

虚实之间：迈向数字时代的谣言社会学分析

范新光 张勇军*

提要：数字时代改变了信息获取和传播的形式，这使得谣言频发已经成为普遍的社会现象。既有研究对于数字时代谣言的识别及社会学涵义缺少系统讨论。本文基于以往的研究思路，探讨数字时代谣言现象的特点，并提出一种基于深度学习的假信息识别策略。我们以疫情期间中文推特社区内关于中国和华人群体假信息为例，分析假信息的主题分布和传播过程，发现华人社区内部存在分化，参与针对中国或华人群体的谣言的转发与其对中国的态度有密切关系。本文希冀结合计算社会科学研究范式为数字时代的谣言社会学研究提供一定的启发。

关键词：谣言现象 假信息 数字时代 计算社会科学

一、引言

进入数字时代，谣言已经成为普遍的社会现象。在国际社会，2016 年美国总统大选期间，美国前总统唐纳德·特朗普批评推特、脸书等在线社交平

* 范新光，北京大学社会学系，北京大学武汉人工智能研究院；张勇军，纽约州立大学石溪分校社会学系和高等计算科学研究院。本文系中央高校基本科研业务经费项目“不平等与态度变迁研究”、东湖高新区国家智能社会治理实验综合基地项目“智能社会核心特征及治理模式研究”和中央引导地方科技发展专项“‘全息东湖’智能社会模拟器研发”（2023EGA035）的阶段性成果。

台和在线媒体纵容虚假和误导性信息^①，随后假新闻（Fake news）被柯林斯词典列为2017年的年度词汇之一^②。国内智能手机的普及和在线社交平台的崛起也催生大量的谣言，不仅影响民众日常生活和社会秩序，也成为威胁社会经济发展和国家安全的重要因素。中国互联网联合辟谣平台的数据显示，官方辟谣的谣言数量从2011年的26条骤升至2023年的3960条，所涉及的谣言类别也涵盖了历史、科学、社会热点、日常生活等各个方面。^③ 谣言现象的泛滥加速了世界各国网络安全相关法规的设立和完备。2016年我国通过《中华人民共和国网络安全法》，其中明确规定不得“编造、传播虚假信息扰乱经济秩序和社会秩序”，为保障网络信息依法有序自由流动提供了法律依据。

谣言伴随着人类社会的发展历史，但数字时代的谣言现象为何如此普遍？传统意义上，信息获取的途径无外乎社会关系、以报纸等为代表的媒体等。谣言现象不仅依赖虚假信息的生产，也取决于大众对虚假信息的传播。一方面，数字社会的来临深刻改变了信息生产和传播方式，尤其体现为获取渠道的多样化以及信息覆盖面前所未有的广度。在网络空间生产和传播虚假或不可靠信息的成本大大降低。信息真实性的甄别成本提高，这导致虚假信息难以在网络空间得到有效清理。以生成式人工智能为代表的新技术的出现也使得网络空间的信息真假莫辨^④。另一方面，大众在社交媒体平台的参与已经越来越普遍。截至2023年6月，我国网民规模达到10.79亿，微博月活跃用户数量也已突破6亿。^⑤ 在线社交媒体已经成为重要的信息获取渠

① Allcott, H. & Gentzkow, M., "Social Media and Fake News in the 2016 Election", *The Journal of Economic Perspectives*, 2017, 31(2): 211-235.

② Collins Dictionary, *Collins 2017 Word of the Year Shortlist*. Collins Dictionary Language Blog, <https://blog.collinsdictionary.com/language-lovers/collins-2017-word-of-the-year-shortlist/>.

③ 关于谣言的统计来自于“中国互联网联合辟谣平台”（<https://www.piyao.org.cn/>）。除此之外，中央网信办（国家互联网信息办公室）以及各社交媒体平台均有举报谣言等互联网违法和不良信息的渠道，中国互联网联合辟谣平台的统计主要针对产生较大社会舆论的谣言，是对当前互联网谣言现状的保守估计。

④ Sandrini, L. & Somogyi, R. "Generative AI and Deceptive News Consumption", *Economics Letters*, 2023, 232: 111317.

⑤ 中国互联网络信息中心：《第52次〈中国互联网络发展状况统计报告〉——互联网发展研究》，<https://www.cnnic.net.cn/n4/2023/0828/c88-10829.html>。微博用户数据来自微博公司发布2023年三季度财报：Weibo Corporate, *Weibo Reports Third Quarter 2023 Unaudited Financial Results*, 2023。

道。规模庞大的在线社交用户使得网络空间已经成为信息生产和传播的主要场域。

但是,上述回答并没有涉及数字时代谣言现象的动机、过程和后果。如果将关注暂时从“数字时代”这个设定移开,我们能够发现无论是社会学、哲学、经济学、政治学还是新闻传播学从产生机制、社会结构、传播过程等方面都有深入的讨论。尽管这些讨论往往不局限于谣言现象本身,但仍然对认识谣言的社会涵义有重要的价值。例如,社会心理学将谣言定义为对于个体而言有重要性和模糊性的信息,为认识谣言现象提供了基本的概念认识。^①不同学科将谣言视为文化生产和传播的现象,提供了各具侧重的分析思路。社会学结合了权力、结构和场域等概念,认为信息生产本质上并非客观。而经济学和政治学分别从各自学科视角对假信息的生产和传播及其后果展开了分析和探讨。但遗憾的是,尽管数字时代的谣言现象受到广泛关注,但这一现象在中国社会学研究中的关注度远不如数字经济、算法控制等。尽管少数研究关注具体的经验问题^②,但鲜少有研究关注从数字时代的特征出发,对谣言现象识别的社会学涵义进行系统性讨论。

随着在线社交的普及和全球化,对于谣言的研究已经迫在眉睫。新冠疫情以来,西方社会针对我国的谣言层出不穷,无论是对于我国的国际形象,抑或是对于社会稳定和国家安全都造成了极为负面的影响^③。尽管这些谣言随着时间不攻自破,但是对于谣言的深入认识仍然是当前值得关注的议题。

本文尝试对“后真相”时代这一概念的反思出发,提出数字时代的信息呈现出“虚实之间”的特点。在网络空间,常识意义上的真实和虚假信息往

① Allport, G. W. & Postman, L., “An Analysis of Rumor”, *The Public Opinion Quarterly*, 1946, 10(4): 501-517; Prasad, J., “The Psychology of Rumour: A Study Relating to the Great Indian Earthquake of 1934”, *British Journal of Psychology. General Section*, 1935, 26(1).

② 其中,周晓虹对“SARS”期间谣言传播的分析是目前国内对谣言现象较为系统的分析,参见周晓虹:《传播的畸变——对“SARS”传言的一种社会心理学分析》,《社会学研究》2003年第6期,第43—54页。李国武基于心理学以及自我实现预言的研究,运用博弈论对谣言实现机制提供了较为深入的思考,参见李国武:《谣言实现的社会机制及对信息的治理》,《社会》2005年第4期,第13页。

③ Fan, X. & Zhang, Y., “Just a virus or politicized virus? Global Media Reporting of China on COVID-19”, *Chinese Sociological Review*, 2023, 55(1): 38-65.

往难以广泛传播。迅速传播的虚假信息在内容上呈现出超出一般的大众常识,具有新奇(novelty)特点,并且难以被甄别真实性。在此基础上,我们将散布在不同学科的谣言分析思路加以整合,进而探索性提出契合于数字时代的谣言生产和传播的分析框架。假信息的生产和传播往往针对特定的个人、组织和群体。由于数字时代信息的“虚实之间”特点,假信息往往是由难以被个体判断真实性或来源的信息构成。相信假信息与否不再仅仅依赖于信息内容客观意义上的真实性,而取决于个人对于假信息所针对目标的主观态度。进而,个人是否参与到假信息的生产和传播过程也与对谣言目标的态度紧密相关。

本文指出对谣言现象的社会学分析需要整合以往研究思路,并提出人机协同的深度学习假信息识别框架。新冠疫情期海外华人群体内部的谣言生产和传播提供了有意义的经验案例。在去全球化和西方社会挑起中国威胁论的大背景下,新冠疫情催生了大量针对中国和海外华人群体的谣言。对于海外华人群体,其内部在中国和华人身份认同上存在差异,因此在回应针对中国或华人群体的谣言时也可能出现分化。本文以分析疫情期间推特平台中文推特的谣言信息及传播为例,为理解数字时代的谣言现象提供初步的尝试。

二、作为社会现象的谣言:相关研究评述

在讨论针对谣言现象的既有研究思路之前,我们有必要对谣言的概念进行回顾。尽管人类社会自出现以来便伴随着谣言,但直至20世纪学者才开始对谣言进行严谨的社会科学研究。普拉萨德对1934年印度大地震所产生的谣言现象时,从心理学角度总结了谣言产生依赖于情绪混乱、未知且难以验证的信息、群体兴趣以及恐惧等条件。奥尔波特和波斯特曼从社会心理学角度进一步给出谣言较早的定义。他们认为,谣言的生产和传播由信息的两个条件决定:重要性(importance,记为I)和模糊性(ambiguity,记为A)。这一界定强调信息的重要性和模糊性是形成对事件的个体想象和记忆的基

础,谣言的出现是个体对此类信息主动(active)回应——传播和扩散——的结果。皮特森和吉斯特由此出发,认为谣言是大众参与其中的社会现象,反映出了大众的期待、恐惧、焦虑、敌视以及愿景,谣言本质上是在社会语境之下个体对于现实的诠释、讨论、猜测以及创造性的想象。^①这一界定强调了社会情境在谣言生产和传播过程中的重要性。此后不同学科的学者从社会网络、文化生产等角度进一步诠释谣言的涵义,但本质上仍然是上述两种界定的进一步明确化。在此,我们可以将谣言视为重要且模糊的信息通过个体或社会群体再造记忆并传播的社会过程。

以往研究分别从文化生产、社会心理和信息传播角度对谣言现象的原因和过程进行了分析。政治和文化社会学将谣言的生产和传播放置于制度环境和社会语境下考察,认为谣言的生产和传播是涉及多主体的复杂社会过程。对媒体的政治经济学和社会学研究已经形成了对新闻生产的成熟认识。新闻和电视内容生产是不同场域或利益主体之间互动的结果,在本质上是对事实具有目的性的框架化(framing)^②,受到不同场域或利益主体(如政府、党派、族群、媒体等)的影响。新闻能否见诸于媒体(新闻覆盖)以及内容偏见往往不取决于新闻的直接生产者(如记者),而是由媒体主编、政府或党派决定。^③和新闻和其他媒体产品类似,谣言的生产和传播也并非个人的创造,而是具有不同利益诉求的主体之间博弈的结果。例如,以推特和脸书为代表的社交巨头被广泛批评的原因在于其对平台上虚假信息的放纵,其重要原因在于谣言往往能够拉动社交平台活跃用户规模。^④此外,通过有选择性地放

① Peterson, W. A. & Gist, N. P., "Rumor and Public Opinion", *American Journal of Sociology*, 1951, 57(2): 159-167.

② Entman, R. M., "Framing: Toward Clarification of a Fractured Paradigm", *Journal of Communication*, 1993, 43(4), 51-58; Chong, D. & Druckman, J. N., "Framing Public Opinion in Competitive Democracies", *The American Political Science Review*, 2007, 101(4): 637-655.

③ Bourdieu, P., *On Television and Journalism*, London: Pluto, 1998; Benson, R., "News Media as a 'Journalistic Field': What Bourdieu Adds to New Institutionalism, and Vice Versa. Political Communication", 2006, 23(2): 187-202.

④ Frenkel, S., Alba, D. & Zhong, R., "Surge of Virus Misinformation Stumps Facebook and Twitter—The New York Times", *New York Times*, 2020, <https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html>.

宽对虚假信息的监管,西方社会的社交平台也被批评为影响民意和操纵选举以实现其政治利益诉求的工具。^①在社交平台的场域,个体通过生产更新奇(novel)的谣言以获取社交媒体关注度。^②在国家场域亦是如此,出于意识形态、国家安全等国家利益的考虑,针对特定国家的谣言在数字时代普遍存在,在缺乏全球层面的制度监管下,谣言已经成为了国家谋取或实现自身利益的工具。^③

社会心理学将关注点则更为关注谣言现象中的个体或群体心理,在微观层面对政治和文化社会学作出重要的补充。信息在传播过程可能出现曲解,因为个体会有意或无意对信息进行创造性想象,是从自身利益或身份出发的再诠释过程。艾因威勒和卡明斯指出对谣言的看法和传播参与行为与个人的身份有显著关联。对于针对特定群体的谣言,处于这一群体且更有群体认同的成员往往会显示对负面谣言的否定;和谣言针对群体存在利益冲突的其他群体则更为相信甚至扩散谣言。^④可见,个体对谣言的看法往往脱离了谣言本身的真实与否,而是基于个体的自身位置和认同对信息的主观判断,以和自身所持有的观念系统是否保持一致,有可能存在偏见。换言之,这是一种防御性动机(defensive motivation)。劳森等的研究发现数字时代的网络社区已经分化为众多在线群体,从属于某一群体的成员如果不顺从其他成员转发虚假信息的行为,和群体内其他成员的互动频率会逐渐降低。他们将这种现象称为不分享虚假信息的社会成本(social cost)。^⑤从社会

① 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.

② Vosoughi, S., Roy, D. & Aral, S., "The Spread of True and False News Online", *Science*, 2018, 359(6380): 1146-1151.

③ Mazepus, H., Osmudsen, M. & Bang-Petersen, M., et al., "Information Battleground: Conflict Perceptions Motivate the Belief in and Sharing of Misinformation about the Adversary", *PLOS ONE*, 2023, 18(3).

④ Einwiller, S. A. & Kamins, M. A., "Rumor Has It: The Moderating Effect of Identification on Rumor Impact and the Effectiveness of Rumor Refutation", *Journal of Applied Social Psychology*, 2008, 38(9): 2248-2272.

⑤ Lawson, M. A., Anand, S. & Kakkar, H., Tribalism and Tribulations: "The Social Costs of Not Sharing Fake News", *Journal of Experimental Psychology: General*, 2023, 152(3): 611-631.

心理学出发的谣言研究往往将谣言的生产视为在个体或群体层面的诠释过程,谣言的传播是一种由于信息的重要性和模糊性并存以及集体顺从导致的结果。

计算社会科学研究范式下的网络科学研究为谣言现象的生产和传播提供了形式社会学的回应,从信息级联等角度模拟谣言传播的社会机制以讨论谣言何以可能。和前面两种思路不同,社会网络分析认为谣言的出现和消失本质上是扩散过程,在形式意义上和集体行动以及病毒的扩散过程存在相似性。^① 格兰诺维特提出的集体行动模型也可以解释谣言的传播过程。^② 如果将虚假或不可靠信息的生产者类比于集体行动的发起者,那么信息传播行为近似于参与集体行动的决策。由于个体的观念和态度会受到其社会网络的影响,如果个体社会网络内大部分个体相信或者传播谣言,那么个体相信或传播谣言的可能性也会显著提高。^③ 近些年来,大量研究发现谣言的内容往往更加具有新颖性,这导致谣言的传播速度和广度要超过真实新闻。^④ 尽管网络科学的研究思路对谣言产生的动机缺乏足够的关注,由此出发的计算机科学和计算社科学者进一步将研究的方向转向对谣言的识别和预测^⑤,为谣言监管和相关政策制定提供了重要的支撑。

上述研究思路各有侧重但互为补充。政治和文化社会学关注虚假信息传播的动机和社会、制度环境,社会心理学思路对应于个体对模糊信息的诠释过程,社会网络研究关注虚假信息内容的特点以及在社会网络中传播的过

-
- ① Friggeri, A., Adamic, L., & Eckles, D., et al., "Rumor Cascades", *Proceedings of the International AAAI Conference on Web and Social Media*, 2014, 8(1); Trpevski, D., Tang, W. K. S. & Kocarev, L., "Model for Rumor Spreading over Networks", *Physical Review E*, 2010, 81(5).
 - ② Granovetter, M., "Threshold Models of Collective Behavior", *American Journal of Sociology*, 1978, 83(6): 1420-1443.
 - ③ Ceron, W., Sanseverino, G. G., de-Lima-Santos, M.-F. & Quiles, M., "COVID-19 Fake News Diffusion across Latin America", *Social Network Analysis and Mining*, 2021, 11(1): 47.
 - ④ Ecker, U. K. H., Lewandowsky, S. & Cook, J., et al., "The Psychological Drivers of Misinformation Belief and its Resistance to Correction", *Nature Reviews Psychology*, 2020, 1(1).
 - ⑤ 谣言和虚假信息识别已经有大量文献。从社会网络理论出发,不少学者基于虚假和真实信息在网络空间传播特点的差异进行反向识别,例如 Zhou, X. & Zafarani, R., *Network-based Fake News Detection: A Pattern-driven Approach*, *ACM SIGKDD Explorations Newsletter*, 2019, 21(2): 48-60.

程。但是基于上述思路的既有研究存在有两点局限。首先,这些研究思路共同构成了谣言现象相对完整的图景,但不同思路之间的侧重不同。社会心理学研究更强调个体对虚假或不可靠信息的诠释,但是忽略了社会制度、关系和互动。而社会结构和互动恰恰是谣言生产和传播得以出现的社会基础,因为谣言生产和传播的动机来自个体或群体在其结构位置上的利益诉求和身份认同。其次,上述研究思路对于真相(truth)的讨论一直处于暧昧的态度。社会心理学研究认为对于信息的感知偏差和主观诠释促使了谣言的产生,因此谣言性在传播过程中通过有偏见的诠释会越来越偏离真相,但谣言的模糊性往往在于个体难以判断谣言真实性,但作为具有预期色彩的谣言在一定条件下也会转化为事实。数字时代的信息,其真实和虚假之间的边界越来越模糊,网络空间信息的广度和传播深度更增加了个体对信息真实性感知的判断。

三、虚实之间和“后真相”时代之辩

“后真相”这一术语直至2015年之后才为人所熟知。随着2016年美国大选将“假新闻”推向了大众舆论的中心。唐纳德·特朗普以“真相社交”(Truth Social)为名创建新的社交平台,以示对脸书和推特“假新闻”泛滥的不满。早在2004年,拉尔夫·凯耶斯以《后真相时代》为题出版了最早关于这一问题的讨论。尽管数字时代此时仍处于萌芽时期,但书中开篇引用丹尼尔·波斯汀的一句话精妙反映数字时代信息社会的根本特点:“真相”已经被“相信”所替代(“Truth” has been displaced by “believability”)。^①这引发了对于进入数字时代之后“真相是什么?”(what is truth?)的反思。本文认为“后真相”时代这一概念并没有指出数字时代信息的本质特点,反而

^① Keyes, R., *The Post-Truth Era: Dishonesty and Deception in Contemporary Life*, New York: Macmillan, 2004, p. 3.

以对真相的暧昧态度弱化了对谣言现象社会涵义的系统分析。数字时代的信息究竟呈现出何种特点是接下来讨论的问题。

网络空间和现实空间之间存在连接在于个人身份 (identification)。在数字时代,面对面的社会互动逐渐被网络空间的虚拟个体之间的线上社交所替代,这导致了数字时代的自我 (self) 两重属性:首先是活生生的、处于社会关系之中的现实身份;其次是以信息(网名、性别、年龄等)界定的数字身份。^①两重属性之间并不完全存在对应的关系,因为网络空间的身份信息可能有别于现实世界。其中的连接反应出数字时代网络空间的谣言区别于以往时代的重要特点:动机来源于现实中个体的社会经历,传播、纠正主要通过网络的数字身份进行。相较于以面对面或真实的身份参与谣言现象之中,匿名化的网络社区往往为个体提供了参与谣言生产和传播过程的额外动机,因为虚拟空间的匿名性降低了参与谣言而被惩罚的风险。当然,如果谣言针对的是现实世界的特定群体,那么处于这一群体且具有认同的个体在网络空间也会更积极参与到谣言的更正过程。^②无论是谣言的生产、传播或是更正,数字时代自我的两重属性都会导致谣言现象的可能性远远高于传统时代。

数字时代谣言现象的主体也因此发生了改变。一方面,谣言的生产和传播者在网络空间往往被符号化为群体形象,但谣言的对象是现实世界中的个体或群体。在网络空间,虚拟身份实际上成了生产和传播谣言的一重保护外壳。同时,更为极端的观点和意见也更可能在网络空间具有可见度。受此影响,个体的观点在复杂的网络环境下也可能更为极端。观点动力学研究从复

① Belk, R. W., "Extended Self in a Digital World", *Journal of Consumer Research*, 2013, 40(3): 477-00; Yee, N., Bailenson, J. N. & Ducheneaut, N., "The Proteus Effect: Implications of Transformed Digital Self-Representation on Online and Offline Behavior", *Communication Research*, 2009, 36(2): 285-312; Zhao, S., "The Digital Self: Through the Looking Glass of Telecopresent Others", *Symbolic Interaction*, 2005, 28(3): 387-405.

② Song, Y., Kwon, K. H., Lu, Y., et al., "The 'Parallel Pandemic' in the Context of China: The Spread of Rumors and Rumor-Corrections During COVID-19 in Chinese Social Media. *American Behavioral Scientist*, 2021, 65(14): 2014-2036.

杂网络视角已经为此提供了证据支持。^①这意味着网络空间基于观念或身份的群体分化程度可能远远大于传统时代。另一方面,在线社交平台已经成为了数字时代谣言现象重要的行动主体,因为其自身也具有维持活跃用户和提高收益等利益诉求,在对社交平台上谣言的监管和处理上往往难以保持客观和公正。尤其当在线社交平台自身具有特定的价值观念或代表特定的利益群体,往往会将谣言的处理作为表达立场乃至带动社会舆论的工具。因此,在线社交平台和网络空间在缺少监管的情况下,真假莫辨的信息大量存在。

在数字时代,信息的来源更为复杂,广度也远超以往。个体所接触的大部分信息,往往是经过在线媒体或社交平台诠释的,个体自身对信息真实性乃至源头的判断不仅成本极高,而且在多数时候也极为困难,这导致个体往往从自身已有的知识和观念出发去选择相信或否定所接触的信息。此外,不少学者也从谣言生产和传播的动机角度提出了新颖假设(novelty hypothesis),认为广泛传播的谣言往往虚实难辨又具有内容新颖的特点,个体会通过对这类信息的创造和传播以博取他者的关注。虚假或不可靠信息的生产试图在真实和虚假之间寻求平衡点,而非那些通过常识即可判断的信息。

由于甄别信息真伪困难,个体对信息是否选择相信取决于两个机制。首先,个体往往会依赖可信的个体、机构或信息来源以评估信息的真实性。然而,随着政治、专家和科学信任在全球范围内处于下降趋势^②,这导致辟谣的可信度也日趋下降。玛戈林等的研究发现,真相确认(fact checking correction)对遏制谣言扩散的影响极为有限。^③谁来承担辟谣的责任极为

① 观点动力学对观念极化的讨论多见于复杂系统和计算社会科学领域,已经取得了重要的研究推进。其中社会网络和互动对观念极化的影响是重要的研究议题,可参见: Baldassarri, D. & Bearman, P., "Dynamics of Political Polarization", *American Sociological Review*, 2007, 72(5): 784-811; Mei, W., Bullo, F. & Chen, G., et al., "Micro-foundation of Opinion Dynamics: Rich Consequences of the Weighted-median Mechanism", *Physical Review Research*, 2022, 4(2).

② Wang, Z. & You, Y., "The Arrival of Critical Citizens: Decline of Political Trust and Shifting Public Priorities in China", *International Review of Sociology*, 2016, 26: 1-20.

③ Margolin, D. B., Hannak, A. & Weber, I., "Political Fact-Checking on Twitter: When Do Corrections Have an Effect?", *Political Communication*, 2018, 35(2): 196-219.

重要。对于个体而言,客观意义的真相在数字时代难以寻觅;取而代之,对于辟谣的个体或组织的信任构成了个体选择继续相信或否定谣言信息的重要依据。^①

其次,如果谣言现象频发且谣言内容超出了个体的知识或认知范围,那么个体观念系统和社会网络决定了对谣言的态度。个体对于特定事物的态度不仅是其观念系统的结果,也会受到他者态度的影响。^②在这一过程中,个体的观念、价值和认同会决定其对谣言接纳、传播以及纠正与否。这意味着,在线社交平台更大意义上是作为大众舆论场,因为无论是媒体还是个人所传递的信息都是经过筛选(filter)、框架化(frame)的结果。在这种意义上,个体或群体之所以参与到谣言现象之中可能不仅仅是为明确事实。承接社会心理学和社会网络研究的思路,我们可以从身份和观念系统来看到谣言对于社会分化的影响。艾克尔等在讨论个体对虚假信息的态度背后的心理动机时指出,如果冲击了个体的世界观和社会文化认同等根本性观念,那么辟谣不再有效,反而会被认为是之于个体乃至社会群体的潜在威胁。^③

综上所述,数字时代更可能陷于谣言的狂欢和真相的沉默。这并非意味着真相不存在,而是由于现实和网络的个体二重性、信息甄别的复杂性以及个体对网络空间信息的主观诠释导致追求真相不再是个体参与谣言生产和传播的主要目的,谣言的产生和传播是个体身份和观念影响下的主动应对(包括传播或辟谣)过程。

这由此带来了对于谣言现象进行社会学分析的两个挑战。首先,谣言现象本质上是社会现象,对谣言的研究需要分析不同主体在谣言过程之中的决策

① Lewandowsky, S., Ecker, U. K. H. & Cook, J., "Beyond Misinformation: Understanding and Coping with the "Post-truth" Era", *Journal of Applied Research in Memory and Cognition*, 2017, 6(4): 353-369.

② Friedkin, N. E. & Johnsen, E. C., "Social Influence and Opinions", *The Journal of Mathematical Sociology*, 1990, 15(3-4): 193-206.

③ Ecker, U. K. H., Lewandowsky, S. & Cook, J., et al., "The Psychological Drivers of Misinformation Belief and its Resistance to Correction", *Nature Reviews Psychology*, 2020, 1(1): 13-29.

和回应,但现有的研究框架往往从单一视角理解这一现象,而缺乏系统性的讨论。其次,基于机器学习方法的假信息识别策略往往依赖专家或“可靠”信息来源创建训练集,但社会学研究已经表明信息的生产本身并非客观中立,那么如何在海量信息中识别假信息是需要应对的方法挑战。接下来,我们以新冠疫情期间海外华人群体为例进行进一步的分析和讨论,并从谣言现象的多重视角出发,提出一种基于深度学习的假信息识别策略,进而基于这一数据分析疫情期间海外华人社区内关于中国和华人群体谣言的主题和传播特点。

四、以新冠疫情期间海外华人群体中的 谣言传播现象为例

(一)背景

我们以新冠疫情期间海外华人群体内部的谣言生产和传播为例进行进一步的分析和讨论。新冠疫情时期,西方社会针对中国的舆论和谣言攻击频繁。在之前的研究工作中,我们关注海外媒体对中国的报道以及海外中文社群对中国的态度。^①我们从全球在线舆情数据库 GDELT 项目的多语言 COVID-19 在线新闻叙事数据库中抽取了超过 1000 万篇涉及中国的媒体文章,系统性分析了全球不同国家在疫情期间对中国的报道,发现海外在线媒体将新冠疫情政治化(politicization)是普遍的现象。需要注意的是,我们发现使用歧视性语言跟意识形态并没有直接关联,而往往跟媒体所在国的疫情严重程度以及应对政策有关。换句话说,针对中国的抹黑在一定程度上是为了转嫁国内的疫情压力。此外,我们利用疫情以来的中文推特社区发布的

^① Fan, X. & Zhang, Y., “‘Just a Virus’ or Politicized Virus? Global Media Reporting of China on COVID-19”, *Chinese Sociological Review*, 2023, 55(1): 38-65.

推文文本系统分析了中文社群在疫情期间对中国的态度变化。^①我们发现绝大部分涉及中国的推文包含着一定的负面情绪,这些情绪往往跟疫情、政治以及宗教议题相关。上述两项研究为本文的实证分析奠定了基础,由于新闻生产本身具有目的性,疫情期间针对中国的抹黑或虚假报道在网络泛滥,成为社交媒体上针对中国或华人群体的虚假信息来源。在线社交平台的华人社区内部也存在对中国的态度分化,那么考虑到网络存在大量针对中国和华人群体的虚假信息,对中国的态度和是否参与谣言的生产和传播之间的关系成为接下来自然的研究议题,也直接回应了前文关于谣言生产和传播过程的探讨。

新冠疫情期间海外华人群体的谣言生产和传播过程尤其契合本文对后真相时代谣言现象的讨论。全球大变局的背景下,海外华人群体的规模以及内部分化都呈现出一些特点。自改革开放以来,海外华人群体规模迅速扩大。以美国为例,1980年在美华人仅有37万,这一数字在2021年已经上升至238万。^②随着规模的扩大,海外华人群体在寻求当地社会融入和身份认同上呈现分化。考虑到在大部分国家,华人仍属于少数群体,身处西方社会的海外华人也往往受国家之间角力的影响,承受着指向中国的谣言。一部分海外华人在价值观念上呈现出趋近西方社会的特点,往往对中国有负面的态度,也会加入到指向中国或华人群体的谣言生产和传播之中。也有一部分华人对中国仍然有积极的态度和文化情节,但由于西方价值观念的主导,在面对谣言时往往难以发声。如前文所述,在后真相时代,在线社交媒体上谣言的生产和传播往往不再以真相为目的,而和个体自身的态度和价值观念紧密相关。可以认为,海外华人对中国的态度与其是否参与谣言的生产和传播存在相关,对中国持有负面态度的海外华人会更加积极参与谣言生产和传播过程。

① Zhang, Y., Lin, H., Wang, Y. & Fan, X., "Sinophobia was Popular in Chinese Language Communities on Twitter during the Early COVID-19 Pandemic", *Humanities and Social Sciences Communications*, 2023, 10(1).

② Terrazas, A. & Batalova, J., "Chinese Immigrants in the United States", 2010, <https://www.migrationpolicy.org/article/chinese-immigrants-united-states-2008>.

这一经验案例也体现了在实证层面对上述判断进行验证面临方法和数据的挑战。尽管识别假信息的方法在最近几年取得了重要进展,但是如何识别海外华人群体内部针对中国或华人群体的谣言仍然存在方法的挑战。如何判断新冠病毒流行期间针对中国的假信息?以往假信息识别方法往往基于CNN、BBC等所谓的客观媒体或专家评估进行新闻真实性的判断,但我们之前的研究已经指出针对中国的新闻报道本身具有偏向性和选择性,那么基于这些方法判断海外华人群体内部针对中国或华人群体的假信息也会存在误导。此外,谣言现象的研究不只包括信息的文本分析,也需要描述信息的传播网络,那么在识别假信息的基础上也需要从社会网络视角分析信息的传播速度和模式。

接下来,我们结合深度学习的方法,基于推特平台的中文推文大数据,提出一种人机协同的迁移学习框架,通过微调大语言模型识别疫情期间中文推特中的假信息,进一步结合回归分析和社会网络分析认识谣言的生产和传播过程。需要注意的是,本文是一个探索性研究,旨在提出利用机器学习和大数据来探讨疫情早期推特中文社群谣言生产和扩散的可能性方向。

(二)基于深度学习的假信息识别策略

基于推特开放的学术API,我们使用一系列跟中国、共产党、政府、华裔、亚裔等相关词汇从推特的历史全文数据库中抓取了2019年12月至2021年4月之间全部中文社群中和中国以及新冠疫情相关的讨论,采集了约132万推特用户的2500万中文推文以及中文推文之间的转发信息,构建了CNTweets的数据库。

为了从海量的推文中准确识别假信息,我们提出一个供社会科学研究参考使用的人机协同深度学习(Human-in-the-loop deep learning)谣言识别框架。我们在2021年底开始创建训练集。我们首先识别那些最有可能发布假信息和传播谣言的推特账户,采集这些账户过去两年发布的推文。另外我们使用支持或反对中国和华人群体的关键词和主题标签(hashtag)来构建了

中文推文数据库。最后，我们用分层抽样的方法从这些获得的数据中抽取了 12,000 条推文以及 3000 条来自 CNTweets 的随机推文作为我们初始的待标注训练集——CNFakeNews。我们聘请了多位研究生助理来标注推文，每条推文至少被不同的学生标注两次；如果两位学生的标注存在差异，由第三位学生独立标注并将之作为最后的结果。

我们使用这些人工标注的训练集来微调预训练的中文大语言模型 (RoBERTa)^①，然后使用最优模型来预测 CNTweets 中的 2500 万推文用以提取潜在的假信息。为了进一步提高模型的预测准确度，我们从第一轮预测的潜在假信息中随机抽取 5000 条推文进一步进行人工审核，然后更新我们的训练集用以继续微调我们的模型来提高准确度。最终，在数据清理后，我们的训练集包含 18,435 条推文，其中有 2599 条假信息。^②

以往的研究中通常侧重于使用可靠信息源（如第三方打假网站和专家）来对假信息进行客观识别 (fact-checker)，但是本文侧重于受众对假信息的主观感知。前文已经在理论和实证层面论述在中文推文社区依靠西方社会的“可靠”信息源以及专家进行信息真实性判断的问题。同样，考虑到假信息的虚实模糊性，因此通过海外华人社群之外、具有对设涉及中国或华人群体的中文信息内容具有甄别能力的人员进行主观感知判断成为目前最切实的选择。因此，我们选择在中国某知名大学的学生参与训练集的创建。在人工标注推文的过程中，我们使用了如下的主观定义：假信息是一种形式上看上去像是媒体内容的伪造信息；假信息缺乏正规媒体的编辑规范和流程以确保信息的准确和可靠性；假信息通常界定为误导性信息 (misinformation) 和虚假信息 (disinformation)。^③ 我们要求人工标注者不通过第三方渠道去检查推文，而是依据他们自己的主观判断来决定这一条推文内的信息是否是看

① Liu, Y., Ott, M. & Goyal, N., et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019.

② Zhang, Y., Liu, S., Wang, Y. & Fan, X., *Detecting Chinese Fake News on Twitter during the COVID-19 Pandemic*, 2023.

③ Lazer, D. M. J., Baum, M. A. & Benkler, Y., et al. "The Science of Fake News", *Science*, 2018, 359(6380): 1094-1096.

上去像是假信息。

这里需要指出的是,虽然人工标注数据仍然是社会科学处理数据的黄金标准,但是基于计算机辅助的数据标注已经被社会学、政治学等学科广泛使用,并已经被证明能够用来标注复杂的社会学概念。^①当然该框架可以进一步结合最新的大语言模型进展,利用生成式人工智能(例如OpenAI公司推出的GPT-4)来标注数据。如果在预算充分的情况下,研究者可以使用GPT模型来标注数据;但在有限预算的情况下,研究这可以尝试使用开源大模型,比如基于Meta AI的LLaMa2等。尽管已有研究表明生成式人工智能在开放式标注数据方面有很大的优势,但是基于特定领域微调的大预言模型仍然优于生成式人工智能^②。

对于社会科学家而言,可用的训练集有限,尤其是针对具有特定需求的训练集往往并不存在。我们没有从头开始训练一个特定领域的大语言模型,这些模型通常都是过亿的参数量。因此,我们通过迁移学习结合有限的训练集来微调预先在海量文本上训练过的大语言模型。具体说来,我们从开源AI社区Huggingface导入提前训练好的中文语言表示模型BERT的变种RoBERTa^③,然后利用人工标注的假信息训练集来微调最后的假信息分类器。我们微调了三个预训练的中文大语言模型:Chinese-bert-wwm-ext、Chinese-roberta-wwm-ext、和Chinese-macbert-large。因为RoBERTa模型优于其他两种(准确度为93.2%, F1分数为0.663),因此在最后的模型中我们使用了RoBERTa来识别所有的潜在假信息。在CNTweets的2500万条中文推文中,我们识别出约87万被标定为假信息的推文,大致占4%左右,

① Nelson, L. K., "Computational Grounded Theory: A Methodological Framework", *Sociological Methods & Research*, 2020, 49(1): 3-42.

② 依赖生成式人工智能进行判断预设了大模型自身具有价值无涉。但考虑到针对中国的谣言在西方社会比较普遍,完全基于欧美国家生成式人工智能的数据标注也存在不可控的标注偏误问题。参见Ziems, C., Held, W. & Shaikh, O., et al., *Can Large Language Models Transform Computational Social Science?*, 2023; Taubenfeld, A., Dover, Y., Reichart, R. & Goldstein, A., *Systematic Biases in LLM Simulations of Debate*, 2024。

③ 关于RoBERTa模型的技术细节,请参考Liu, Y., Ott, M. & Goyal, N., et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019。

该数据库我们称之为CNFakeTweets。图1给出了本文所采用的假信息识别框架。

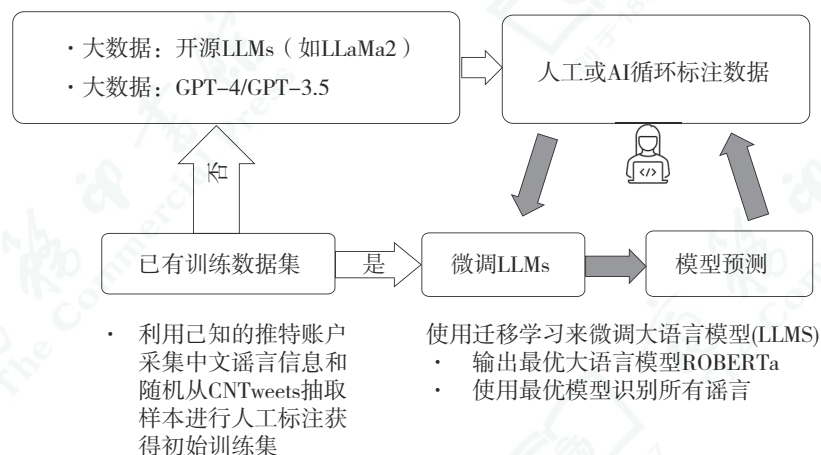


图1 人机协同的深度学习假信息识别框架

（三）一些实证发现

基于识别的假信息数据集，我们探索性解答前文对于谣言传播的几个问题。

首先，什么内容更可能是假信息？我们首先比较CNTweets数据库中关于假信息和非假信息的每日动态变化趋势。在图2中，黑色曲线展示的是假信息的动态变化情况，而灰色呈现的是非假信息的趋势图。总体来看，非假信息的推文尽管有波动，但是相对比较稳定；而假信息则波动比较大，特别是在美国确认首例新冠疫情后，中文社群中的假信息数量呈爆发式增长。接下来我们对中文社群推文进行内容分析以探索何种推文更容易被标定为假信息。如下表格呈现了我们数据库中的假信息主题分布。需要指出的是，这些主题是基于大语言模型RoBERTa的识别，每个推文可能包含多个主题。不难看出，在中文推特社群中，涉及疫情、政治、美国和宗教的议题通常容易成为假信息；在政治议题中，涉及美国的政治议题比香港和台湾议题要更加容易成为假信息。

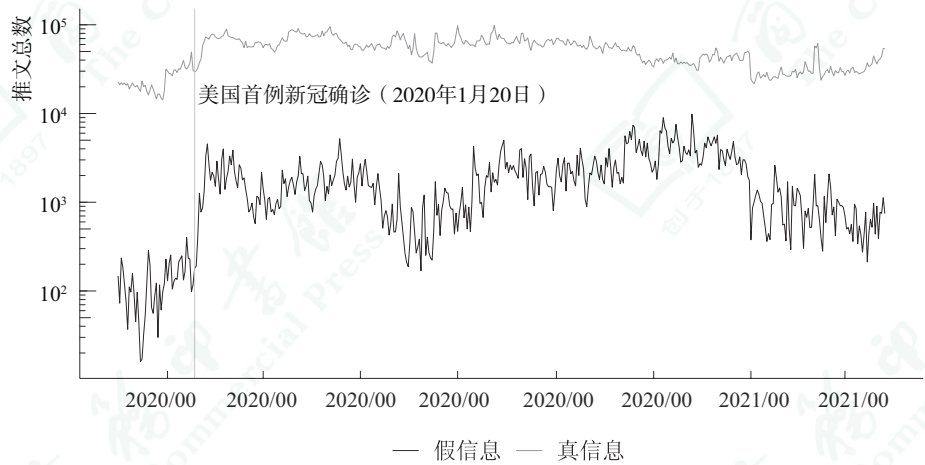


图2 CNTweets 数据库中假信息和非假信息的每日动态趋势

表1 月份固定效应回归模型预测假信息

	模型 1	模型 2
主题		
疫情	1.932*** (0.241)	
政治	1.552*** (0.130)	
美国	1.156*** (0.196)	
宗教	0.413* (0.172)	
文化	-2.264*** (0.173)	
经济	-1.223*** (0.118)	
香港议题		-1.069*** (0.157)
台湾议题		-1.607*** (0.109)
美国议题		1.754*** (0.204)
N	9553219	9553219
数据源	CNTweets	CNTweets
R2	0.178	0.106

注：分析仅保留原始推文，因此回归分析样本量和CNTweets的数据量存在差异。

表2 假信息主题分布

主题	比例
疫情	0.4616
政治	0.9449
香港议题	0.0183

续表

主题	比例
台湾议题	0.0116
经济	0.0166
文化	0.0018
宗教	0.0129
美国	0.6693

注：主题是基于RoBERTa模型识别；N=87万。

其次，什么主题的假信息更容易传播成为谣言？假信息只有经过传播才会被定义为谣言。我们分析了假信息的公众互动情况。我们首先创建谣言指数，即系统测量一条假信息的转发、回复和引用的总数。这里我们把传播了至少100次的假消息定义为谣言。表3呈现了Logistic回归结果用以预测什么样的假消息更容易成为谣言。我们发现，在中文推特社区中，涉及疫情、美国和经济的议题的谣言指数会更加高。在政治议题方面，涉及美国的议题谣言指数要比其它类型的扩散性要强。

表3 月份固定效应回归模型预测谣言形成

	模型3	模型4
疫情	0.434*** (0.067)	
政治	0.185 (0.211)	
美国	0.538*** (0.079)	
宗教	-0.523 (0.291)	
文化	0.320 (0.584)	
经济	0.392** (0.125)	
香港议题		-0.373 (0.241)
台湾议题		-0.785* (0.368)
美国议题		0.440*** (0.087)
N	141247	141247
数据源	CNFakeTweets	CNFakeTweets
R2	0.011	0.007

注：我们移除了转发的推文，仅保留原始推文。

我们进一步探索了这些谣言在3天内的转发情况。下图A比较了谣言

和非谣言的 72 小时内一条推文的平均转发次数。下图 B 呈现的是不同类型的推文在 72 小时内的转发情况。总体而言,假信息比非假信息在初始阶段的平均转发量要大,但是在 24 小时的时候,两种的转发基本持平。在我们的数据库中,涉及政治、美国和疫情相关的假信息转发数量最大,其次是经济、台湾、香港和宗教议题,最后是文化议题。

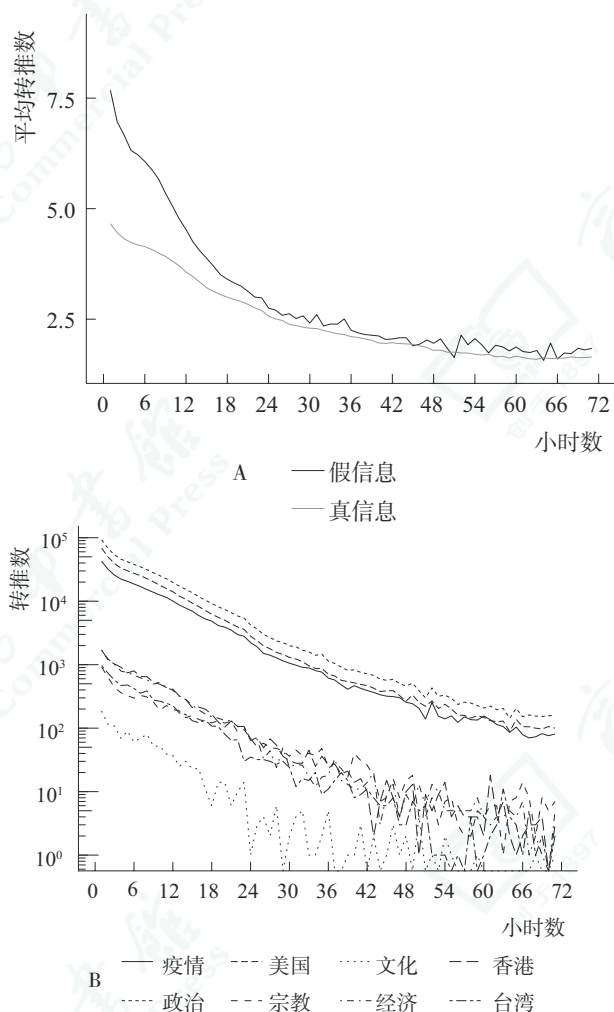


图3 中文推特社区不同类别和主题信息在72小时内转发趋势

最后,转发假信息和对中国或华人群体的态度之间有无关系?我们将关注点放到转发假信息的用户对中国持有何种态度这一问题上。CNTweets

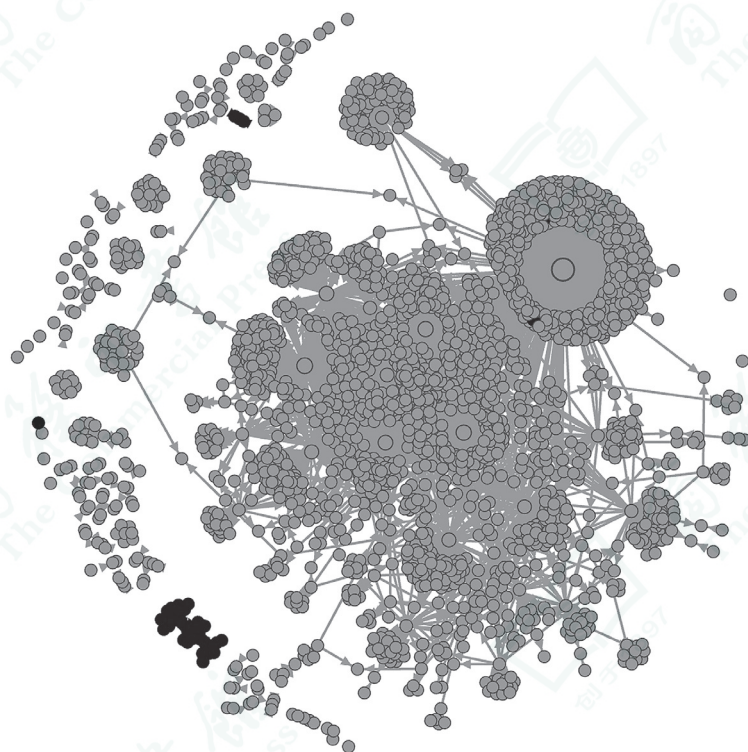
数据库利用微调的大语言模型对所有的推文进行了一系列的标注,这些标签包含对中国的情感分析(支持/反对/中立),情感对象(广义的中国、政府、中国人民)以及各种议题,包含疫情、经济、政治、香港、台湾、美国、文化、宗教等。^①基于推文对中国或华人群体态度情感分析的结果,我们把数据库中的推特用户分为对中国或华人群体持有正面和负面态度的两种类型。如果他们所发表的推文中正面的比例大于 0.6,我们标定为正向用户;如果负面的比例大于 0.6 则标定为负向。在约 130 万用户群体中,我们识别出 459,821 负向和 496,504 正向用户。表 4 呈现了 Logit 回归分析结果用以探索推文的情感和对象分析与转发假消息的关系。不难看出,正向推文成为谣言的可能性要远远低于负向推文(即更少被转发),同时如果针对的是中国人民和广义的中国,假信息的扩散性要低。

表 4 月份固定效应回归模型预测谣言形成

	模型 5	模型 6
用户对中国或华人态度		
中立	-0.135 (0.081)	
正向	-0.753*** (0.142)	
假信息针对对象		
针对中国政府		0.020 (0.097)
针对中国人民		-0.759* (0.327)
针对广义中国		-0.511*** (0.089)
<i>N</i>	141247	141247
<i>R</i> ²	0.004	0.006

图 4 呈现出假信息的转发网络,清晰表明参与假信息转发的用户绝大部分为负向用户(98.72%),参与假信息转发的正向用户极少。此外,假信息分别在正向或负向用户内部传播,正向和负向用户之间的假信息转发现象较为少见。

① 每条推文可能包含多个标签。



注：黑色为对中国态度为正向用户；灰色为对中国态度为负向的用户。

图4 CNFakeTweets 假信息转发网络

五、余论

美国著名汉学家孔飞力的著作《叫魂》向世人呈现了在清代乾隆年间关于“叫魂”的妖术谣言在民间社会所引起的恐惧和不安，以及在社会和国家层面带来的巨大冲击。无论是由于观念和科学知识的局限，抑或是对于周围的普遍怀疑，孔飞力对谣言传播的剖析已经超越了“叫魂”妖术的真实或虚幻，而是转向了对人的关注，也即为什么人们会相信这种邪恶的妖术？对于数字时代的谣言现象来说，这需要社会学研究回答两个层面的问题以理解数字时代的谣言的生产和传播：如何在网络空间的海量信息中识别出假信息，以及如何认识谣言现象的产生和传播过程乃至社会后果？

与其说本研究是对上述两个问题系统解答的尝试,不如说此研究意图从研究思路和方法上呈现计算社会科学研究范式 and 传统社会科学研究思路相结合对于数字时代谣言现象研究的意义。其中,我们在文中提出数字时代的信息“虚实之间”特征本质上仍然承接了以往研究关于谣言现象的社会心理学视角。但是,我们尝试从政治和文化社会学视角剖析数字时代所带来的社会结构和互动变化之于谣言现象频发的重要意义。通过对推特华人社区谣言的传播分析,我们进一步强调了谣言过程中个体的态度和关切对于其参与传播行为的重要性,这实际上是融合了文化社会学和社会心理学研究的一种尝试性分析。

我们希望通过本文做出两个方面的摸索,以对未来相关研究提供一定的参考。一方面,数字时代的谣言社会学研究需要融合不同学科的视角以呈现谣言现象的社会涵义及机制。尽管最近几年对假信息和谣言的研究已经成为讨论的热点,但是多数研究的侧重点往往放在了谣言现象的形式分析之上,例如谣言内容的识别方法等。承接于不同学科的研究脉络,本文呼吁未来的谣言社会学研究能够将重点从谣言的内容本身转向对人的研究。这种尝试当然需要基于既有对假信息的识别等方法发展基础之上,但考虑到数字时代谣言现象频发及其巨大的社会影响,对谣言现象的社会学分析需要探索数字时代如何形塑谣言现象的生产和传播过程,凸显对社会结构和社会心态研究的关注。另一方面,本文在方法上的尝试也希望能够为计算社会科学对于假信息识别这类具有社会学意涵的研究议题提供参考。无论是社会学还是其他学科,对信息或新闻报道具有偏向性已有共识,这导致在计算社会科学研究范式基础上开展社会学研究的过程中恰恰面临“去社会学”的悖论,也就是完全依赖于机器学习等新技术,而忽视这些新技术的应用过程是否需要社会学理论和知识的指引。这一问题在大语言模型迸发的今天尤其重要。当然,本文所提出的方法仍然是一种折中方案,也并没有就此问题展开深入讨论,因此期待未来研究能够围绕这一问题开展具有启发性的研究和讨论。

Between Truth and Falsehood: A Sociological Analysis of Rumors in the Digital Age

Abstract: The digital age has altered the methods of information acquisition and dissemination, rendering the frequent occurrence of rumors a common societal phenomenon. There is a lack of systematic discussion on the identification of rumors and their sociological implications in the digital era. This article, building on previous research, discusses the characteristics of rumors in the digital age and proposes a deep learning-based strategy for fake information identification. Taking the fake information about China and the Chinese community on Chinese-language Twitter during the pandemic as an example, we analyze the thematic distribution and dissemination process of the misinformation. We find that there is a division within the Chinese community, and that engagement in spreading rumors targeting China or the Chinese community is closely related to their attitudes towards China. This paper aims to combine the research paradigm of computational social science to offer insights for the sociological study of rumors in the digital age.

Keywords: rumor, fake information, digital age, computational social science

责任编辑: 李代