社交机器人驱动的计算宣传:社交机器人识别及其行为特征分析

卢林艳,李媛媛,卢功靖,刘熠,王成军

(南京大学,南京 210023)

摘要:近年来,社交媒体上的计算宣传行为愈演愈烈,社交机器人被广泛用于操控舆论、制造话题、转移注意力,成为社交媒体操纵的工具。准确识别社交机器人、分析社交机器人行为模式并了解社交机器人操控舆论的模式对社交机器人的治理至关重要。本研究选取2019-2020年11个热议的社会公共事件,从计算宣传的技术属性和社会属性入手,综合使用人工特征提取与深度学习方法对220万微博账号的行为特征进行提取,基于模型融合等方法建立社交机器人识别模型。在此基础上,本研究根据每个事件中的机器人占比,抽样出一个由14万微博用户构成的数据集,并建立逻辑回归模型。研究发现社交机器人识别算法的平均AUC得分为0.88。微博上社会公共事件中的机器人既聪明又傻瓜。一方面,社交机器人借助低发文率、低活跃度逃避平台管制;另一方面,它们又喜好在同一时间段群体行动,同时发布高同质性文案。从操纵策略上看主要采用扩音方式(72.2%)而非引导方式;从实际效果来看,扩音作用也显著强于引导作用。综上,目前微博上的社交机器人在热点社会事件中表现出明显的特征和较高的可预测性,其社交媒体操纵策略主要是扩大声量而非引导舆论。

关键词:计算宣传;社交机器人识别模型;社交媒体操纵;机器人行为特征

中图分类号:G2 **文献标识码:**A **文章编号:**1673-4793(2021)02-0035-10

Computational propaganda driven by social bots: social bots detection and characterization

LU Linyan, LI Yuanyuan, LU Gongjing, LV Yi, WANG Cheng-Jun

(Nanjing University, Nanjing 210023, China)

Abstract: In recent years, computational propaganda on social media has become increasingly fierce, and social bots have been widely used to manipulate public opinion, create topics and divert attention, becoming a tool for social media manipulation. Therefore, it is important to detect them and analyze the patterns of their behaviors. From both a technical and social perspective of computational propaganda, this study extracts both artificial and deep-learning features of 2.2 million Weibo accounts to establish a social bots detection model for 11 topical issues in 2019-2020. By using several methods such as dataset split and model integration, the average AUC of the model is over 0.88. Controlling the proportion of bots in each event, the study samples 140, 000 Weibo users for logistic regression and finds that the bots in social topical issues are both "smart" and "stupid". On the one hand, they conceal themselves with low activity. On the other hand, they prefer collective actions and highly homogeneous texts. They prefer to playing an amplifying role rather than a leading role, and their impact on opinion amplification is stronger than their impact on opinion leading. To sum up, social bots on microblog show obvious characteristics and high predictability in hot social events, and their strategy is mainly to expand the voice rather than to lead public opinion.

Key words: computational propaganda; social bots detection; social media manipulation; characterization

第一作者:卢林艳(1997-)女,汉族,河南省沈丘人,南京大学新闻传播学院硕士研究生。

通讯作者:王成军(1986-)男,汉族,山东省滕州市人,南京大学新闻传播学院副教授,E-mail:wangchengjun@nju.edu.cn

1 问题的提出

基于社交机器人的计算宣传(Computational Propaganda)目前已经进入公众视野并被广泛运用到各种舆论宣传中。计算宣传是综合使用算法、自动化和人工有目的地在社交网络中散布误导性信息。社交机器人是在社交网络中扮演人的身份、拥有不同程度人格属性,且与人进行互动的人工智能应用(张洪忠,2019)。已有的研究显示,社交机器人占所有Twitter账户的9%-15%(Bence,2016)。通过模仿和模拟人类在社交媒体中的行为,社交机器人有组织地与人类用户交互,依照人类操纵者的意图影响目标受众(郑晨予,范红,2020)。基于社交网络建立的社交机器人具有天然的传播基因或优势,已经被广泛地应用到宣传与舆论传播当中。社交机器人通过系统化的传播行为支持或者抹黑特定主体(师文,2020)。不同于水军,机器人或人工智能技术取代了人类在舆论宣传中的角色(赵爽,2017)。社交机器人为了达到目标会去学习与模仿人类的行为,包括但不限于传播信息和影响目标(kevin, 2016)。

政治领域的计算宣传是目前研究者主要关注的方向。为了达到目的,社交机器人向社交媒体用户传递误导性的信息或发送垃圾消息、伪造政治关注、攻击政治对手、制造趋势话题,以此来制造共识并操纵舆论(罗昕 & 张梦,2019)。社交机器人通常被用来营造虚假人气、推送政治消息、传播虚假或垃圾政治信息,制造烟雾遮蔽效应混淆公众视听(张洪忠,2019)。计算宣传同时出现在经济和文化领域。在经济方面,计算宣传常被用于操纵股票和广告市场。托马斯(2017)发现欺诈者通过在市场上传播虚假或误导信息来兜售公司股票,人为地暂时提高股价。在文化方面,Alice和Rebecca(2015)认为,互联网亚文化群体正在利用当前的媒体生态系统来操纵新闻框架、设定议程和传播思想,通过增加主流媒体上的错误信息数量的方式来降低主流媒体的可信度。然而,无论计算宣传具体应用于何种用途,其本质都在于操控舆论(罗昕 & 张梦,2019)。采用社会网络方法,研究者发现社交机器人只需要占特定讨论参与者的5%-10%,就可以改变公众舆论,使它们的观点最终成为占主导地位的观点(Cheng, Luo & Yu,2020)。

计算宣传研究对政治以外的其他领域关注较少。例如,在商业和社会事件当中往往也会卷入大量的社交机器人。另外,目前已有的研究以国外研究为主,但国内外舆论场和公众参与有较大区别(魏少华,2017)。在国内社交媒体当中,社交机器人在非政治类事件当中的影响力更大。本研究将研究视野从政治领域转移至社会领域,重点关注微博上的社会类公共事件。本研究主要致力于解决解答以下困惑:在非政治类的社会热点事件中,社交机器人驱动的计算宣传对舆论演化是否起到操纵作用以及如何发挥作用?

本文认为社交机器人作为媒介生态中的重要组成部分,已成为改变公共话语和舆论议程的一只看不见的手。这要求我们能够通过算法精准识别机器人并进行有效治理。同时,社交机器人的人格化特征又区别于一般意义上的真人,针对社交机器人这一新的行动者(agent),需要重新分析其传播模式(张洪忠等,2019)。鉴于社交机器人在社交媒体中的普遍存在,以及社交机器人可能在社会事件中对舆论起到影响乃至操控的作用,有必要对这个“无形的手”进行深入剖析。

2 理论框架

计算宣传是一种特殊的宣传形式。对于宣传的研究起源于第一次世界大战。拉斯韦尔在1927年提出宣传是指借助故事、谣言、报道、图片以及社会传播的其他形式实现意见控制的目的(拉斯韦尔,1927)。拉斯韦尔(1937)进一步将宣传视为一种操纵的技术,宣传是一种通过操纵文本的表述方式来影响人类行为的技巧。随着互联网快速发展,宣传的意义和内涵随之发生转变。伴随2016年美国总统大选,计算宣传(Computational Propaganda)作为一种新的传播方式开始出现。计算宣传以大数据和算法为支撑,由机器人假扮人类向社交媒体用户传递诱导性或欺骗性信息,试图制造共识并操纵舆论(罗昕 & 张梦,2019)。

本文将从技术属性和社会属性两个维度构建分析框架,研究微博社会热点事件中的计算宣传行为。目前关于计算宣传研究的主要知识领域包括计算机科学和社会科学两个方面。这两个研究领域分别对应着计算宣传的双重属性:技术属性和社会属性。技术属性来源自以人工智能、算法为代表的现代信息传播技术(ICT)的发展,主要关注计算宣传背后的技术问题,主要关注检测、识别、追踪社交机器人的存在;而社会属性则重点关注宣传控制舆论的内涵,将计算宣传视为利用人工智能技术操纵舆论的新途径,探讨计算宣传在政治、文化和经济领域产生的影响(罗昕 & 张梦,2019)。计算宣传既是一种影响政治和社会的技术力量,也是一种操纵舆论的宣传方式。计算宣传必须基于算法等技术手段,但主要是为了达到背后操纵者的特定目标。从传播手

段而言，计算宣传往往逾越伦理的边界故意歪曲符号、诉诸情感和偏见来绕过理性思维（Bolsover& Howard，2017）。

2.1　技术属性：社交机器人的特征与识别算法

目前社交媒体机器人账号检测致力于寻找区分度明显的检测特征，以及准确率高综合代价小的检测算法（刘蓉等，2017）。构建社交机器人识别模型首先需要基于机器人和人类行为的差异提取特征。一般而言，社交机器人的表现比起人类同质性更强（Cresci, 2019）。目前社交机器人识别主要集中在以下几个维度：

账户元特征指账户的元数据，如账号ID昵称、账号的注册时间、账号的简介、账号关注的人数、账号的粉丝数、账号的转发数量、账号发布内容的数量以及账号采用的头像图片等（王雅晗, 2019）。Beskow和Carley (2019）提出社交账号的用户名可以分为随机和非随机两种，在创建非随机的社交机器人用户名时往往遵循一套固定的命名逻辑。还有学者发现机器人账户总是在很短时间内被批量创建，且发文时定位的地理位置呈现出很强的离散性，甚至均匀分布在大西洋和无人区（Echeverria, J.，2017）。除此之外，用户的认证类型、等级和勋章以及账户发布推文的API接口等属性也是重要的账户元特征，机器人账号大多缺乏会员认证或绑定信息，且由于存活时间较短，因而等级往往较低；API接口用于判断其发文时使用的终端平台（如iPhone、Android、网页端）。机器人发文的API接口往往指向同一个平台（Bolsover, 2019）。

网络结构可以反映不同节点间的通信特征（Das A，2016），根据信息扩散模式多个维度，对于机器人识别的网络特征也分为基于关注与被关注关系的社交关系网络，基于转发、提及（@关系）和标签（#hashtag）的行动网络，建立网络并提取了他们的统计特征，例如度分布、聚类系数和中心度。Dorri, Abadi & Dadfarnia(2018）提出可以根据社交网络的同质性进行社交机器人识别。如果两个账户在一个社交网络中连接在一起，那么他们就可能具有相似的属性。一方面，机器人为了骗取人类关注自己往往会关注更多的真实人类；另一方面，机器人也会相互关注，并且关注机器人的账号大多也是机器人。

微博内容可以反映出一个账号的行为习惯及写作风格（Ghosh R, 2011）。因此，微博内容特征也是目前社交机器人机器学习检测技术中的一类常用特征，针对文本本身，如，微博内容长度、微博内容中所含的词汇数量（Jr, S. B, et al., 2018). 、微博内容相似性和重复性（Bara, Fung，& Dinh, 2015）、不同的符号数量（例如“#”等）、微博内容所含的URL链接数量以及微博内容使用的语言、语法、语义（Dickerson, Kagan，& Subrahmanian, 2014）等；除了上述文本本身特征，对文本进行进一步分析，可以挖掘出微博内容的情感特征，Loyola-Gonzalez等人（Loyola-Gonzalez, Monroy, Rodriguez, Lopez-Cuevas，& Mata-Sanchez, 2019）利用情绪分析构建了一个使用基于对比模式的分类器检测推特机器人的模型，并取得了较好的识别效果。机器人的发帖数量、规模、信息的爆发性被认为是高于人类用户。在2016年美国大选中，每天在某一特定议题下发表50篇推文被用做识别社交机器人的分界线（Howard，et al. 2016）；机器人的发文间隔最快为几秒，而人类往往需要更久的时间；社交机器人账号数量小，但发布的内容数量极多，遇到大事件时，会突然集中爆发（张洪忠等, 2020）。

综上，我们可以从社交媒体账号的元特征、网络特征、内容特征、时间特征四个维度就社交机器人进行识别，基于此提出第一个研究问题：

Q1:在社会热点事件中，社交机器人和人类在不同维度的特征上有何差异？

监督式机器学习方法是更为常见的社交机器人识别算法（Subrahmanian et al.，2016），该方法从样本中选择特征参数，建立判别函数，对未被识别的样本进行分类，能够有效利用先验数据信息，减少人为因素的干预，形成符合特征的分类模型。该方法依赖有标记的数据集，这些标签通常来自人工编码（Varol, Ferrara, Davis, et al., 2017）、自动化方法（Lee et al., 2011），或者是暴露出可疑行为的僵尸网络（Echeverria & Zhou, 2017a, 2017b）。

BotOrNot是Twitter公开的第一个检测社交机器人的接口。该系统利用Twitter的搜索接口，收集待检测账号最近的200个帖子和最近被提及的100个帖子，从网络、用户、好友、时间、内容和情感等6类特征入手，判断该账号属于恶意机器人的可能性，经过十折交叉验证后发现随机森林模型的分类效果最好（Davis et al., 2016）。此外，朴素贝叶斯算法、K近邻算法、C4.5决策树、支持向量机、随机森林算法等都已被用于识别社交机器人。Echeverria 和 Zhou（2017）选取发文内容、发文数量、粉丝和好友数量、推文来源、用户注册时间、地理位置信息等7个特征，使用朴素贝叶斯的方法对Twitter上真实用户和星球大战僵尸机器人进行研究，发现机器人账户与真实用户在地理距离和连接属性上呈现明显差异，真实用户的推文

数据呈幂率分布,而机器人呈现出均匀分布的特征。谈磊等(2012)将朴素贝叶斯模型与K近邻模型结合起来,提出了一种基于复合分类模型的算法来检测社交网络中的恶意用户。还有学者提出了多种检测网络水军的算法,包括基于黑名单的算法(Grier et al., 2010)、基于用户特征的算法(Irani, Webb & Pu,2010)以及基于文本的方法(Lau et al., 2012)等。

社交机器人处于不断演化的过程中,是否具有可预测性呢?微博热点事件中的机器人识别仍具有挑战,他们可能更隐蔽也可能更笨拙。针对性地找到合适的识别方法成为我们要解决的问题。在构建社交机器人识别特征的同时,我们想进一步追问社交机器人识别问题的可预测性。基于此,提出第二个研究问题:

Q2: 社会热点事件中社交机器人具有可预测性吗?或者说是否可以精准识别的社交机器人?

2.2 社会属性:社交媒体操纵

从社会属性来看,计算宣传与社交媒体操纵(Social Media Manipulation)紧密相联。社交媒体操纵也被称为"社交媒体宣传操纵",主要是指在社交媒体上使用自动化程序或者机器人进行蓄意宣传和虚假消息传播(Woolley & Howard,2017)。这一概念目前已被广泛应用到政治、经济、文化等多个领域。在政治领域,计算宣传通常被用来左右国家选举、营造虚假人气、煽动公众抗议、开展国际攻击(罗昕 & 张梦,2019)。2016年美国大选中特朗普和希拉里阵营利用社交机器人提高候选人支持率(Bence et al,2016)、英国公投期间大量社交机器人参与推动其退出欧盟(Cadwalladr,2017)、乌克兰危机运动中社交媒体上出现大量假新闻(Khaldarova & Pantti,2016)、巴西公众抗议运动中社交机器人号召人们上街游行(Oliveira et al,2016)、叙利亚战争中社交机器人通过发布大量无关的推文弱化抗议者的声音并转移公众注意力(Michael, 2017)。2017年,包括美国在内,至少有18个国家在选举中遭遇了线上操纵和虚假信息(House,2017)。

微博平台凭借庞大的用户群和公开性成为计算宣传的理想场所。本文立足于国内,以新浪微博为研究对象,新浪微博被称为中国的推特,为用户提供了一个参与讨论的平台,同时也为计算宣传提供了空间(Bolsover & Howard, 2019)。Bolsover在2013年以新浪微博为研究对象对一项新闻的传播网络进行分析,发现6%的新闻报道由虚假账号转发,30%的意见领袖实则是用来操纵信息的机器人账号。2018年,Bolsover和Howard(2019)研究了推特上110万条与中国政治标签相关的帖子以及国内官方新闻微博下150万条评论。他们发现推特上存在大量反中国政府的社交机器人,这些社交机器人致力于传播对抗中国政府的虚假信息。

从以上研究我们可以发现不同类型事件中的主体不同、计算宣传的目的不同,社交媒体操纵的表现也不相同。另一方面,多数研究集中在政治领域,而忽略了对非政治领域的社交媒体操纵问题。本研究认为针对社会公共事件中社交媒体操纵进行研究对更好理解计算宣传具有重要意义。基于此,我们提出以下问题:

Q3:微博社会热点话题事件中是否存在社交媒体操纵现象?可以分为哪些基本类型?哪一种类型的操纵作用最强?

3 研究方法

本文首先利用2019-2020年备受关注的社会公共事件数据集训练社交机器人识别算法模型,为更多数据打上是否是机器人的标签;在此基础上借助逻辑回归模型和统计检验分析社交机器人的行为特征,区分社交机器人的行为表现以及对人类的影响。

3.1 数据

本文选取了2019-2020年11个的社会公共事件,包括"西安奔驰漏油"、"上海特斯拉自燃"、"视觉中国版权风波"、"北大弑母案"等引起公众广泛关注与讨论的热点事件。所有数据集总计包含220万个微博用户,除了微博文本内容外,还包括根微博内容、账号信息、发布时间等辅助性信息。由于机器人在不同性质事件中参与行为存在差异,我们将事件分为正面、负面、中性三个类型加以控制。

在保证分析的维度后,我们对数据集进行规范性处理。由于微微博内容本是一种非正式的文本,其内容可能仅仅是一句话、一个表情或是一张图片,还可能包括"话题标签#"、"提及@"以及网页链接等不规范的文本内容,而这些不规范的文本会给文本内容识别带来阻滞。为了去除文本噪音,更好识别微微博内容本真正的文本意涵,我们对微博内容进行包括停用词在内等常规的预处理,同时还进行以下的额外处理:

(1)对表情包与网址进行语言化处理

微微博内容本中存在大量的表情包等图片形式语言,同时有研究发现:机器人可能会较多地使用表情包,表达某种情绪来模糊视点(Yang,2019)。这些表情包表

达的意义各不相同，如不对其处理，将无法捕捉到该文本特征。因此，我们将所有文本中的表情包替换成对应表情的文字，例如“吃惊”的表情包被替换成“吃惊”；文本中的网页链接被替换成“网址”。

（2）“转发微博”字样文本针对性处理

由于微微博内容本包含大量的带有“转发微博”字样的文本，实则是没有意义的空转发。因此，在计算与“其他微博的文本相似性”特征时将剔除“转发微博”文本；但在计算“账号自身文本相似性”时，由于有部分机器人只发布空转发，因此将保留该部分数据。

（3）处理类别不平衡问题

由于各数据集均存在不同程度的类别不平衡问题，这会导致模型无法有效识别少数类。通过调整正样本权重，在保留样本的真实分布的同时，缓解类别不平衡问题，并选取 AUC 作为模型评价指标。

3.2　特征提取

除了账号属性、参与程度、时间特性与文本内容四类特征之外，结合微博数据特点，构造出根微博特征，同时将BERT模型的输出结果作为特征之一输入模型，以期在更大范围内识别机器人用户的立场和动机。具体来说，账户属性主要包括性别、粉丝数、关注数、认证类型等属性信息。参与程度指账号发布微博条数、用户使用平均数量、用户使用总数量。在时间特征上，不同于以往研究中常使用的发文时间，本模型从时间维度的“爆发性”维度建构，提取平均一天发文数量、同时发布的平均微博数两个特征。而根微博即原创微博，是转发微博的最初来源，我们使用平均根微博转发数量、平均根微博不同微博占比、微博来源同质性三个变量衡量根微博特征。此外，本模型采用 TF-IDF 来构建文本内容的特征，包括“平均词汇多样性”、“其他用户与自己微博平均重复次数”、“与机器人微微博内容本相似性”(BERT特征）等。

3.3　模型结构

区别于其他社交机器人识别模型，本算法基于人为提取特征与BERT特征，将数据集根据转发率拆分为原创集、转发集、全集并分别进行特征筛选与建模。最后，通过数据集融合与模型融合两个步骤，得到最终的输出结果。关键步骤流程为：

（1）Cleanlab标签清洗

通过观察微博样本发现，部分机器人标签标注并不是那么准确。而保证训练集标签的准确性是提升模型准确度的关键，我们借助Cleanlab进行训练集的标签清洗，以各个类的平均预测概率为概率阈值，估计出已知的有噪声标签和未知的无噪声标签之间的联合概率分布，并对可能错误的标签进行清洗（Northcutt C G，2019）。

（2）数据集拆分

在社交机器人识别中，转发率高账号与转发率低账号的行为模式与关键特征显著不同，如将两类账号混合训练会降低模型的分类性能。因此，本模型将数据集拆分为全集、原创集和转发集，对三类数据集分别建模并将结果融合。在特征重要性方面，本文发现转发集中根微博特征最为重要，原创集则以BERT特征最为重要，全集中BERT特征重要性提高。通过拆分数据集，整体特征重要性有较大变化，这表明数据集拆分有其合理性。

本模型采用嵌入法筛选特征，利用 LightGBM 克服特征多重共线性。每一个数据集最终均输出 6 个结果，分别为全集、转发集、原创集在 LightGBM 和 CatBoost 模型上的结果，依次进行数据集融合和模型融合，得到最终的结果。具体模型流程图1所示。

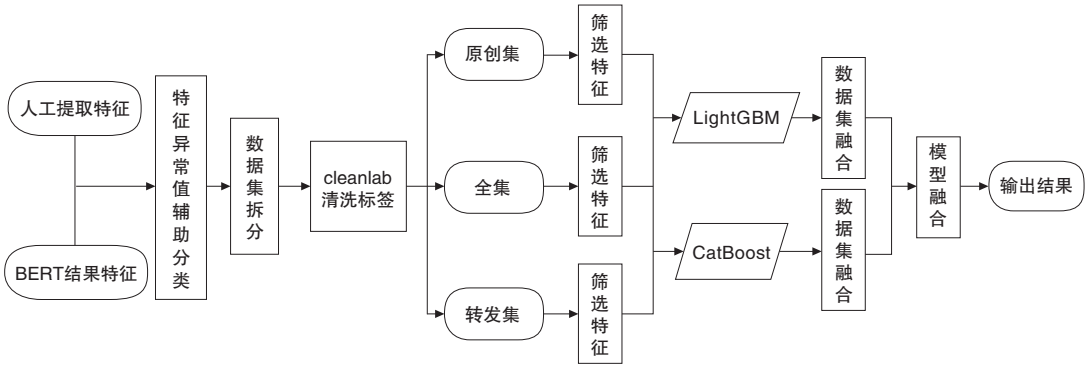

图1　模型流程图

4 研究发现

4.1 机器学习模型结果

因为数据集存在类别不平衡问题,本模型采用AUC作为模型评价指标。经过模型融合,在不修改验证集标签的情形下,最终我们的模型在数据集上的AUC在0.815-0.967之间,平均AUC为0.88,AUC数值的波动可能与数据集大小、数据集分布、数据集标签准确度有关。如果使用Cleanlab清洗验证集可能错误的标签,模型在各个数据集上的AUC均可达到0.9以上,平均AUC为0.928。最终,模型能够较为精准识别社会热点事件中的社交机器人。因此,对于Q2,本研究发现社交机器人具有较高的可预测性。

4.2 回归模型

为了回答Q1,本研究构建了逻辑斯蒂回归模型。借助社交机器人账号识别模型,我们给微博用户打上是否是机器人的标签,总计2205025个微博用户(含机器人),按照每个事件原本的机器人比例进行随机抽样,获得了144111个微博用户(含机器人),在此基础上采用逻辑回归模型对抽样数据进行分析。逻辑回归可以理解为参数θ在已知x的条件下比较$P(y=1|x,\theta)$和$P(y=0|x,\theta)$概率大小,选择较大的概率作为分类结果。逻辑斯蒂回归(logistic regression)的概率公式:

$$P=\frac{1}{1+e^{-w^{T}x}} \quad (1)$$

其中是否是机器人二分类标签是因变量,账号属性、参与程度、时间特性、根微博、文本内容五大类共14个特征为自变量,同时引入事件性质虚拟变量(正面、负面、中性)进行控制(表1)。

(1)、在账号属性方面,机器人的认证类型基本上全部为普通用户(β=9.15, P<.001)。此外,社交机器人大多伪装成男性参与公共事件讨论(β=0.66, P<.001),这可能是因为男性为建立微博账号时的默认选项。

(2)、在参与程度方面,社交机器人账号在发布微博数量方面没有人类积极(β=-0.04, P<.01),机器人账号发布的平均微博条数(1.17)比人类账户(1.41)低。机器人在模仿人类"@"行为与"参与话题#"行为上表现不积极。社交机器人不像人类用户一般建立好友关系网、与其他账户互动并产生联系。机器人几乎不参与"@"行为(β=-1.76, P<.001),转发率也相对较低(β=-0.56, P<.001)。

(3)、从时间维度上看,社交机器人平均一天发文量却显著高于人类(β=0.18, P<.001)。

(4)、在转发的根微博方面,社交机器人转发的根微博独特文本占比较低(β=-1.40, P<.001),转发的根微博来源同质性水平更高(β=1.12, P<.001)。

(5)、在文本内容方面,社交机器人微博平均词汇多样性低(β=-0.42, P<.001),与其它社交机器人微博文本相似性高(β=5.77, P<.001)。

(6)、在事件属性方面,社交机器人更倾向于关注高情感极性的事件。例如,与中性事件相比,社交机器人对负面事件(β=0.3, P<.001)和正面事件(β=0.41, P<.001)的关注度更高。

表1 罗杰斯蒂回归结果

	变量	回归系数
	常数	−12.36 (1.01)
账号属性	认证类型(普通用户=1)	9.15 (1.00)
	性别(男性=1)	0.66 (0.02)
参与情况	微博条数	−0.04 (0.02)
	@平均数目	−1.76 (0.07)
	#平均数目	−0.01 (0.01)
	转发率	−0.56 (0.06)
时间特性	平均一天发文数量	0.18 (0.03)
	同时发布的平均微博数	0.00 (0.00)
根微博	根微博转发微博独特文本占比	−1.4 (0.04)
	根微博累计被转发平均次数	0.00 (0.00)
	微博来源同质性水平	1.12 (0.08)
文本内容	他人与自己微博平均重复次数	0.00 (0.00)
	平均词汇多样性	−0.42 (0.07)
	与机器人微博文本相似性	5.77 (0.05)
事件属性 (以中性事件为对照组)	负面事件	0.3 (0.03)
	正面事件	0.41 (0.03)

4.3 热点事件中的社交媒体操纵水平

以机器人和人类都转发过的根微博为样本,以小时为单位,分别统计机器人和人类转发同一条根微博的高峰时间,机器人转发高峰早于人类转发高峰意味着该条根微博一开始集中被机器人转发,后期大量人类参与转发,机器人在影响人类;机器人转发高峰等于人类转发

高峰即表示不存在明显的前后影响，机器人与人共同参与话题讨论，推动事件热度升级；机器人转发高峰晚于人类转发高峰则表示人类率先关注某话题，机器人跟风参与，机器人扮演扩声器而非意见领袖。

最后的结果如图 2 所示，在 11 个热点事件中，机器人转发高峰早于人类的占比低，高峰时间一致的占比高，人类转发高峰早于机器人的占比次之，由此可知，热点事件中的机器人并未像政治事件中一样以较高比例主动发起话题，更多用来浑水摸鱼，以乌合之众的身份和人类一同推动事件热度发酵。

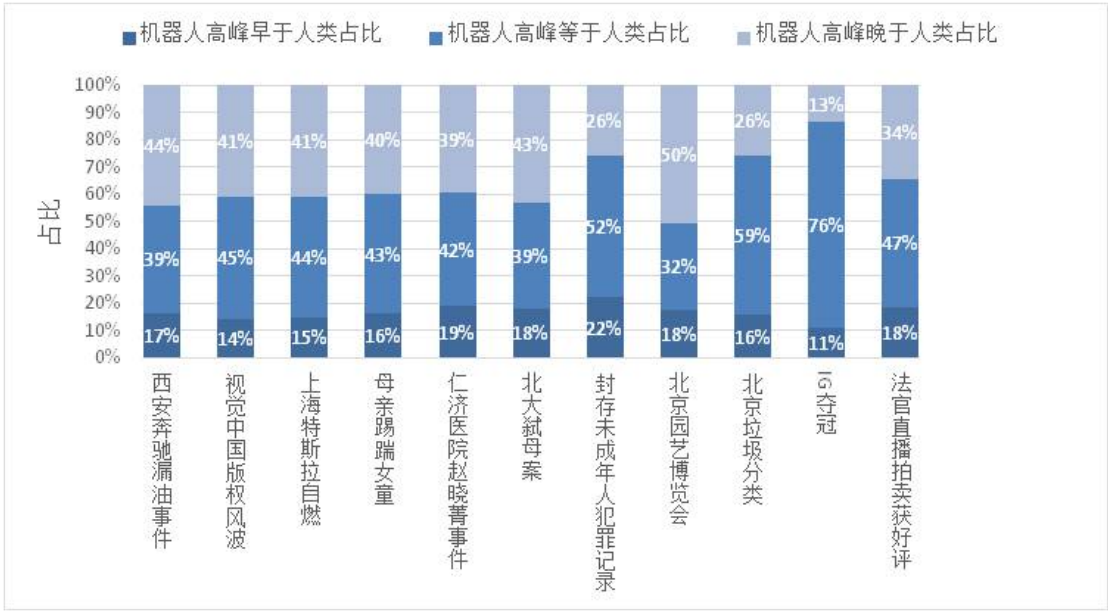

图 2　机器人发文高峰与人类发文高峰比较

虽然机器人和人类转发根微博的时间先后能够说明机器人究竟是在引导人类还是充当人类的扩音器，但单纯的时间差并不能说明机器人的宣传效果大小。比如，在机器人转发高峰早于人类转发高峰的情况下，当一条根微博率先被 50 个机器人转发，但最终过了 5 个小时只吸引到了一个人类转发，而另一条根微博同样率先被 50 个机器人转发，在 1 小时内就吸引到了 40 个人类转发，引发了人类转发高峰，那么很显然第二条根微博的机器人转发对引导人类转发起到了更大的作用，宣传效果更大；而在机器人转发高峰晚于人类的情况下，当一条根微博首先被 50 个人类转发，而后被 1 个机器人转发，另一条根微博同样首先被 50 个人类转发，而后一个小时内被 100 个机器人转发，那么第二条根微博机器人转发对于扩大人类声量的作用更大，宣传效果更强。

首先，我们设置了机器人引导效果这个指标，通过计算人类高峰转发次数所对应的单位时间机器人的转发数量表示机器人的宣传效果，计算公式如下：

$$\text{引导效果} = \frac{R_{human}^{Peak}}{(t_{human}^{Peak} - t_{robot}^{Peak}) \times R_{robot}^{Peak}} \quad (2)$$

其中，R_{human}^{Peak} 代表人类转发高峰时的转发次数与人类参与者数量之间的比例，R_{robot}^{Peak} 表示机器人转发高峰时的转发次数与机器人数量之间的比例，t_{human}^{Peak} 代表人类转发高峰时间，t_{robot}^{Peak} 代表人类转发高峰时间。

当机器人转发某一根微博的高峰时间早于人类转发的高峰时间时（即人类与机器人的转发高峰时间差为正时），我们即认为机器人对人类产生了引导作用；单位时间单位机器人转发影响的人类越多，引导效果的值越大。

当该根微博被机器人转发的高峰时间晚于人类转发的高峰时间时（即人类与机器人的转发高峰时间差为负时），机器人失去引导效果，此时我们可以采用类似的公式计算机器人扩音效果（或跟随效果）：

$$\text{扩音效果} = \frac{R_{robot}^{Peak}}{(t_{robot}^{Peak} - t_{human}^{Peak}) \times R_{human}^{Peak}} \quad (3)$$

对 11 个微博热点事件，我们共分析了被机器人和人类共同转发过的 14546 条根微博。排除 5818 条机器人和人类同一小时转发的根微博之后（无明显操纵作用），发现其中 2426 条根微博的机器人宣传效果为引导作用，剩下 6302 条根微博的机器人宣传效果为扩音作用。这说明社交机器人主要采用的是扩音策略（占比约为 72.2%）。进一步比较了引导作用和扩音作用哪一个更大，结果发现社交机器人的扩音作用也明显高于引导作用（$t(8726)=-3.64$, $P<.000$）。

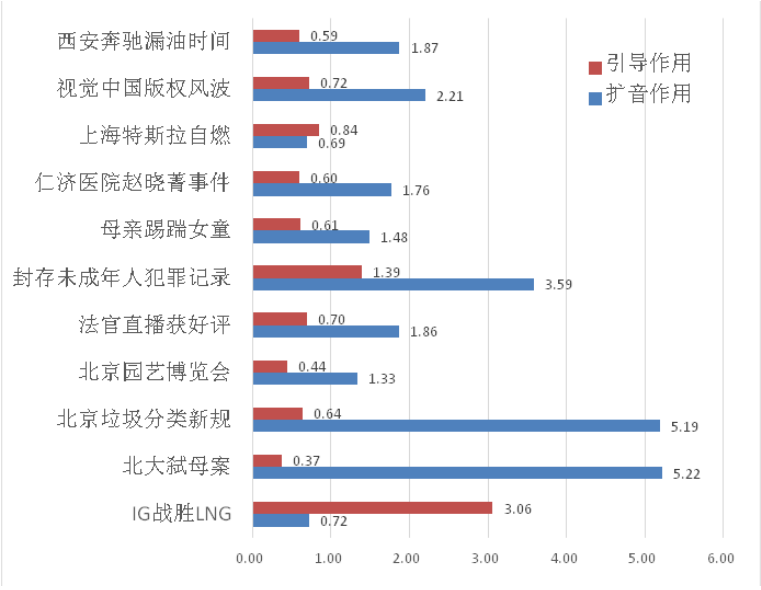

图3 机器人宣传效果

此外,本研究还采用格兰杰因果检验的方式分析每条微博扩散的过程中,究竟是社交机器人影响人类(引导作用)还是人类影响社交机器人(扩音作用)。基于此,可以得到不同宣传效果的根微博在该事件中所占的比例。对以上格兰杰检验的结果进行T检验也证实扩音作用强于引导作用($t(20)=-3.16$, $P<0.000$)。综上,社交机器人的扩音作用引导作用使用更广、效果更强。

5 结论

一方面,社交机器人不是传播中的意见领袖,更接近喧哗的大众。社交机器人试图将自己以普通用户(或大众)的身份隐藏起来。但为了完成其舆论操纵作用,社交机器人又必须表达观点。社交机器人避免过多的互动,行为表现上却像是大众社会里孤立的个体。机器人发声的目的不是创造某种观点,而是让某种观点变得引人瞩目或者通过关注特定话题模糊视线。机器人账号使用者另辟角度切入热点话题,通过提高话题暴风眼外围微博的曝光量,吸引公众注意力,例如在"上海特斯拉自燃事件"中,正常人类用户主要在讨论和吐槽特斯拉汽车质量问题和新能源电池安全问题,而机器人主要关注和转发特斯拉自燃事件车主状况和起火原因调查。从行为方面来看,社交机器人在某些程度上比人类表现更加懒惰。通常我们认为社交机器人会不眠不休的全天候进行工作,直到操纵者下令停止。但研究发现,社交机器人往往具有更少的发文数量和更低的活跃度。这可能是因为微博等社交平台的反垃圾信息机制的制约,低活跃率可以避免被现有的算法识别并封禁。

另一方面,社交机器人会在特定时间段内更加活跃,同时行动完成任务目标。例如在"2019北京园艺博览会"相关话题微博中,很多机器人在转发同一根微博时频繁使用表情包、"赞"、"不错"、"支持"、"太喜欢了"等文案,呈现出整齐划一的排队阵势。此外,社交机器人参与的微博内容的同质性和文本相似性水平较高,而词汇多样性低。

从社交机器人的宣传策略来看,社交机器人普遍采用的是舆论扩音策略而非引导策略,但从效果来看扩音作用也明显强于引导作用。进一步,我们发现机器人主要起引导作用的事件是那些需要改变人们观点的事件,如IG战胜LNG夺冠、封存未成年人犯罪记录等;而那些机器人主要起扩音器作用且扩音作用比较强的事件中,主要为北大弑母案、北京垃圾分类这些主要目的在于进行信息普及的事件。由于目的的不同,机器人会采取不同的操纵策略。

计算宣传现象的兴起影响着人们公共生活的参与,准确识别、捕捉社交机器人,掌握社交机器人的行为特征、动机诉求对于我们更好地治理社交机器人至关重要。与仅使用人工特征的社交机器人识别模型不同,我们构建的机器人识别模型在使用人工特征的基础上,引入BERT模型并将模型结果特征化使用,同时,借助Cleanlab清洗训练集标签,结合特征异常值辅助分类、数据集拆分、模型融合等方法进一步提高了模型性能,最终建立平均AUC为0.88的社交机器人识别模型。

与国外社交媒体上常见的政治机器人不同,社会公共事件中的机器人既"聪明"又"傻瓜"。一方面,机器人

借助低发文率、低活跃度掩饰其本质；另一方面，社交机器人总是群体行动、以缺乏互动的“傻瓜”行径暴露其身份。在社交媒体操纵影响方面，大量社交机器人发文高峰早于人类，但其舆论引导作用弱于扩音作用。需要注意的是，社交机器人识别与反识别是一场持续的博弈。随着技术的进步和社会需求，社交机器人也会不断进化，本文的研究发现在更大程度上应该被看作研究的起点。

参考文献（References）：

［1］ 方师师.社交媒体操纵的混合宣传模式研究[J].现代传播(中国传媒大学学报), 2018, 40(10):143-150.

［2］ 刘蓉, 陈波, 于泠, 刘亚尚, 陈思远.恶意社交机器人检测技术研究[J].通信学报, 2017, 38(S2):197-210.

［3］ 罗昕, 张梦.西方计算宣传的运作机制与全球治理[J].新闻记者, 2019(10):63-72.

［4］ 师文, 陈昌凤.分布与互动模式：社交机器人操纵Twitter上的中国议题研究[J].国际新闻界, 2020, 42(05):61-80.

［5］ 谈磊, 连一峰, 陈恺.基于复合分类模型的社交网络恶意用户识别方法[J].计算机应用与软件, 2012, 29(12):1-5.

［6］ 魏少华.对话理论视域下的中国社交媒体“话题”功能研究[D].华东师范大学, 2017.

［7］ 王雅晗.社交机器人检测技术研究及实现[D].北京邮电大学, 2019.

［8］ 赵爽, 冯浩宸.“机器人水军”发展与影响评析[J].中国信息安全, 2017, (11):88-89.

［9］ 郑晨予, 范红.从社会传染到社会扩散：社交机器人的社会扩散传播机制研究[J].新闻界, 2020, (03):51-62.

［10］ 张洪忠, 段泽宁, 韩秀.异类还是共生:社交媒体中的社交机器人研究路径探讨[J].新闻界, 2019, (02):10-17.

［11］ 张洪忠, 赵蓓, 石韦颖.社交机器人在Twitter参与中美贸易谈判议题的行为分析[J].新闻界, 2020, (2): 7.

［12］ Alice Marwick and Rebecca Lewis. Media Manipulation and Disinformation Online[Z]. Data & Society Research Institute, 2015.1.

［13］ Bara I A, Fung C J, & Dinh T. Enhancing Twitter spam accounts discovery using cross-account pattern mining[A]. 2015 IFIP/IEEE international symposium on integrated network management (IM), IEEE, 2015. 491-496.

［14］ Bence Kollanyi, Philip N Howard, and Samuel C Woolley. Bots and Automation over Twitter during the Third US Presidential Debate[EB/OL]. http://comprop. oii. ox. ac. uk/wp-content/uploads/sites/89/2016/10/Data-Memo-Third-Presiden - tial-Debate.pdf

［15］ Beskow D M, & Carley K M. Its all in a name: detecting and labeling bots by their name[J]. Computational and Mathematical Organization Theory, 2019, 25(1), 24-35.

［16］ Bolsover G, & Howard P. Chinese computational propaganda: automation, algorithms and the manipulation of information about chinese politics on twitter and weibo[J]. Information, Communication & Society, 2018, 1-18.

［17］ Cadwalladr C. The Great British Brexit Robbery: How Our Democracy Was Hijacked[EB/OL]. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-rob - bery-hijacked-democracy，2017—05—07.

［18］ Cheng C, Luo Y, & Yu C. Dynamic mechanism of social bots interfering with public opinion in network[J]. Physica A: Statistical Mechanics and its Applications, 2020: 124163.

［19］ Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. Emergent properties, models, and laws of behavioral similarities within groups of twitter users[J]. Computer Communications. (DOI:10.1016/j.comcom.2019.10.019)

［20］ Das A, Gollapudi S, Kiciman E, et al. Information dissemination in heterogeneous-intent networks[A].//Proceedings of the 8th ACM Conference on Web Science. ACM, 2016. 259-268.

［21］ Dickerson J P, Kagan V, & Subrahmanian V. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? [A]. Proceedings of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining [C]. IEEE Press, 2014.620-627.

［22］ Dorri A, Abadi M, & Dadfarnia M. Socialbothunter: Botnet detection in Twitter-like social networking services using semi-supervised collective classification[A]. 2018 IEEE 16th international conference on dependable, autonomic and secure computing. IEEE, 2018. 496-503.

［23］ Echeverria J, & Zhou S. Discovery, Retrieval, and Analysis of the 'Star Wars' Botnet in Twitter[A]. In Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining, 2017. 1-8.

［24］ Ghosh R, Surachawala T, Lerman K. Entropy-based classification of retweeting activity on twitter[J/OL]. arXiv preprint, arXiv:1106.0346, 2011.

［25］ Gillian Bolsover & Philip Howard. Chinese computational propaganda: automation, algorithms and the manipulation of information about Chinese politics on Twitter and Weibo[J]. Information, Communication & Society, 2018, (5):1-19. (https://www.tandfonline.com/doi/pdf/10.1080/136 9118X.2018.1476576? needAccess=true)

［26］ Grier C, Thomas K, Paxson V, & Zhang M. spam: the underground on 140 characters or less[A]. In Proceedings of the 17th ACM conference on Computer and communications security [C]. 2010. 27-37.

（下转第 53 页）