群群体检测条件,用于识别一个或多个满足条件的行为集合。工程实践中常将规则、无监督和有监督方法组合发现可疑的优化活动。例如,基于网络图的技术可以检测可疑账号链接模式[154-155],无监督学习中的异常检测可以发现账号组的异常行为[156],通过计算账号活动时间序列之外的距离度量可以发现账号组在推文和转发行为中的异常模式[115]等。现阶段,优化活动检测还面临从通用机器学习算法到专门设计的用于检测某次特定协调活动的临时性算法的转变。因此,重点也从传统的特征工程转向学习有效的特征表示和设计全新的定制算法[157]。优化活动识别方法根据其使用的特征不同分为了四类,分别是作者归属、网络流量特征、群体和个人行为特征、语义和非语义特征[141]。

5.3.1　基于作者归属的优化活动检测

基于作者归属的优化活动检测指的是通过分析账号的个人资料和评论等信息来推断优化活动的作

① 国家网信办指导北京市网信办依法约谈处罚新浪微博. www.cac.gov.cn/2020－06/10/c 1593350719478753.htm。

者，Peng 等人[158]收集了社交媒体账号的个人资料和评论，对个人资料所做的评论进行了位级 n-gram 分析，然后执行 KNN 分类，分析作者归属。他们的研究在两个不同的新闻网站上揭示了一个可能的 Astroturfing 案例。

5.3.2 基于网络流量特征的优化活动检测

基于网络流量特征的优化活动检测指的是通过监测账号的信息传播和互动流量，识别优化行为的流量模式，假设是优化行为可能表现出不自然的信息扩散速度。Liu 等人[109]利用网络上的信息传播模式来检测微博客环境中的“众包”活动，作者提出 DetectVC 算法，通过使用基于网络图的模型来检测自愿追随者，利用网络图的邻接矩阵检测“购买追随者”活动。

5.3.3 基于行为特征的优化活动检测

基于行为特征的优化活动检测指的是通过分析个人用户行为指标和特定用户组的行为模式来识别优化活动。这类方法与受控账号检测中基于用户行为的检测方法类似，但使用的场景略有差异。Lee 等人[108]对 Fiverr 和 Twitter 中的“众包”行为开展分析，通过从 Fiverr 网站上抓取的众筹任务，根据人口统计特征和内容特定特征，利用机器学习算法对活动进行分类，研判用户是否属于“众包”工作者。这种分析方法揭示了优化活动中的一致性行为模式，为检测和识别优化活动提供重要线索。

5.3.4 基于语义和非语义特征的优化活动检测

语义特征指的是内容的含义和表达方式，方法与恶意优化内容识别中基于内容的方法相似，非语义特征包含账号属性、账号活动模式、社交关系等属性特征。Varol 等人[139]使用监督学习框架利用数百个随时间变化的特征来捕获 Twitter 趋势榜话题不断变化的网络和扩散模式、内容和情感信息、时间信号和用户元数据，区分流行是有机的还是通过宣传活动推广的，Keller 等人[159]研究了韩国国家信息服务局在 2012 年韩国总统大选期间的虚假宣传活动，最能将受控账户与普通用户区分开来的特征是 Astroturfing 活动之间协调留下的痕迹，包括转发网络、共同发文网络和共同转发网络等时间协调特征和消息协调特征。

6 社交媒体优化度量方法

在识别社交媒体优化的基础上，测量和评估社交媒体优化的影响范围、程度及对抗能力也是社交媒体优化研究的方向之一，但由于数据缺失、道德风险、度量准确性难以评价等原因，目前开展的度量研究相对有限。经汇总已有研究，本文将优化度量分为四个角度，分别为优化行为曝光规模度量、现实事件影响程度度量、群体认知影响程度度量和平台对抗优化能力度量，并汇编了针对社交媒体优化度量的相关文献，详情如表 6 所示，接下来从四个角度分别分析当前进展及面临的难点问题。

表 6 社交媒体优化度量相关研究

评估对象	数据背景
假新闻的覆盖程度[29]	2016 年美国总统选举（调查回复与假新闻网络流量）
假新闻和错误信息的覆盖范围[160]	欧洲虚假信息网站流量数据
Twitter 话题中的被调控流量[162]	2017—2018 年 Twitter 热搜词
Twitter 用户对错误信息的暴露程度[161]	5000 名 Twitter 用户及其推文数据
Facebook 中动员信息对用户行为的影响[163]	2010 美国总统选举（Facebook 中动员信息）
俄罗斯的恶意活动对用户态度和投票行为的影响[8]	2016 美国总统选举（Twitter 调查数据与 2016 年竞选期间的社交媒体数据）
竞选广告对候选人好感度和投票的影响[164]	2016 年美国总统选举（对 34000 人投放 49 个选举广告的用户数据）
假新闻对受众信息行为的影响[165]	两个假新闻消费数据：(1)希拉里向伊斯兰国出售武器，(2)奥地利总统候选人 Alexander 患癌和痴呆症
虚假、极端内容的比例和潜在的意识形态两极分化情况[166]	2016 年至 2020 年十亿 Twitter 数据
假新闻曝光对认知的影响[142]	招募志愿者参与对假新闻准确性看法的实验
社交媒体平台的回音室效应[167]	在 4 个社交媒体平台上的有争议的话题数据
社交媒体平台导致“回音室”效应[168]	欧盟政治案例（欧洲委员会的 Eurobarometer86.2 调查数据）
社交媒体中观点分歧导致的两极分化[169]	美国党派人士的态度变化
社交媒体平台内容推广算法在“回声室”出现中的作用[170]	不同平台（YouTube 和 Facebook）中相同内容的用户交互行为数据
“回音壁”内信息的接触程度对认知和意识的影响[9]	Facebook 在 2020 美国大选的对照组实验数据（23377 用户）
社交媒体平台对抗优化的能力[171]	从俄罗斯优化服务商购买的优化服务在不同社交媒体平台的表现

6.1 优化行为曝光规模度量

曝光规模度量研究最早可追溯到纸质和电视媒体时代，但相较于传统的纸质媒体和电视媒介，社交媒体的信息传播方式更为复杂，因此在评估社交媒

体的曝光度和影响规模时,需在原有统计度量方法基础上综合考虑付费与非付费媒体、个性化推荐、排名算法以及定向广告等多种因素的综合影响[145,34]。现有研究主要针对某一具体优化事件的内容、用户和流量三个维度展开分析。

(1)内容维度的度量。统计优化内容的曝光规模,Fletcher 等人[160]收集了法国和意大利最受欢迎的假新闻和虚假信息发布网站的使用统计数据,以此衡量假新闻和虚假信息在欧洲的影响规模,统计数据包含网站的覆盖范围、被关注程度,以及在 Facebook 上的互动次数等。与法国和意大利新闻媒体的等效数据对比后,得出结论为基于当前的证据,虚假新闻的影响范围比预期小,这一发现与社交媒体影响美国选举事件的研究结果基本一致[28-29]。

(2)用户维度的度量。与内容维度度量的研究不同,用户维度度量是通过用户信息评估用户接收到的优化内容占比及范围。Mohsen 等人[161]检查了 Twitter 用户在多大比例上关注了发布虚假或不准确声明的重要账户(这些声明的准确性基于 PolitiFact 事实核查网站的评级),从而评估 Twitter 用户对来自公众人物和组织的误导信息的曝光程度。

(3)流量维度的度量。从流量维度上开展度量不明确分辨哪些内容是优化内容,而是将流量的异常波动看作优化活动中的流量调控并开展度量研究。例如,Nimmo 等人[162]将流量控制定义为"少数用户试图产生大量流量,与涉及的用户数量成比例不符",并提出了流量影响系数指标来计算给定 Twitter 流量受影响程度的方法。

然而,优化行为曝光规模的度量工作依托于社交媒体数据来开展,社交媒体数据的获取遇到数据访问限制、数据丢失、私密数据获取困难等问题的困扰。一方面受用户的隐私保护政策约束,难以获取内容传播情况和内容涉及账户的完整数据,同时有研究者认为社交媒体平台粗暴地删除优化行为的证据,使得独立评估和审计变得不可能,因为社交媒体平台可能大量删除被识别为恶意优化内容和恶意账户的信息①;另一方面还存在用户独立于主站点分享、通过私人渠道传递的信息无法获取,导致优化内容在社交媒体平台的覆盖范围实际上远高于互动数据所能检测到的范围,数据难以获取成为度量工作的最大限制。

6.2　现实事件影响程度度量

社交媒体优化对现实活动,特别是选举活动的过程和结果进行影响程度分析,本文称为现实事件影响度量,它是四个度量角度中研究相对深入和广泛的。然而,这些研究得出的结论存在较大的分歧。例如,关于俄罗斯通过社交媒体优化对 2016 年美国总统大选产生的实际影响,至今仍是学术界争议的焦点。一些研究者认为,由于与俄罗斯相关的虚假信息和假新闻的传播规模相对较小[28,6],且说服力有限[164,8],所以这些在社交媒体上的优化行为并未改变选举结果。然而,另一些研究者如 Jamieson 等则认为,俄罗斯的网络恶意活动成功为特朗普赢得了更多选票,最终导致其获胜[172]。在巴西、瑞典和印度的选举,以及英国的脱欧公投等其他投票中也存在类似的争论。

近年来,研究者们持续提出衡量社交媒体影响现实世界的新方法,不断有新的评价框架和社会实验出现。例如,Aral 等人[173]提出如何衡量社交媒体改变用户购物、阅读和锻炼习惯的方法。Bond 等人[163]在 2010 年美国总统竞选期间对 6100 万 Facebook 用户发送了动员信息,开展了随机对照试验,结果表明这些信息直接影响了数百万人的立场表达、信息搜索和投票行为,可能为选举增加了数十万张额外选票。Coppock 等人[164]通过对 34000 人进行的 59 个独立实验,分析了竞选广告的影响,得出对候选人影响较小的结论。Aral 等人提出了[66]衡量社交媒体对选举结构影响的研究框架,该框架包括评估信息内容和范围、评估信息接收者和曝光程度、评估因果行为变化、评估对投票行为的影响等几个步骤。然而,评估网络空间的优化行为对现实事件的影响程度依然面临一些挑战。首先,社交媒体数据是个性化的,传统媒体的评估方法很难直接应用于社交媒体数据;其次,数据获取受到隐私法规的限制,无法获取用户消费内容的数据,公民的现实行为也同样难以获取,如投票结果;最后,社交媒体优化并非只在一个平台上进行,用户也会接触到来自不同平台的信息。因此,只有推动社交媒体平台的合作及公共事务数据的公开,才能更有效地获取各类数据,为这项研究提供全面的视角。

6.3　群体认知影响程度度量

当前社交媒体成为信息主要传播阵地,同时其开放性分布式信息传播模式和算法技术的广泛参与改变了人们接收信息的方式。社交媒体对认知的影响度量一般指的如何衡量社交媒体本身及社交媒体

① https://help.Twitter.com/en/rules-and-policies/platformmanipulation。

上的优化行为在加剧认知两极分化中发挥的作用。目前对认知影响的度量集中在两个视角，一个是社交媒体平台对认知的影响程度，另一个是优化内容对认知的影响程度。这两个视角都采取相对一致的方法，结合实验设计和观察法，通过设置不同的实验分组，以统计分析方法对比不同分组策略下群体的认知情况。

(1)从社交媒体平台视角看，分组策略主要有两种，一种是社交媒体平台用户数据和其他用户数据，另一种是社交媒体平台中争议性话题数据和非争议性话题数据。但是根据研究结果，社交媒体平台对认知是否有影响尚无定论。例如 Cinelli 等人[167]通过比较分析了来自 Gab，Facebook，Reddit 和 Twitter 等平台上的超过一亿篇争议性话题(如枪支管制、疫苗接种、堕胎)内容，研究平台上用户互动的差异，并通过衡量互联网中信息的同质化程度和信息扩散的倾向性两个主要因素来定量化社交媒体上的“回音室”现象，实验结果证实了社交媒体上“回音室”现象的存在。Suhay 等人[169]也开展了两项实验研究，发现网络环境中的观点分歧暴露确实会加剧两极分化。但也有研究者持相反的观点[9]，Ngyuen 等人[168]利用来自 28 个欧洲国家的调查数据发现，相比于依赖其他新闻来源的公民，通过社交媒体网站获取政治新闻的公民并没有表现出更明显的立场两极分化。

(2)从优化内容视角看，分组策略为社交媒体平台上被优化内容影响的用户数据和其他用户数据对比。根据研究结果，恶意优化内容对认知产生影响的结论相对统一，普遍认为社交媒体优化能够通过“回音室”效应放大负面信息和极端主义思想，最终影响群体认知。以 Zimmer 等人[165]的研究为例，他们通过分析在线评论和回应，研究假新闻对受众信息行为的影响，研究结果发现，大约三分之一到五分之二的评论可被归类到“回音室”内，这种行为直接导致了认知偏差的产生。Mosleh 等人[161]通过分析错误信息的共享者和共享网络发现了“虚假回音室”的存在，即那些频繁接触错误信息的用户更有可能关注一组相似的账户，并在一组相似的内容中分享信息。

分析社会认知引导影响的主要挑战在于，目前尚缺乏对“回音室”“信息茧房”“两极分化”“群体极化”等认知概念的公认量化方法和指标。因此，要在社交媒体平台上开展一致的认知影响度量非常困难，同时缺乏公认的评判标准，对度量的准确性无法评判，不同的研究得出不同结果，可能是由于使用的量化方式不同导致的。

6.4 平台对抗优化能力度量

社交媒体平台已经成为一种公共基础设施，它们是社交媒体优化的主要发生场所和防控中心，社交媒体平台对抗恶意优化的能力是一项具有社会价值的研究方向，这项度量的手段是综合性的，包括平台政策核查、技术手段检测和响应策略评估等。

优化对抗能力度量主要由公立的第三方机构开展，公开发表的研究并不多。2021 年，北约战略通信中心设计了一项开创性的实验[171]，他们直接在社交媒体平台上进行真实的恶意优化活动，这些优化活动购买于俄罗斯社交媒体优化服务商，实验定义了包括成功阻止创建虚假账户、检测和删除虚假账户的能力、检测和删除虚假活动的能力、优化活动的购买成本、交付速度、对虚假活动的举报反应以及行动的透明度共七项具体的评价指标。根据在 2021 年 9 月至 11 月期间对 Facebook、Instagram、YouTube、Twitter、TikTok 和 VKontakte 开展的实验结果显示，与一年前相比，社交媒体优化活动的实施速度更快，成本更低，且成功率极高，四周后 96％的优化活动仍未被社交媒体平台删除。各平台在对抗恶意优化的能力方面存在显著的差异，Twitter 在 2021 年仍然是行业领导者，TikTok 和 Facebook 紧随其后。

现阶段，优化对抗能力度量主要通过模拟真实优化活动来开展，这种评估方法的难点在于，无法保证模拟的优化活动具备与真实优化的一致性和完整性，同时购买优化服务和开展优化行为可能会带来法律和道德风险。

7 社交媒体恶意优化对抗手段

随着对社交媒体优化行为的揭露，有效对抗和打击社交媒体恶意优化行为成为研究者、政策制定者的重点关注议题，本章节从技术对抗、政策监管两个角度讨论针对社交媒体优化的治理手段。

7.1 技术措施

技术手段在对抗社交媒体优化方面发挥着不可替代的作用，研究者也致力于相关技术的研发和使用，代表性的对抗技术分为以下三类：

(1)人工智能技术识别优化行为。不引起注意是部分社交媒体优化能如预期一样发挥作用的前提，因此有效识别是阻止、控制和预防优化的关键。

研究者致力于开发社交媒体优化的检测方法,具体的技术方法在第 3.3 章恶意优化内容识别、第 4.3 章受控账号识别、第 5.3 章优化活动识别中已有详细分析。社交媒体平台通过使用这些方法对平台上的信息、账户和活动进行识别和剔除,以 Facebook 为例,自 2017 年以来,Facebook 对 200 多个违反“协调不真实行为”政策的社交媒体活动进行了执法,涉及 68 个国家①。Twitter 也在积极识别、删除和关闭恶意优化内容和账户②。2019 年,Twitter 收购了专注于机器学习和虚假信息检测的公司 FabulaAI,帮助其改善平台信息环境。

(2)事实核查辨别信息真伪。近年来,大语言模型的出现使生成信息与改写人造信息变得更容易,导致通过文字模式鉴别内容真假更加困难,事实核查通过对事件的真实性、描述与事实的匹配程度等查验开展,为解决虚假信息问题提供了新思路和更可靠的检测解释性。事实核查依赖第三方事实核查机构对信息的调查核验,比如法新社在 2017 年推出的数字验证服务 Factcheck③,路透社的事实核查网站 ReutersFactCheck④,普林斯顿大学运营的网站 PolitiFact⑤,以及由事实核查人员组织的国际事实核查网络论坛 IFCN⑥,这些机构关注有影响力的网络信息和恶意的网络活动,开展针对性的调查和反驳。

(3)评估和提升算法可信度。除了使用外部技术对抗恶意优化行为,也有研究者致力于研究高可信的推荐算法,通过研究如何让推荐系统识别和抵抗攻击,进一步提升平台推荐算法的可信度和安全性来预防优化行为,有的研究者在推进系统中使用差分隐私[174]、联邦学习[175]等技术手段保护用户隐私数据;有的研究者通过定位推荐系统偏差和去除偏差的角度[176],弱化推荐信息的“茧房极化”现象和提升算法公平性;有的研究者致力于算法的审计及评估方法研发[177],提出衡量算法公平性、隐私风险和价值观的方法,实现对智能算法表现出的偏见、歧视、隐私泄露、人类价值观偏离等算法滥用问题的多维度量化评估。

7.2　社会治理

社会共治通过调动政府、企业和公众的力量,形成协作机制,将政府的监管能力、企业的技术能力和公众的监督参与结合在一起,提高治理的综合效能。

(1)政府制定政策及实施监管。政府对社交媒体平台开展监管被视为解决社交媒体优化问题的一种可能路径。在欧洲,欧盟的《通用数据保护条例》和德国的《网络执行法》,已经颁布并施行。澳大利亚也出台了《散播邪恶暴力内容法》。2021 年 6 月,美国参议院商业、科学和交通委员会重新提出了《过滤气泡透明度法案》,旨在赋予用户对获取信息的更多选择和控制权。在中国,已经发布了《网络信息内容生态治理规定》《互联网信息服务算法推荐管理规定》《网络暴力信息治理规定》等法规,旨在规范互联网信息生产、传播、消费等各个环节。除了法律法规的完善,政府也采取了一系列监管行动打击社交媒体恶意优化。例如,美国联邦政府发起了媒体扫盲运动,旨在揭示社交媒体上的误导信息。欧洲对外行动署设立了“东方战略通信特别工作组”,作为欧盟的快速警报系统,协调和制定应对虚假宣传的措施。在中国,开展了互联网信息服务算法备案、新技术新应用评估、“清朗”系列专项行动等,规范互联网信息服务算法,着力研究破解网络生态新风险。

(2)社交媒体平台规范及信息透明。社交媒体优化的分散性使得高效的对抗手段依赖于社交媒体平台在信息分发过程中的核心地位,政府也通过设定激励或惩罚措施,鼓励平台在打击恶意优化活动方面发挥更积极的作用,具体包括调整社区规则和提高信息透明度,社交媒体平台通过持续调整规则,适应复杂的网络环境,Facebook 为例,自 2016 年后,公司多次对新闻信息规则实施重大调整,包括增加人工审查力度、调整新闻推送算法、简化假新闻申诉流程、标记有争议的内容等;社交媒体平台也试图通过提高信息透明来缓解平台优化行为,2018 年 4 月,Facebook 公开了社区标准的具体内容和执法报告,Google 也发布了第一份“社区准则执行报告”,分别揭示了平台上恶意优化的具体行为和执法情况。各社交媒体平台都在竞选广告方面增加了透明度措施。如 Facebook 和 Twitter 都要求在投放与竞选相关的广告必须显著标注并公开额外的信息,如广告商、广告活动的花费、面向的目标群体等。

(3)社会公众监督。公众对可疑内容和账号的“举报”是净化网络环境、对抗社交媒体恶意优化的有效方式,社交媒体平台为此设置了完整的受理流程和监督系统。例如,在新浪微博上,用户可以报告

① https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements/。
② https://blog.x.com/en_us/topics/company/2018/enablingfurther-research-of-information-operations-on-twitter。
③ https://factcheck.afp.com/。
④ https://www.reuters.com/fact-check。
⑤ https://www.politifact.com/。
⑥ https://www.poynter.org/ifcn/。

和标记可疑内容，并指出该内容违反了哪些社区规则，最终根据微博社区委员会的投票结果作出裁定。公众的监督机制有助于提高信息透明度并促使政府和企业履行治理责任。

8 挑战及未来展望

从 2006 年研究者首次表示担忧社交媒体被用于影响公众舆论，到 2016 年美国总统大选期间大量虚假新闻涌现引发人们对社交媒体优化的警觉，再到 2023 年生成式 AI 在创作任务中展现的卓越性能和创作的高逼真虚假信息在社交媒体广泛传播，当前社交媒体优化已是网络空间安全领域的热点问题。为了提供社交媒体优化的宏观视角，本文明确定义了社交媒体优化并提出社交媒体优化的三要素，分别是优化内容、受控账号和优化活动，围绕三要素，分别从分类、产生、识别角度对社交媒体优化问题进行了深入剖析；同时，为了及时跟进优化研究的最新进展，围绕优化行为曝光规模、现实事件影响程度等 4 个维度梳理了社交媒体优化的度量方法；另外，为了能够对解决社交媒体恶意优化发挥助力作用，从技术措施和社会治理两个角度介绍了社交媒体恶意优化对抗手段。除此之外，本文总结了社交媒体优化研究面临的挑战，并分析了未来发展趋势。

8.1 面临挑战

社交媒体优化这一研究领域的重要性得到了广泛认可，但仍然面临一些限制和挑战，主要包括以下两点：

(1)优化行为的持续“进化”。大量证据显示，为了规避检测技术，社交媒体优化一直在持续“进化”中。随着人工智能技术，特别是生成式人工智能技术的发展，优化实施者们获取这些技术的门槛越来越低，它们被广泛应用于社交媒体优化中，例如，根据 Facebook 发布的报告，2022 年被发现的所有”协调的不真实活动”网络中，超过三分之二的账户使用了生成对抗网络生成的个人资料图片，优化实施者将此视为一项模仿真实用户的手段。除了在生成用户资料上的应用，本文也分析了深度伪造技术在生成难以辨别真伪的虚假内容，以及构建社交媒体机器人网络方面的使用。这些技术通过增加生成内容的逼真程度和机器人账户的类人特性，提升内容鉴伪和机器人检测的难度，现有的自动化审核机制难以完全覆盖所有技术革新下的审核需求。同时，社交媒体优化实施者并非对优化检测方法一无所知，他们一直有针对性地持续更新优化方式，面对“进化”的优化手段，检测技术是否有效应对是一个持续性的挑战。

(2)获取社交媒体数据集的困难。一方面，社交媒体研究所需的关键数据往往由私营公司掌握，这导致研究者在获取重要的社交媒体数据集时面临困难。本文收集的大部分社交媒体优化研究主要 Twitter 数据的分析，然而 Twitter 并非最具影响力或优化行为发生最频繁的平台，这种研究上的不平衡主要是因为相较于其他平台，Twitter 的数据更易获取。对于全球最受欢迎的社交媒体平台 Facebook，关于其优化的分析多数是由 Facebook 公司的研究人员或与 Facebook 合作的学者开展，往往还需要公司的预发布审批。此外，社交媒体平台对优化行为的立场与研究者并不完全一致，例如，当平台发布优化数据时，他们可能会淡化其产品在社交媒体优化中的重要性。这都使得保证数据集的完整性和可靠性变得更为困难。另一方面，处理用户社交媒体数据时，法律与道德的考量至关重要，为确保所有的数据分析和共享都在尊重用户隐私的前提下进行，这让研究变得艰难。2014 年，剑桥大学的一位研究人员以个人身份在 Facebook 平台上发起了一项心理学问卷调查，参与此项调查的用户同意提供他们自身以及其朋友们在 Facebook 上的个人信息和行为数据用于学术研究。然而，该数据被用于剑桥分析公司辅助特朗普的选举活动。这一事件不仅是一场商业丑闻，也是一场学术丑闻，常在政策讨论中被引用。尽管这一事件是否违反了 Facebook 的服务条款仍在争议中，但其未经朋友同意就侵犯其隐私数据的行为产生了系列道德问题，该行为对分析社交媒体数据的学术研究产生了寒蝉效应。

8.2 未来研究展望

由于社交媒体优化是一个新兴的研究领域，本文还列出了一些有趣的潜在问题以供未来探索：

(1)如何识别和预防结合大模型技术的恶意优化行为？大语言模型与内容创作、社交机器人融合，极大提升了错误、虚假信息的生成与传播速度，强化了社交机器人的内容生成能力和对目标受众影响的精准度，大模型技术已经成为优化的重要工具，未来的潜力还将进一步释放，如何识别和预防结合大模型技术的优化行为成为新的挑战。可能的一个解决思路是，将针对大模型发展的生成内容检测、生成内容标识、价值观对齐和评估技术与现有优化识别、度

量、对抗技术融合,以应对技术革新下的挑战。

(2)如何早期识别社交媒体优化,并预测其可能产生的影响?当前的社交媒体优化检测和分析方法主要侧重于事后检测和分析,即对已经发生的社交媒体优化行为进行识别和度量,在优化行为影响力还不明显的初期,这些方法并不适用。在实际应用中,早期发现并监控一个正在进行的优化活动,对于预防和阻断恶意优化行为有着至关重要的作用,这也能够最大限度地降低恶意优化行为的负面影响。

(3)如何有效评估优化识别和度量的准确性?相对于社交媒体优化案例分析或检测方法来说,对优化的深入剖析和量化研究工作较欠缺,一个最主要的原因在对优化行为开展检测和度量影响分析时,缺乏公认的量化评估方法和标准,导致很多工作难以对比,也无法给出可信的依据来源,因此建立评估方法是提升优化研究的科学性、透明性的必经之路,也是增加决策者对采取相关技术的信任基础。

(4)如何解释并归因社交媒体优化行为?除了将某些特定的内容归类为宣传,或者将某个账户标记为虚假账户之外,理解和识别社交媒体优化行为的背后意图对于深入研究社交媒体优化具有重要的意义。目前这个问题基本上还没有得到解决,解决这个问题可能需要计算机、传媒学、心理学、社会学等多学科的联合努力。

参 考 文 献

[1] Howard P N, et al. New media campaigns and the managed citizen. Cambridge, UK: Cambridge University Press, 2006

[2] Ratkiewicz J, Conover M, Meiss M, et al. Detecting and tracking political abuse in social media//Proceedings of the International AAAI Conference on Web and Social Media: Volume 5. Barcelona, Spain, 2011: 297-304

[3] Metaxas P T, Mustafaraj E. Social media and the elections. Science, 2012, 338(6106): 472-473

[4] Howel L. Digital wildfires in a hyperconnected world. Global risks report//Proceedings of the World Economic Forum, Cologny. Geneva, Switzerland, 2013: 23-27

[5] Fourney A, Racz M Z, Ranade G, et al. Geographic and temporal trends in fake news consumption during the 2016 US presidential election//Proceedings of the CIKM: Volume 17. Singapore, 2017: 6-10

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

[7] Bovet A, Makse H A. Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications, 2019, 10(1): 1-14

[8] Eady G, Paskhalis T, Zilinsky J, et al. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 2023, 14(1): 62. DOI:10.1038/s41467-022-35576-9

[9] Barberá P, Chen A, Allcott H, et al. Like-minded sources on Facebook are prevalent but not polarizing. Nature, 2023, 620(7972): 137-144

[10] Guess A M, Malhotra N, Pan J, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? Science, 2023, 381(6656): 398-404

[11] Guess Am, Malhotran,Panj, et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. Science, 2023, 381(6656):404-408

[12] Gonzalez-Bailons, Lazerd, Barberá P, et al. Asymmetric ideological segregation in exposure to political news on Facebook. Science, 2023, 381(6656):392-398

[13] Lee K, Tamilarasan P, Caverlee J. Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media//Proceedings of the International AAAI Conference on Web and Social Media: Volume 7. Massachusetts, USA, 2013: 331-340

[14] Bradshaw S, Campbell-Smith U, Henle A, et al. Country case studies industrialized disinformation: 2020 global inventory of organized social media manipulation. Oxford, UK: Oxford Internet Institute, Technical Report, 2021

[15] Bridgman A, Merkley E, Loewen P J, et al. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. Harvard Kennedy School Misinformation Review, 2020, 1(3), DOI: 10.37016/mr-2020-02

[16] Wu l, Morstatter F, Carley Km, et al. Misinformation in social media: Definition, manipulation, and detection. ACM SIGKDD Explorations Newsletter, 2019, 21(2):80-90

[17] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 2017, 19(1):22-36

[18] Zubiaga A, Aker A, Bontcheva K, et al. Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR), 2018, 51(2):1-36

[19] Woolley S, Shorey S, Howard P. The bot proxy: Designing automated self-expression//A Networked Self and Platforms, Stories, Connections. New York, USA: Routledge, 2018: 59-76

[20] Assenmacherd,Clever L,Frischlichl, et al. Demystifying social bots: On the intelligence of automated social media actors. Social Media+Society, 2020, 6(3):2056305120939264

[21] Schafer F, Evert S, HEINRICH P. Japan's 2014 general election: Political bots, right-wing internet activism, and prime minister Shinzō Abe's hidden nationalist agenda. Big Data, 2017, 5(4):294-309

[22] Rogers R, Niederer S. The politics of social media manipulation. Amsterdam, Netherlands: Amsterdam University Press, 2020

[23] Marwick A E, Lewis R. Media manipulation and disinformation online. New York: Data & Society Research Institute, 2017, 359: 1146-1151

[24] Cheny, Conroy N J, Rubin V L. Misleading online content: Recognizing clickbait as "false news"//Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. Seattle, USA, 2015: 15-19

[25] Cresci S, Lillo F, Regoli D, et al. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. ACM Transactions on the Web (TWEB), 2019, 13 (2):1-27

[26] Keller F B, Schoch D, Stier S, et al. How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea//Proceedings of the 11th International AAAI Conference on Web and Social Media. Montreal, Canada, 2017, 11(1): 564-567

[27] Arif A, Stewart L G, Starbird K. Acting the part: Examining information operations within #BlackLivesMatter discourse. Proceedings of the ACM on Human-Computer Interaction, 2018, 2 (CSCW): 1-27

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

[29] Guess A, Nyhan B, Reifler J. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. European Research Council, 2018, 9(3):4

[30] Shao C, Hui P M, Wang L, et al. Anatomy of an online misinformation network. PLOS ONE, 2018, 13(4): e0196087

[31] Howard P N, Bolsover G, Kollanyi B, et al. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. Oxford, UK: Oxford Internet Institute, Technical Report: 2017.1, 2017

[32] Hindman M, Barashv. Disinformation, 'fake news' and influence campaigns on Twitter. Washington, USA, Technical Report,2018

[33] Shao C, Ciampaglia G L, Varol O, et al. The spread of low-credibility content by social bots. Nature Communications, 2018, 9(1):1-9

[34] Bessi A, Ferrara E. Social bots distort the 2016 US presidential election online discussion. First Monday, 2016, 21(11). DOI:10.5210/fm.v21i11.7090

[35] Abu-El-Rub N, Mueen A. Botcamp: Bot-driven interactions in social campaigns//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 2529-2535

[36] Faris R, Roberts H, Etling B, et al. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. Cambridge, MA 02138, USA: Berkman Klein Center Research Publication, Technical Report: 6, 2017

[37] Howard P N, Kollanyi B. Bots, #strongerin, and #brexit: Computational propaganda during the UK-EU referendum. arXiv preprint arXiv:1606.06356, 2016

[38] Bastos M T, Mercea D. The Brexit botnet and user-generated hyperpartisan news. Social Science Computer Review, 2019, 37(1): 38-54

[39] Bastos M, Mercea D. The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. Philosophical Transactions of the Royal Society A, 2018, 376(2128):20180003

[40] Del Vicario M, Gaito S, Quattrociocchi W, et al. News consumption during the Italian referendum: A cross-platform analysis on Facebook and Twitter//Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Tokyo, Japan, 2017: 648-657

[41] Del Vicario M, Gaito S, Quattrociocchi W, et al. Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the Italian referendum. arXiv preprint arXiv:1702.06016, 2017

[42] Guarino S, Trino N, Chessa A, et al. Beyond fact-checking: Network analysis tools for monitoring disinformation in social media//Proceedings of the International Conference on Complex Networks and Their Applications. Madrid, Spain, 2020: 436-447

[43] Ferrara E. Disinformation and social bot operations in the run up to the 2017 French presidential election. arXiv preprint arXiv:1707.00086, 2017

[44] Brachten F, Stieglitz S, Hofeditz L, et al. Strategies and influence of social bots in a 2017 German state election-a case study on Twitter. arXiv preprint arXiv:1710.07562, 2017

[45] Morstatter F, Shao Y, Galstyan A, et al. From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election//Proceedings of the Companion Web Conference. Lyon, France, 2018: 621-628

[46] Jones M O. The Gulf information war Propaganda, fake news, and fake trends: The weaponization of Twitter bots in the Gulf crisis. International Journal of Communication, 2019(13): 1389-1415

[47] Cardoso F, Luceri L, Giordano S. Digital weapons in social media manipulation campaigns//Proceedings of the 14th International AAAI Conference on Web and Social Media. Georgia, USA, 2020: Workshop

[48] Giglietto F, Iannelli L, Rossi L, et al. Mapping Italian news media political coverage in the lead-up to 2018 general election. SSRN Electronic Journal, 2018. DOI:10.2139/ssrn.3179930

[49] Resende G,Melo P,Sousa H, et al. (Mis)information dissemination in WhatsApp: Gathering, analyzing and countermeasures//Proceedings of World Wide Web Conference. San Francisco, USA, 2019: 818-828

[50] Pierri F, Artoni A, Ceri S. Investigating Italian disinformation spreading on Twitter in the context of 2019 European

elections. PLOS ONE, 2020, 15(1): e0227821

[51] Caldarelli G, De Nicola R, Del Vigna F, et al. The role of bot squads in the political propaganda on Twitter. Communications Physics, 2020, 3(1):81

[52] Narayanan V, Kollanyi B, Hajela R, et al. News and information over Facebook and WhatsApp during the Indian election campaign. Oxford, UK: Oxford Internet Institute, Data Memo: 2019.2, 2019

[53] Assenmacher D, Clever L, Pohl J S, et al. A two-phase framework for detecting manipulation campaigns in social media//Proceedings of the International Conference on Human-Computer Interaction. Copenhagen, Denmark: Springer, 2020: 201-214

[54] Ferrara E, Chang H, Chene, et al. Characterizing social media manipulation in the 2020 US presidential election. First Monday, 2020, 11(25), DOI:10.5210/fm.v25i11.11431

[55] Chang H C H, Chen E, Zhang M, et al. Social bots and social media manipulation in 2020: The year in review//Handbook of Computational Social Science, Volume 1. New York, USA: Routledge, 2021: 304-323

[56] Rogers R. Marginalizing the mainstream: How social media privilege political information. Frontiers in Big Data, 2021, 4:689036. DOI:10.3389/fdata.2021.689036

[57] Ferrara E. What types of COVID-19 conspiracies are populated by Twitter bots?. arXiv preprint arXiv:2004.09531, 2020

[58] Al-Rawi A, Shukla V. Bots as active news promoters: A digital analysis of COVID-19 tweets. Information, 2020, 11(10):461

[59] Ferrara E, Cresci S, Luceri L. Misinformation, manipulation, and abuse on social media in the era of COVID-19. Journal of Computational Social Science, 2020, 3(2):271-277

[60] Prabhu A, Guhathakurta D, Subramanian M, et al. Capitol (pat) riots: A comparative study of Twitter and Parler. arXiv preprint arXiv:2101.06914, 2021

[61] Shen F, Zhang E, Zhang H, et al. Examining the differences between human and bot social media accounts: A case study of the Russia-Ukraine war. First Monday, 2023. 8(2). DOI:10.5210/fm.v28i2.12777

[62] Geissler D, Bär D, Pröllochs N, et al. Russian propaganda on social media during the 2022 invasion of Ukraine. EPJ Data Science, 2023, 12(1):35

[63] Abdine H, Guo Y, Rennard V, et al. Political Communities on Twitter: Case Study of the 2022 French Presidential Election//Proceedings of the First Workshop on NLP for Political Sciences (politicalnlp2022). European Language Resources Association, 2022: 62-71

[64] Ballings M, Van Den Poel D, Bogaert M. Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. Omega, 2016, 59:15-25

[65] Young S W. Introduction to social media optimization: Setting the foundation for building community. Library Technology Reports, 2016, 52(8):5-8

[66] Aral S, Eckles D. Protecting elections from social media manipulation. Science, 2019, 365(6456):858-861

[67] Bradshaw S, Howard P. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. Oxford, UK: Oxford Internet Institute, Working Paper, 12, 2017

[68] Bradshaw S, Howard P N. Challenging truth and trust: A global inventory of organized social media manipulation. The Computational Propaganda Project, 2018(1):1-26

[69] Islam Mr, Liu S, Wang X, et al. Deep learning for misinformation detection on online social networks: A survey and new perspectives. Social Network Analysis and Mining, 2020(10):1-20

[70] Wardle C, Derakhshan H. Information disorder: Toward an interdisciplinary framework for research and policymaking. Strasbourg, France: Council of Europe Strasbourg, 2017

[71] Del Vicario M, Bessia, Zollof, et al. The spreading of misinformation online. Proceedings of the National Academy of Sciences, 2016, 113(3):554-559

[72] Just M R, Crigler A N, Metaxas P, et al. 'It's trending on Twitter'—an analysis of the Twitter manipulations in the Massachusetts 2010 special Senate election//Proceedings of the APSA 2012 Annual Meeting Paper. New Orleans, USA, 2012,(08):1-22

[73] Molina M D, Sundar S S, Le T, et al. "Fake news" is not simply false information: A concept explication and taxonomy of online content. American Behavioral Scientist, 2021, 65(2):180-212

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

[75] Zhao Z, Resnick P, Mei Q. Enquiring minds: Early detection of rumors in social media from enquiry posts//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 1395-1405

[76] Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying//Proceedings of The Social Mobile Web, Papers from the 2011 ICWSM Workshop. Barcelona, Spain,2011, 5(3): 11-17

[77] Warner W, Hirschberg J B. Detecting hate speech on the World Wide Web//Proceedings of the Second Workshop on Language in Social Media. Montréal, Canada, 2012: 19-26

[78] Yamaguchi S. Why are there so many extreme opinions online?: An empirical, comparative analysis of Japan, Korea and the USA. Online Information Review, 47(1): 1-19

[79] Westerlund M. The emergence of deepfake technology: A review. Technology Innovation Management Review, 2019, 9(11): 39-52

[80] Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 2020, 53(5): 1-40

[81] Dong Y, He D, Wang X, et al. A generalized deep Markov random fields framework for fake news detection//Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Macao, China, 2023: 4758-4765

[82] Guacho G B, Abdali S, Shah N, et al. Semi-supervised content-based detection of misinformation via tensor embeddings//Proceedings of the 18th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona, Spain, 2018: 322-325

[83] Hosseinimotlagh S, Papalexakis E E. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles//Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2). Los Angeles, USA, 2018:1-8

[84] Mitra T, Gilbert E. CredBank: A large-scale social media corpus with associated credibility annotations//Proceedings of the International AAAI Conference on Web and Social Media: Volume 9. Oxford, UK, 2015: 258-267

[85] Wang W Y. "Liar, liar pants on fire": A new benchmark dataset for fake news detection//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada, 2017: 422-426

[86] Shu K, Mahudeswaran D, Wang S, et al. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286, 2018

[87] Thorne J, Vlachos A, Christodoulopoulos C, et al. FEVER: A large-scale dataset for fact extraction and verification. Big Data, 2020, 8(3): 171-188

[88] Derczynski L, Bontcheva K. Pheme: Veracity in digital social networks. South Yorkshire, UK: University of Sheffield, UMAP Workshop: 2014

[89] Liu X, Nourbakhsh A, Li Q, et al. Real-time rumor debunking on Twitter//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1867-1870

[90] Hu X, Guoz, Wu G, et al. CHEF: A pilot Chinese dataset for evidence-based fact-checking. arXiv preprint arXiv:2206.11863, 2022

[91] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs//Proceedings of the 25th ACM International Conference on Multimedia. California, USA, 2017: 795-816

[92] Yu S, Li M, Liu F. Rumor identification with maximum entropy in micronet. Complexity, 2017(1): 1703870

[93] Zhou X, Jain A, Phoha V V, et al. Fake news early detection: A theory-driven model. Digital Threats: Research and Practice, 2020, 1(2): 1-25

[94] Magdy A, Wanas N. Web-based statistical fact checking of textual documents//Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents. Toronto, Canada, 2010: 103-110

[95] Ciampaglia G L, Shiralkar P, Rocha L M, et al. Computational fact checking from knowledge networks. PLOS ONE, 2015, 10(6): e0128193

[96] Shi B, Weninger T. Discriminative predicate path mining for fact checking in knowledge graphs. Knowledge-Based Systems, 2016, 104: 123-133

[97] Jin Z, Cao J, Zhang Y, et al. Novel visual and statistical image features for microblogs news verification. IEEE Transactions on Multimedia, 2016, 19(3): 598-608

[98] Rubin V L, Lukoianova T. Truth and deception at the rhetorical structure level. Journal of the Association for Information Science and Technology, 2015, 66(5): 905-917

[99] Jin Z, Cao J, Zhang Y, et al. News verification by exploiting conflicting social viewpoints in microblogs//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 30. Phoenix, USA, 2016, 30(1):2972-2978

[100] Kwon S, Cha M. Modeling bursty temporal pattern of rumors//Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. Ann Arbor, USA, 2014, 8(1): 650-651

[101] Wu K, Yang S, Zhu K Q. False rumors detection on Sina Weibo by propagation structures//Proceedings of the 2015 IEEE 31st International Conference on Data Engineering. Seoul, Republic of Korea, 2015: 651-662

[102] Alamsyah A, Sonia A. Information cascade mechanism and measurement of Indonesian fake news//Proceedings of the 2021 9th International Conference on Information and Communication Technology (ICoICT). Yogyakarta, Indonesia, 2021: 566-570

[103] Bian T, Xiao X, Xu T, et al. Rumor detection on social media with bi-directional graph convolutional networks//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 549-556

[104] Gupta M, Zhao P, Han J. Evaluating event credibility on Twitter//Proceedings of the 2012 SIAM International Conference on Data Mining. Anaheim, USA, 2012: 153-164

[105] Shu K, Wang S, Liu H. Beyond news contents: The role of social context for fake news detection//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. Melbourne, Australia, 2019: 312-320

[106] Chen L, Chen J, Xia C. Social network behavior and public opinion manipulation. Journal of Information Security and Applications, 2022, 64: 103060

[107] Chu Z, Gianvecchio S, Wang H, et al. Who is tweeting on Twitter: Human, bot, or cyborg? //Proceedings of the 26th Annual Computer Security Applications Conference. Austin, USA, 2010: 21-30

[108] Lee K, Webb S, Ge H. Characterizing and automatically detecting crowdturfing in Fiverr and Twitter. Social Network Analysis and Mining, 2015, 5: 1-16

[109] Liu Y, Liu Y, Zhang M, et al. Pay me and I'll follow you:

Detection of crowdturfing following activities in microblog environment//Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI): Volume 16. New York, USA, 2016: 3789-3796

[110] Ong J C, Cabañes J V A. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. Amherst, USA: University of Massachusetts, Technical Report: 2018

[111] Chavoshin, Hamooni H, Mueen A. On-demand bot detection and archival system//Proceedings of the 26th International Conference on World Wide Web Companion. Perth, Australia, 2017: 183-187

[112] Zannettou S, Caulfield T, De Cristofaro E, et al. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web//Proceedings of the Companion Proceedings of the 2019 World Wide Web Conference. San Francisco, USA, 2019: 218-226

[113] Iii R E G, Butts C T. Manufacturing diversity: Accomplishing mass passing and legitimation within social influence bot networks. Amherst, USA: University of Massachusetts, Technical Report, 2023

[114] Im J, Chandrasekharan E, Sargent J, et al. Still out there: Modeling and identifying Russian troll accounts on Twitter//Proceedings of the 12th ACM Conference on Web Science. Southampton, UK, 2020: 1-10

[115] Mazza M, Cresci S, Avvenuti M, et al. Rtbust: Exploiting temporal patterns for botnet detection on Twitter//Proceedings of the 10th ACM Conference on Web Science. Cambridge, UK, 2019: 183-192

[116] Kaghazgaran P, Caverlee J, Squicciarini A. Combating crowdsourced review manipulators: A neighborhood-based approach//Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. Marina del Rey, California, USA, 2018: 306-314

[117] Chen C, Wu K, Srinivasan V, et al. Battling the Internet water army: Detection of hidden paid posters//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls, Canada, 2013: 116-120

[118] Zhang Guangsheng, Kang Zhao, Tian Ling. User Identity Recognition Technology and Challenges for Network Security Governance. Journal of University of Electronic Science and Technology of China, 2023, 52(3): 398-412(in Chinese)
(张广胜, 康昭, 田玲. 面向网络安全治理的用户身份识别技术与挑战. 电子科技大学学报, 2023, 52(3):398-412)

[119] Cresci S. A decade of social bot detection. Communications of the ACM, 2020, 63(10): 72-83

[120] Cresci S, Di Pietro R, Petrocchi M, et al. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race//Proceedings of the 26th International Conference on World Wide Web Companion. Perth, Australia, 2017: 963-972

[121] Gilani Z, Farahbakhsh R, Tyson G, et al. Of bots and humans (on twitter)//Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. Sydney, Australia, 2017: 349-354

[122] Yang K C, Varol O, Hui Pm, et al. Scalable and generalizable social bot detection through data selection//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 1096-1103

[123] Yang K C, Varol O, Davis C A, et al. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies, 2019, 1(1): 48-61

[124] Feng S, Tan Z, Wan H, et al. Twibot-22: Towards graph-based Twitter bot detection. Advances in Neural Information Processing Systems, 2022, 35: 35254-35269

[125] Mukherjee A, Venkataraman V. Opinion spam detection: An unsupervised approach using generative models. Houston, USA: Texas A&M University, Technical Report, 2014

[126] Clark Em, Williams Jr, Jones Ca, et al. Sifting robotic from organic text: A natural language approach for detecting automation on Twitter. Journal of Computational Science, 2016, 16: 1-7

[127] Webb S, Caverlee J, Pu C. Social honeypots: Making friends with a spammer near you//Proceedings of the Conference on Email and Anti-Spam (CEAS). San Francisco, USA, 2008: 1-10

[128] Zarei K, Farahbakhsh R, Crespi N, et al. Impersonation on social media: A deep neural approach to identify ingenuine content//Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). The Hague, The Netherlands, 2020: 11-15

[129] Kolomeets M, Chechulin A, Kotenko I V. Bot detection by friends graph in social networks. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 2021, 12(2): 141-159

[130] Feng S, Wan H, Wang N, et al. Botrgcn: Twitter bot detection with relational graph convolutional networks//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Virtual, The Netherlands, 2021: 236-239

[131] Chen Kan, Chen Liang, Zhu Pei-Dong, et al. An Online Social Network 'Water Army' Detection Method Based on Interactive Behavior. Journal on Communications, 2015, 36(7): 120-128(in Chinese)
(陈侃, 陈亮, 朱培栋等. 基于交互行为的在线社会网络水军检测方法. 通信学报, 2015, 36(7):120-128)

[132] Rajendran G, Ram A, Vijayan V, et al. Deep temporal analysis of Twitter bots//Proceedings of the Machine Learning and Metaheuristics Algorithms, and Applications: First Symposium, SoMMA 2019. Trivandrum, India, 2020: 38-48

[133] Chavoshi N, Hamooni H, Mueen A. Debot: Twitter bot detection via warped correlation//Proceedings of the ICDM: Volume 18. Barcelona, Spain, 2016: 28-65

[134] Zhao J, Liu X, Yan Q, et al. Multi-attributed heterogeneous graph convolutional network for bot detection. Information Sciences, 2020, 537: 380-393

[135] Yang K C, Ferrara E, Menczer F. Botometer 101: Social bot practicum for computational social scientists. Journal of Computational Social Science, 2022, 5(2): 1511-1528

[136] Alothali E, Zaki N, Mohamed E A, et al. Detecting social bots on Twitter: A literature review//Proceedings of the 2018 International Conference on Innovations in Information Technology (IIT). Al Ain, United Arab Emirates, 2018: 175-180

[137] Wang G, Konolige T, Wilson C, et al. You are how you click: Clickstream analysis for Sybil detection//Proceedings of the USENIX Security Symposium: Volume 9. Washington, USA, 2013: 1-008

[138] Shi P, Zhang Z, Choo K K R. Detecting malicious social bots based on clickstream sequences. IEEE Access, 2019, 7: 28855-28862

[139] Varol O, Ferrara E, Menczer F, et al. Early detection of promoted campaigns on social media. EPJ Data Science, 2017, 6: 1-19

[140] Martino G D S, Cresci S, Barrón-Cedeño A, et al. A survey on computational propaganda detection//Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Japan, 2020: 4826-4832

[141] Mahbub S, Pardede E, Kayes A, et al. Controlling astroturfing on the Internet: A survey on detection techniques and research challenges. International Journal of Web and Grid Services, 2019, 15(2): 139-158

[142] Pennycook G, Cannon T D, Rand D G. Prior exposure increases perceived accuracy of fake news. Journal of Experimental Psychology: General, 2018, 147(12): 1865-1880

[143] Ou Ya, Xia Yue. Covert Persuasion: Computational Propaganda and Its Challenges to China's International Communication. Foreign Communication, 2019(12): 40-42(in Chinese)
(欧亚, 夏玥. 隐蔽的说服: 计算式宣传及其对中国国际传播的挑战. 对外传播, 2019(12): 40-42)

[144] Mathew B, Dutt R, Goyal P, et al. Spread of hate speech in online social media//Proceedings of the 10th ACM Conference on Web Science. Boston, USA, 2019: 173-182

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

[146] Del Vicario M, Vivaldo G, Bessi A, et al. Echo chambers: Emotional contagion and group polarization on Facebook. Scientific Reports, 2016, 6(1):37825

[147] Nikolov D, Oliveira D F M, Flammini A, et al. Measuring online social bubbles. Computer Science, 2015, 1:e38

[148] Del Vicario M, Zollof, Caldarellig, et al. The anatomy of Brexit debate on Facebook. arXiv preprint arXiv:1610.06809, 2016

[149] Recuero R, Araújo R. On the rise of artificial trending topics in Twitter//Proceedings of the 23rd ACM Conference on Hypertext and Social Media. Milwaukee, Wisconsin, USA, 2012: 305-306

[150] Elmas T, Overdorf R, Özkalay A F, et al. Ephemeral astroturfing attacks: The case of fake Twitter trends//Proceedings of the 2021 IEEE European Symposium on Security and Privacy (EuroS&P). Online, 2021: 403-422

[151] Yu L L, Asur S, Huberman B A. Artificial inflation: The real story of trends and trend-setters in Sina Weibo//Proceedings of the Privacy, Security, Risk & Trust. Tarragona, Spain, 2013: 514-519

[152] Grimmec, Assenmacherd, Adam L. Changing perspectives: Is it sufficient to detect social bots? //Proceedings of the 10th International Conference on Social Computing and Social Media (SCSM), Las Vegas, USA, 2018: 445-461

[153] Echeverrï£¡ A J, De Cristofaro E, Kourtellis N, et al. Lobo: Evaluation of generalization deficiencies in Twitter bot classifiers//Proceedings of the 34th Annual Computer Security Applications Conference. San Juan, Puerto Rico, 2018: 137-146

[154] Liu S, Hooi B, Faloutsos C. Holoscope: Topology-and-spike aware fraud detection//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Singapore, 2017: 1539-1548

[155] Pacheco D, Flammini A, Menczer F. Unveiling coordinated groups behind White Helmets disinformation//Proceedings of the Companion Proceedings of the Web Conference. Online, 2020: 611-616

[156] Yu R, Qiu H, Wen Z, et al. A survey on social media anomaly detection. ACM SIGKDD Explorations Newsletter, 2016, 18(1): 1-14

[157] Cai C, Li L, Zeng D. Detecting social bots by jointly modeling deep behavior and content information//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. Singapore, 2017: 1995-1998

[158] Peng J, Choo R K K, Ashman H. Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution//Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA. Tianjin, China, 2016: 121-128

[159] Keller F B, Schoch D, Stier S, et al. Political astroturfing on Twitter: How to coordinate a disinformation campaign. Political Communication, 2020, 37(2): 256-280

[160] Fletcher R, Cornia A, Graves L, et al. Measuring the reach of "fake news" and online disinformation in Europe. Australasian Policing, 2018, 10(2): 25-33

[161] Mosleh M, Rand D G. Measuring exposure to misinformation from political elites on Twitter. Nature Communications, 2022, 13(1): 7144

[162] Nimmo B. Measuring traffic manipulation on Twitter. Ox-

ford, UK: University of Oxford, Project on Computational Propaganda, Technical Report, 2019

[163] Bond R M, Fariss C J, Jones J J, et al. A 61-million-person experiment in social influence and political mobilization. Nature, 2012, 489(7415): 295-298

[164] Coppock A, Hill S J, Vavreck L. The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. Science Advances, 2020, 6(36): eabc4046

[165] Zimmer F, Scheibe K, Stock M, et al. Echo chambers and filter bubbles of fake news in social media. man-made or produced by algorithms. Proceedings of the 8th Annual Arts, Humanities, Social Sciences & Education Conference. Honolulu, USA, 2019: 1-22

[166] Flamino J, Galeazzi A, Feldman S, et al. Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. Nature Human Behaviour, 2023, 7(6): 904-916

[167] Cinelli M, De Francisci Morales G, Galeazzi A, et al. The echo chamber effect on social media. PNAS, 2021, 118(9): e2023301118

[168] Nguyen A, Vu H T. Testing popular news discourse on the "echo chamber" effect: Does political polarisation occur among those relying on social media as their primary politics news source?. First Monday, 2019, 24(5), DOI: 10.5210/fm.v24i5.9632

[169] Suhay E, Bello-Pardo E, Maurer B. The polarizing effects of online partisan criticism: Evidence from two experiments. The International Journal of Press/Politics, 2018, 23(1): 95-115

[170] Bessi A, Zollo F, Del Vicario M, et al. Users polarization on Facebook and YouTube. PLOS ONE, 2016, 11(8): e0159641

[171] Bay S, Fredheim R, Haiduchyk T, et al. Social media manipulation 2021/2022: Assessing the ability of social media companies to combat platform manipulation. NATO Strategic Communications Centre of Excellence. Riga, Latvia, 2022:1-46

[172] Jamieson K H. Cyberwar: How Russian hackers and trolls helped elect a president: What we don't, can't, and do know. UK: Oxford University Press, 2020

[173] Aral S, Nicolaides C. Exercise contagion in a global social network. Nature Communications, 2017, 8(1): 14753

[174] Martx E D, Abadi N, Chu A, et al. Deep learning with differential privacy//Proceedings of the 2016 ACM SIGSAC Conference. Vienna, Austria, 2016: 308-318

[175] Wang Y, Tian Y, Yin X, et al. A trusted recommendation scheme for privacy protection based on federated learning. CCF Transactions on Networking, 2020, 3(3-4): 218-228

[176] Ai Q, Bi K, Luo C, et al. Unbiased learning to rank with unbiased propensity estimation//Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval. Paris, France, 2018: 385-394

[177] Metaxa D, Park J S, Robertson R E, et al. Auditing algorithms: Understanding algorithmic systems from the outside in. Foundations and Trends® in Human-Computer Interaction, 2021, 14: 1-74

WANG Xiao-Shi, Ph. D. candidate. Her research interests include algorithmic security, trustworthy artificial intelligence, algorithmic regulation.

JING Shao-Ling, Ph. D., engineer. Her research interests include natural language processing, large language model evaluation.

SUN Fei, Ph. D., associate professor. His research interests include recommendation algorithm, natural language processing.

YIN Zhi-Yi, Ph. D., senior engineer. Her research interests include cybersecurity, social computing.

SHEN Hua-Wei, Ph. D., professor. His research interests include Web data mining, social network analysis, graph neural networks.

CHENG Xue-Qi, Ph. D., professor. His research interests include data science and big data analytics systems, network science and social computing, web search and minings.

Background

From the first expression of concern by researchers in 2006 about the potential use of social media to influence public opinion, to the emergence of a large number of fake news during the 2016 U. S. presidential election that raised alarms about social media optimization, and to the exceptional performance of generative AI in creative tasks and the widespread dissemination of highly realistic fake information on social media in 2023, social media optimization has become a hot topic in the field of cyberspace security. However, existing research lacks a systematic analysis and comprehensive understanding of social media optimization. This paper provides a systematic review of the theories and research related to social media optimization, aiming to offer strong guidance for analyzing optimization behaviors on social media and addressing the issue of malicious social media optimization.

By conducting in-depth research in the field of social media optimization, this paper begins by defining relevant terms and concepts of social media optimization based on ex-

isting discourse. Social media optimization refers to the organized and intentional dissemination or suppression of specific information on social media platforms to create favorable or unfavorable images or arguments that influence public opinion. The paper clearly defines social media optimization and proposes three elements of social media optimization: optimized content, controlled accounts, and optimization activities. It deeply analyzes the issue of social media optimization from the perspectives of classification, generation, and identification of these three elements. At the same time, in order to keep up with the latest developments in optimization research, the paper organizes measurement methods of social media optimization from four perspectives: the scale of exposure of optimization actions and the degree of impact on real-world events. Furthermore, to address malicious social media optimization, the paper introduces countermeasures against it from two perspectives: technical measures and social governance. In addition, the paper offers some thoughts on the challenges and development trends faced by research on social media optimization.