

情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：政治安全语境下社交媒体虚假叙事传播账号识别研究
作者：曾思诚，杨建林
网络首发日期：2026-09-04
引用格式：曾思诚，杨建林．政治安全语境下社交媒体虚假叙事传播账号识别研究[J/OL]．情报杂志. https://link.cnki.net/urlid/61.1167.G3.20260903.1339.002



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2026-09-04 09:46:46　网络首发地址：https://link.cnki.net/urlid/61.1167.G3.20260903.1339.002

政治安全语境下社交媒体虚假叙事传播账号识别研究*

曾思诚[1,2]　杨建林[1,2]

（1. 南京大学信息管理学院　南京　210023；
2. 江苏省数据工程与知识服务省高校重点实验室（南京大学）　南京　210023）

摘　要：[**目的**]本文旨在构建面向政治安全语境的虚假叙事传播账号识别方法，以提升对相关目标账号的识别能力并揭示其传播规律。[**方法**]面向政治安全相关事件构建领域专属数据集，设计“机器初筛—人工精标”的两阶段标注机制，利用 Bot-Finder 对全量账号进行一级初筛，在此基础上结合人工研判形成目标账号标签；进一步在通用行为特征基础上，引入权威叙事对抗性、政治情绪动员性、叙事框架协同性与关键节点介入性四类特征，构建分层特征体系，并结合不平衡学习模型开展账号识别，通过基线对比、消融实验与特征重要性分析不同特征维度的作用。[**结果/结论**]研究结果表明，在目标账号占比较低的情况下，引入虚假叙事操纵特征后，模型在召回率（Recall）与 F1 值等关键指标上实现明显提升；其中，关键节点介入性与叙事框架协同性对模型性能贡献最为突出。目标账号在关键时间窗口集中发声、持续输出具有煽动性的相似表达，并呈现跨账号内容协同等特征，体现出较强的策略性叙事操纵倾向。

关键词：政治安全；社交媒体；虚假叙事传播；账号识别；叙事操纵；分层特征

中图分类号：G350　　**文献标识码：**A

Research on the Identification of False Narrative Dissemination Accounts on Social Media in the Context of Political Security

Zeng Sicheng　Yang Jianlin

（Nanjing University, Nanjing　210023）

Abstract: [**Purpose**] The study aims to construct an identification method for false narrative dissemination accounts in the context of political security, improve the identification of relevant target accounts, and reveal their dissemination patterns. [**Method**] The study constructed a domain-specific dataset based on political-security-related events and designed a two-stage labeling mechanism that combined machine pre-screening with manual annotation. Specifically, the study used Bot-Finder to conduct the first-stage screening of event-related accounts and then applied manual judgment to generate target-account labels. On this basis, the study introduced four types of features, namely, authoritative narrative opposition, political emotion mobilization, narrative frame coordination, and critical node intervention, in addition to general behavioral features, and constructed a hierarchical feature system. The study then used imbalanced learning models for account identification and examined the effects of different feature dimensions through baseline comparison, ablation experiments, and feature importance analysis. [**Result/Conclusion**] The results show that when target accounts make up only a small proportion of the dataset, the introduction of false narrative manipulation features significantly improves the model's key metrics, such as recall and F1 score. Among these features, critical node intervention and narrative frame coordination make the most prominent contributions to model performance.

基金项目：南京大学新时代文科卓越研究计划“中长期研究专项”项目“‘数智赋能’背景下的情报学理论、方法与应用研究”（编号：010814914214）；南京大学中国特色哲学社会科学自主知识体系建构“引领工程”重大研究专项培育项目“中国自主的情报学知识体系构建研究”（编号：2024181013）研究成果。

作者简介：曾思诚，男，2003 年生，硕士研究生，研究方向：国家安全情报，网络舆情传播；杨建林，男，1970 年生，博士，教授，博士生导师，研究方向：情报学基础理论，信息分析与科学评价，国家安全情报。

通信作者：杨建林

Further analysis shows that target accounts tend to post intensively during critical time windows, continuously produce similar inciting expressions, and exhibit cross-account content coordination, indicating a strong tendency toward strategic narrative manipulation.

Key words: political security; social media; false narrative dissemination; account identification; narrative manipulation; hierarchical features

网络空间中的信息流动、舆论演化与价值建构,已成为政治安全保障的重要内容。从情报学视角看,网络舆情安全治理正在由单一信息监测转向多源数据融合、态势感知、风险理解与主动塑造相结合的综合治理过程[1]。从非传统安全视角看,网络意识形态安全所面临的风险集中体现为多元舆论竞争、价值观冲突加剧以及复杂信息环境对主流认知结构的冲击[2]。在此背景下,认知战逐渐成为政治安全竞争的重要表现形式,其核心在于通过信息干预、意义建构与情绪塑造影响公众的认知框架与价值判断,争夺对象也由“信息控制权”转向“认知塑造权”与“解释主导权”[3]。已有研究表明,认知安全风险会沿着“个体认知—社会共识—国家政治安全”的路径逐级放大[4]。由此,虚假信息已不再只是离散的信息失真单元,而日益演化为服务于认知塑造的叙事资源;其通过符号重组、情绪强化与责任归因等方式,围绕政治敏感议题构建具有策略性与欺骗性的虚假叙事,并借助算法放大、情绪极化与群体共识建构等机制,在网络空间形成具有误导性的舆论结构[5]。

社交媒体平台及其技术体系则为上述过程提供了关键基础。一方面,平台算法推荐、社交关系链、视觉化传播与情绪扩散机制,使特定叙事得以在“回音室”与“信息茧房”中被持续放大,并增强其对用户感知和判断的渗透力[6];另一方面,社交机器人与生成式人工智能技术的发展,进一步降低了虚假内容生产与传播的门槛,使相关操纵行为呈现出更强的规模化、隐蔽化与拟人化特征[7]。由此,在政治事件的风险传播主体中,相较于传统以自动化行为异常为主要表现的异常账号,一类更具欺骗性的传播主体正在受到关注。其往往以发布、转发等方式实质性参与讨论,并传播与官方通报或权威事实核查存在偏离的叙事内容;其异常性更多体现为围绕特定议题持续发声、强化情绪表达、实施内容协同与策略性介入。基于此,本文将其界定为“虚假叙事传播账号”。对此,传统基于通用行为特征或单一内容判别的方法已难以实现有效识别,有必要在政治安全语境下,从叙事操纵的角度对相关账号开展识别研究。

基于上述问题,本文围绕政治安全相关网络事件,构建领域数据集并设计分层特征体系,在目标账号占比较低的条件下开展识别实验,并进一步分析其操纵规律。本文旨在同时从识别方法与行为机制推进相关研究:一方面提升政治安全语境下虚假叙事传播账号的识别能力,另一方面揭示其在议题传播中的策略性操纵特征。

1 相关研究

1.1 社交机器人与异常账号识别相关研究

围绕社交媒体平台中的社交机器人与异常账号识别问题,学界已形成较为系统的研究基础。早期研究主要将其视为具有自动化发布、关系异常或行为失衡特征的传播主体,并围绕账户资料、发布行为、文本内容和社交网络结构等维度展开识别,逐步形成了基于特征、基于图结构、基于异常模式、众包标注和逆向工程等主要技术路径。Ferrara E 等归纳了社交机器人检测体系,指出网络特征、用户特征、时序特征、内容特征和情感特征是机器学习识别的重要依据[8];Javed D 等进一步指出,当前研究已扩展到异常检测、图学习和深度学习等更复杂的方法体系[9]。

在具体方法上,基于特征的识别方法通过构建账户属性、行为模式与内容特征等多维指标体系进行判别,如 Zhou J 等利用账户静态属性、动态行为和发布内容等特征,结合 XGBoost 模型实现社交机器人的检测[10]。基于图的识别方法则通过关注、转发、提及等互动关系构建网络结构,并利用图论算法或图神经网络识别异常主体,如 Guyan Q 等通过增强外围网络信息利用提升了中心节点的检测性能[11];杜瑞成等进一步引入关系图保留机制与语义邻域融合,提升了复杂结构中的识别稳健性[12]。此外,Pan J 等仅利用异常突发性参数和时间间隔熵值即可区分人机账户[13]。Alarifi A 等通过众包标注构建高质量训练数据[14],Bello B S 等则采用逆向工程从英国脱欧后的 Twitter 数据中提取机器人行为规则[15]。近年来,相关研究进一步引入联邦学习、多模态融合和类别不平衡处理等策略,以提升跨平台识别能力、鲁棒性与少数类识别效果[16-18]。总体来看,现有研究已在异常账号与社交机器人识别方面形成较成熟的方法体系,但其研究重心多集中于账号属性、行为模式和网络结构等通用异常特征。面对政治事件中兼具异常传播表征与叙事操纵功能的目标账号,相关研究仍有必要进一步强化面向政治安全语境的特征设计与任务适配。

1.2 虚假信息传播与操纵行为相关研究

另一类研究更多从虚假信息传播、舆论操纵与叙

事干预的角度,关注异常账号如何在具体议题中塑造舆论结构与认知导向。许灵毓等基于舆论形成过程模型指出,社交机器人可在舆论出现、影响力争夺和舆论形成等阶段持续介入,并对信息传播速度和网民意见倾向形成产生显著影响[19]。

在具体政治议题研究中,已有文献揭示出操纵主体的若干典型行为特征。Stella M 等发现,社交机器人会在政治事件中,有针对性地向关键用户推送负面和煽动性内容,以提升冲突性叙事的触达程度[20];Bello B S 等人的分析也表明,相关机器人在脱欧议题中呈现出明显的策略协同性和定向传播特征[15]。张梦晗等人的研究指出,不同立场的社交机器人能够围绕俄乌冲突中特定的战争叙事展开主题引导和议程竞争[21]。在中文语境下,围绕涉华议题[22]和公共卫生突发事件[23]的研究同样表明,推动特定内容扩散社交机器人更倾向于参与争议性话题,并通过转发、提及和标签操作试图提升议题可见度,尤其能推动带有情感倾向的特定内容扩散。

进一步看,相关研究也关注异常账号如何在具体议题中通过情绪和协同机制参与舆论操纵。研究发现:社交机器人评论的情感极性会显著影响用户参与行为[24];共情传播、情绪感染与群体认同是推动网络舆情快速聚集和极化的重要机制[25];且网络匿名性等去抑制因素会降低用户对网络评论行为后果的关注,并强化情绪驱动下的表达冲动,提升非理性评论发生的可能,从而推动舆论的进一步发酵[26]。此外,既有研究表明,异常账号正逐步转向兼具情绪动员、议题设置与叙事强化功能的操纵主体[27-29]。不过,现有研究多停留于影响机制解释和个案分析层面,尚未将情绪动员、内容协同等传播规律系统转化为面向账号识别的可操作特征体系。

1.3 现有研究不足

现有研究已分别在社交机器人与异常账号识别、虚假信息传播机制以及舆论操纵行为分析等方面取得了较为丰富的成果,但围绕政治安全语境下“虚假叙事传播账号识别”这一问题,仍存在进一步深化空间。

其一,在研究对象上,现有识别研究多以高频自动化、关系异常或具有典型机器人表征的账户为主要目标,对那些能够嵌入公共议题讨论、具备较强拟人化特征,并通过发布、转发等方式持续参与虚假叙事传播的目标账号关注仍显不足。其二,在特征体系上,既有研究虽已逐步引入文本、情感和网络结构等变量,但总体仍偏重账户属性、行为模式和一般内容特征,对政治安全语境下更具解释力的叙事操纵性特征缺乏系统刻画。其三,在研究路径上,现有研究尚未形成在政治安全相关场景下从异常账号识别到传播机制分析的统一分析链条,现有数据集也难以适应政治安全这一具体研究场景。

2 研究框架与方法设计

围绕政治安全语境下社交媒体平台虚假叙事传播账号识别的问题,本文构建了由“领域数据集构建—分层特征提取—不平衡模型识别—操纵规律分析”构成的研究框架,如图 1。考虑到虚假叙事传播账号并不必然表现为传统意义上的低拟人化机器人,其异常性更多体现在特定政治事件中的叙事偏离、情绪动员、内容协同与时间策略性介入,仅依赖通用行为特征难以实现有效识别。基于此,本文首先围绕政治安全相关网络事件构建领域专属数据集,并采用“机器初筛—人工精标”的两阶段标注机制形成目标账号标签,同时提升标签质量与任务适配性;其次,在通用行为特征基础上引入权威叙事对抗性、政治情绪动员性、叙事框架协同性与关键节点介入性四类特征,构建分层特征体系;再次,针对目标账号占比较低的类别不平衡问题,引入多种不平衡学习模型开展比较与验证;最后,在识别结果基础上进一步分析目标账号的操纵规律。

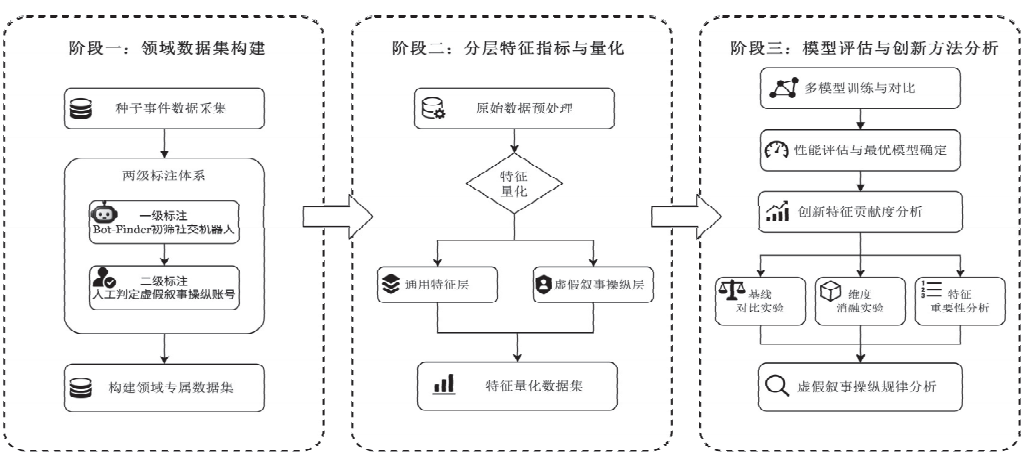


图 1 研究技术路线图

2.1 领域数据集构建

虚假叙事传播行为具有较强的情境依赖性与语义复杂性,通用社交机器人数据集虽然能够较好刻画自动化行为异常,却难以充分反映政治事件中目标账号在叙事操纵层面的功能特征。因此,围绕具体政治事件构建领域专属数据集,是开展虚假叙事传播账号识别的基础。本文从数据来源、预处理与标签生成三个环节出发,构建与研究任务相匹配的数据集。

2.1.1 数据来源与预处理

在数据来源上,本文以政治安全相关网络舆情事件为切入点,综合考虑议题的政治关联性、政治敏感度、社会影响面以及叙事演化复杂性,选取具有明确事实依据和权威结论支撑的典型事件作为数据采集对象。研究以微博平台为主要数据源,围绕事件关键词进行检索,通过网络爬虫技术对相关事件的贴文进行全量抓取。在此基础上,进一步获取参与事件讨论用户的公开个人资料信息,并追溯其历史行为,采集时间范围覆盖自回溯起点至数据采集截止时间的完整发文记录,同时采集用户的互动信息,包括转发、评论及点赞等社交行为数据。

为提高数据质量并满足后续特征提取和一级初筛的需要,本文对原始数据进行了系统化预处理。首先,在账号层面,剔除因封禁、注销或不可访问而无法获得完整信息的失效账号。其次,对账号基本资料、贴文内容和互动记录进行结构化整理,统一字段格式。再次,在文本层面,对 URL 链接、@ 提及、话题标签、空白字符等噪声信息进行清洗,并对表情符号和特殊字符作统一编码处理。最后,对发文时间进行标准化处理,构建连续发文序列。

2.1.2 两阶段标注机制

针对虚假叙事传播账号识别中标注成本高、任务判定依赖语境的问题,本文采用“机器初筛—人工精标”的两阶段标注机制。

第一阶段为机器初筛。本研究采用本地化改造后的 Bot-Finder 工具对事件样本中的账号进行自动化筛选。Bot-Finder[23] 是面向中文社交媒体环境开发的社交机器人检测工具,基于 XGBoost 算法构建识别模型,能够依据账号属性、内容发布和行为模式等 24 维特征识别具有社交机器人式异常行为表征的账号。原研究以微博平台中的社交机器人账号与真人账号为二分类对象构建测试集,XGBoost 模型在该任务中的准确率(Accuracy)为 92.93%、召回率(Recall)为 92.34%、精确率(Precision)为 93.72%、F1 值为 92.97%。其中,准确率指测试集中社交机器人账号与真人账号被正确分类的总体比例;召回率指测试集中实际为社交机器人账号的样本中被模型成功识别出的比例;精确率指模型判定为社交机器人账号的样本中实际为社交机器人账号的比例。

本研究在实际使用中并未采用该工具的实时在线爬取功能,而是将前期围绕特定政治事件爬取并整理后的账号公开资料、历史发文与互动信息等结构化数据,以批量 JSON 文件形式输入本地化改造后的程序,程序进一步在将这些数据整理为 Bot-Finder 模型所需的数据结构之后,送入工具模型进行预测。模型输出取值分布于-10~10 之间的 Bot Score;本文依据工具原有判别规则,将 Bot Score 大于 0 的账号纳入“可疑账号集合”。需要说明的是,Bot-Finder 输出结果仅表明账号具有一定通用异常行为表征,并不等同于最终确认其为虚假叙事传播账号。因此,本文仅将其作为缩小人工研判范围的前置步骤,以降低后续人工标注工作量并提高标签构建的针对性。

第二阶段为人工精标。在一级初筛结果基础上,本文依据账号在特定政治事件中的行为参与情况与叙事内容属性,对可疑账号进行逐一判定。为降低人工研判的主观性,本文以官方通报、权威媒体报道和事实核查结论作为权威事实参照,并将目标账号判定细化为三个方面:其一,账号需以发布、转发等方式实质性参与事件讨论;其二,账号传播内容需与权威事实参照在事实判断、责任归因或事件解释上存在明显偏离;其三,账号行为需体现出对相关偏离性叙事的扩散、重复或情绪强化。对于仅表达一般情绪、提出合理疑问或虽参与事件讨论但未形成明确叙事偏离的账号,不纳入目标账号。最后,对于语义含混或证据不足的边界样本,本文采取保守判定原则,并结合账号事件相关贴文、上下文语境和权威事实参照进行多轮回查,以尽可能减少人工判断偏差。

基于上述标注路径,本文研究对象主要聚焦于那些具有异常传播表征,并在具体政治事件中承担虚假叙事传播功能的目标账号。该标注机制既保留了机器初筛在效率上的优势,也通过基于权威事实参照的人工研判增强了标签与研究问题之间的匹配程度,为后续识别模型训练提供了较为可靠的数据基础。

2.2 分层特征体系构建

考虑到目标账号并不必然表现为典型的自动化传播主体,其异常性不仅体现在一般行为模式上,也体现在围绕具体议题实施叙事操纵的过程中,本文构建“通用特征层—政治事件虚假叙事操纵特征层”的分层特征体系,以实现对目标账号的多维刻画。

2.2.1 通用特征层

通用特征层主要借鉴社交机器人与异常账号识别研究中广泛采用的指标体系,从身份伪装、社会关系、内容基础与活动强度等方面刻画账号的一般行为模

式,如表1。该层特征主要用于识别账号是否具有异常传播表征,为后续识别提供基础支撑。

表1　通用特征层(＊为创新特征)

一级指标	二级指标	三级指标	来源文献
Layer 1:通用特征层	身份伪装	个人信息完整性	[7][10][30][31]
		昵称自主性	[7][10][31]
		头像异常度	[7][14][31]
		认证状态与阳光信用等级	[7][10][23]
		账号等级	[23]
	社会关系	粉丝数	[23]
		关注数	[23]
		粉丝-关注比	[7][23]
		获赞、评论、转发的累计/均值/方差	[23]
	内容基础	帖子字数均值和方差	[7][10][23][31]
		帖子@数量均值和方差	[14][31]
		同时@二人及以上的帖子数量	[14][31]
		帖子话题标签数均值和方差	[14][31]
		特殊符号与链接占比	[30][31]
		句子数量均值/方差	[30]
		多模态程度	[7][10]
		话题多样性	＊
	活动强度	日均发帖量	[7][10]
		帖子编辑率	[10]
		转发/原创比	[23][31]
		转发频率	[7]
		活动来源多样性	[32]
		账号活跃天数比例	[7]
		异常互动频率	[10]

在完成上述分层特征提取后,为降低冗余信息对模型训练的干扰,并缓解高维特征可能引发的过拟合问题,需要在保持关键解释变量的前提下,剔除与标签相关性较低、区分能力有限或信息重复度较高的指标,以提升模型训练效率与泛化能力。

2.2.2　政治事件虚假叙事操纵特征层

在通用特征层基础上,本文进一步构建政治事件虚假叙事操纵特征层,用于刻画目标账号在政治事件中的叙事特征异常。前文“相关研究”部分指出,政治议题中的异常账号会通过推送负面和煽动性内容提升冲突性叙事的触达程度[20],通过话题引导[21]、定向传播和策略协同[15]增强特定叙事的可见度,并在舆论形成、影响力争夺和事件发酵等关键阶段持续介入[19]。由此可见,政治事件中的虚假叙事操纵通常同时涉及内容偏离、情绪动员、框架协同和节点介入等多个层面。基于此,本文将政治事件虚假叙事操纵特征概括为四个维度:权威叙事对抗性、政治情绪动员性、叙事框架协同性与关键节点介入性,如表2所示。

其中,权威叙事对抗性用于衡量账号发布内容相对权威叙事的语义偏离程度,反映其是否与主流解释框架形成对抗性表达;政治情绪动员性用于刻画其是否通过负向情绪、政治情感异常偏差和情感强度波动放大事件争议性,进而增强传播感染力并影响公众认知判断;叙事框架协同性用于识别其是否围绕特定政治议题持续发声,并与其他账号形成内容复用、表达呼应或框架趋同;关键节点介入性则用于衡量其是否在舆情高潮、事件转折或讨论集中阶段呈现集中介入和活跃节奏异常的传播特征。

表2　政治事件虚假叙事操纵特征层(＊为创新特征)

一级指标	二级指标	三级指标	计算说明	来源文献
Layer 2:政治事件虚假叙事操纵特征层	权威叙事对抗性	权威叙事语义距离	政治事件贴文与“权威通报”语义向量的曼哈顿距离	＊
		负面情感极性密度	政治事件贴文中负面情感词占比	[7]
	政治情绪动员性	政治情感异常偏差	政治事件贴文的情感得分与该话题平均情感得分的差值	[33]
		情感强度波动	账号历史帖子情感得分的方差	＊
		政治议题持续参与度	账号在该事件发帖数量是否超过1	＊
	叙事框架协同性	政治议题聚焦度	账号历史帖子两两之间语义向量的平均余弦相似度	＊
		政治议题内容复用度	政治事件贴文与其他账号所发布贴文的最大余弦相似度	＊
		发文突发参数	历史发帖时间突发程度:$B=(\sigma-\mu)/(\sigma+\mu)$,其中$\sigma$和$\mu$为该账号历史发贴的时间间隔序列的均值和标准差	[13]
		时间间隔熵	历史发帖时间间隔随机程度:$E=-\sum_{i=1}^{M}p(\tau_i)\log(p(\tau_i))$,其中$\tau_i$为该账号发帖时间间隔	[13]
	关键节点介入性	深夜发帖率	历史发帖中深夜发帖的比例	[10]
		政治议题舆情高潮窗口介入度	(1)舆情高潮耦合度:政治事件贴文发帖时间与事件发展高潮中心的耦合程度,取值范围[0,1],越接近1表示发帖越集中在高潮中心。(2)峰值时间距离:政治事件贴文发帖时间到峰值的最近时间:取值范围$[0,+\infty)$。	＊

注:表2中语义相似性相关指标均基于SBERT预训练模型生成文本向量后计算;负面情感极性密度基于大连理工大学情感词典识别,情感得分采用经微调的Chinese RoBERTa-wwm-ext-base模型量化;舆情高潮窗口依据事件相关贴文每小时发帖量确定。

2.3　模型构建与验证

2.3.1　模型选择与训练方法

由于虚假叙事传播账号在整体样本中的占比较

低,识别任务具有典型的类别不平衡特征。为提升模型对少数类目标账号的识别能力,本文选取多种不平衡学习方法构建候选模型集合,包括 EasyEnsemble、Balanced Random Forest、RUSBoost、以 Logistic Regression 为基学习器的 Balanced Bagging 模型、以 Decision Tree 为基学习器的 Balanced Bagging 模型和 XGBoost 等,并在统一特征体系下开展训练与比较。

2.3.2 模型验证与结果解释方法

为评估模型在虚假叙事传播账号识别任务中的表现,本文采用五折交叉验证,并从整体区分能力与少数类识别能力两个层面选取 ROC 曲线下面积(AUC)、精确率—召回率曲线下面积(PR-AUC)、精确率(Precision)、召回率(Recall)和 F1 值作为主要评价指标;其中,精确率(Precision)是指模型判定为虚假叙事传播账号的样本中,经人工精标确认为目标账号的比例;召回率(Recall)是指全部人工确认的虚假叙事传播账号中,被模型成功识别出的比例。由于目标账号占比较低,PR-AUC 对少数类样本更加敏感;同时,考虑到政治安全相关场景更强调对目标账号的尽早识别,本文在模型比较中相对更重视召回率及 PR-AUC 表现。

在模型验证与可解释性方面,本文主要从三个层面展开分析:一是通过基线对比实验检验政治事件虚假叙事操纵特征对识别效果的增益作用;二是通过消融实验比较不同叙事操纵维度对模型性能的贡献;三是结合特征重要性分析识别影响分类结果的关键变量,从而揭示模型判别的主要依据。

2.3.3 虚假叙事操纵规律分析

在实现对虚假叙事传播账号的有效识别之后,本文进一步关注其在政治安全语境下的行为机制问题,即目标账号如何通过特定策略参与信息传播并影响舆论演化。具体而言,本文结合虚假叙事操纵特征,从权威叙事对抗性、政治情绪动员性、叙事框架协同性与关键节点介入性四个维度,对目标账号的特征分布与变量关系进行分析。具体采用相关性分析与结构归纳相结合的方式,总结虚假叙事传播账号的主要操纵特征。

3 实证研究与结果分析

3.1 事件选取与领域数据集描述

本文选取两个具有代表性的政治安全相关公共舆情事件作为实证研究对象。两个事件均涉及公共治理与制度运行问题,在网络空间中引发了较高讨论度和明显争议;同时,事件本身具有较为明确的官方通报、权威媒体报道或事实核查依据,可为虚假叙事传播账号的人工研判提供相对稳定的参照。另外,二者分别属于公共安全和社会保障这两个不同的治理维度,有助于检验识别模型在异质化语境下的泛化能力。

在数据获取方面,本文以微博平台为数据来源,围绕事件关键词采集相关贴文,并通过用户溯源获取参与讨论账号的公开资料,同时回溯采集其自事件发生前一个月起至数据采集截止日的历史发文及互动信息。具体而言,公共安全争议事件 A 的事件相关贴文采集时间范围为 2025 年 8 月 2 日至数据采集截止日;在识别参与事件讨论的账号后,进一步获取其公开个人资料,并回溯采集其自 2025 年 7 月 2 日至数据采集截止日的历史发文及互动信息。社会保障议题事件 B 的事件相关贴文采集时间范围为 2025 年 7 月 22 日至数据采集截止日;在识别参与事件讨论的账号后,进一步获取其公开个人资料,并回溯采集其自 2025 年 6 月 22 日至数据采集截止日的历史发文及互动信息。经数据清洗、失效账号剔除与样本去重后,最终形成包含两个事件子集的领域数据集。其中,公共安全争议事件 A 包含账号样本 943 个,社会保障议题事件 B 包含账号样本 2055 个,共计 2998 个账号样本;进一步采集到参与事件讨论账号的历史微博数据共 2,492,777 条。

在标签构建方面,本文基于“机器初筛—人工精标”的两阶段标注机制形成目标账号标签。具体而言,先通过 Bot-Finder 对全量账号进行一级初筛,得到具有异常传播表征的可疑账号集合;再依据权威事实参照,进一步判断账号是否同时具备前文设定的三项账号判定标准:事件参与、叙事偏离和传播强化,对于仅表达一般情绪、提出合理疑问或证据不足的边界样本,本文采取保守判定原则,不纳入目标账号。不同阶段账号样本的具体数据见表 3:经一级初筛,共有 341 个账号进入可疑账号集合,占总样本的 11.37%;经二级人工精标后,最终确认虚假叙事传播账号 176 个,占总样本的 5.87%,占可疑账号集合的 51.61%。这一结果表明,目标账号在整体样本中占比较低,识别任务具有典型的类别不平衡特征。需要说明的是,本文将研究对象限定为具有通用异常行为表征且可被有效观测的虚假叙事传播账号。因此,未进入可疑账号集合的账号在本文标签构建中均被作为非目标样本处理。

表 3 两阶段标注流程中的账号样本分布

标注阶段	公共安全争议事件 A	社会保障议题事件 B	合计	占总样本比例
事件参与账号总样本	943	2055	2998	100.00%
一级初筛可疑账号	75	266	341	11.37%
二级精标目标账号	52	124	176	5.87%

3.2 模型训练与表现

在完成分层特征提取后,本文首先依据特征与标签之间的相关性及特征间共线性,对区分能力较弱或信息冗余较高的指标进行筛选。最终剔除的特征包括

粉丝—关注比、获赞均值与方差、获评方差、获转发均值与方差、“@”两人及以上的帖子数量、特殊符号与链接占比、转发与原创比、负面情感极性密度及发文连续性等指标。经筛选后的特征进入后续模型训练。

考虑到目标账号占比仅为 5.87%,本文未采用简单过采样或文本生成式数据增强方法扩充正类样本。其原因在于,虚假叙事传播账号标签依赖具体事件语境、叙事偏离程度与行为参与特征,若通过文本生成或文本替换方式扩展样本,可能破坏原始叙事结构并引入标签噪声。因此,本文基于真实样本,引入多种不平衡学习模型开展比较实验,并统一采用五折交叉验证进行评估。

如表 4 的实验结果表明,各模型均表现出较强的识别能力,说明所构建的特征体系具有良好的判别效力。其中 EasyEnsemble 综合表现最佳,PR-AUC 为 0.5738,F1 值为 0.5997,召回率(Recall)达到 72.22%,精确率(Precision)为 51.95%。相较之下,XGBoost 虽在精确率上表现较优,但召回率偏低。说明 EasyEnsemble 在处理高度不平衡样本时,能够更有效地学习少数类特征,在降低漏报风险方面具有明显优势。

表 4　各模型性能评估

模型	AUC	PR_AUC	F1 值	召回率	精确率
EasyEnsemble	**0.9532**	**0.5738**	**0.5997**	**0.7222**	**0.5195**
BalancedRF	0.9481	0.5418	0.5266	0.5179	0.5507
RUSBoost	0.9342	0.4418	0.4758	0.5911	0.4042
EnsembleLR	0.9339	0.5187	0.4917	0.5914	0.4367
BalancedBagging_DT	0.9425	0.4606	0.5105	0.5410	0.4933
XGBoost	0.9532	0.5951	0.5603	0.5065	0.6518

进一步从不平衡学习视角分析,随机分类器在 PR 空间中的期望精确率约为正例占比,即 5.87%。本文最优模型精确率(Precision)达到 51.95%,约为随机基线的 8.9 倍,表明模型已具备显著的少数类识别能力。同时,模型 PR-AUC 为 57.38%,较随机基线提升近 10 倍,说明模型能够稳定地区分正负样本。

3.3　虚假叙事特征有效性分析

为系统验证本文所构建分层特征体系在政治事件虚假叙事传播账号识别任务中的有效性与必要性,本节从基线对比实验、叙事维度消融实验以及特征重要性分析三个层面展开分析。通过实证检验,揭示虚假叙事操纵特征在模型判别过程中的作用机制,并进一步验证其在政治安全语境下的理论解释力与实践价值。

3.3.1　基线对比实验

为验证虚假叙事操纵特征对目标账号识别的增益效果,本文设置基线模型与创新模型进行对比。其中,基线模型仅使用通用特征层(Layer 1),即账号身份伪装、社会关系、内容基础与活动强度等通用行为特征;创新模型则在 Layer 1 基础上进一步引入政治事件虚假叙事操纵特征层(Layer 2),形成“通用行为特征+叙事操纵特征”的分层特征体系。实验结果如表 5 所示。

表 5　基线对比实验

实验	AUC	PR_AUC	F1 值	召回率	精确率
基线模型(Layer1)	0.9483	0.5637	0.5512	0.6600	0.4850
创新模型(Layer1+Layer2)	0.9532	0.5738	0.5997	0.7222	0.5195
指标提升	**+0.0049**	**+0.0101**	**+0.0485**	**+0.0622**	**+0.0345**

表 5 的实验结果表明,引入 Layer 2 特征后,模型在各项指标上均实现不同程度提升。其中召回率的提升幅度最大,达到 6.22%。这证明了仅靠通用社交机器人特征难以捕捉社交媒体上政治领域深层的叙事操纵逻辑,而 Layer 2 特征补齐了这一短板。上述结果具有重要方法论意义:在虚假叙事传播账号识别这一高度不平衡任务中,召回率的提升意味着模型对风险账号的捕捉能力显著增强,有助于降低漏报风险,从而更契合政治安全领域“风险预警优先”的现实需求;而 F1 值与 PR-AUC 的同步提升表明,引入叙事特征并未以牺牲精确率为代价,而是在整体判别性能上实现优化。

进一步分析可发现,单纯依赖 Layer 1 特征虽可识别部分具有典型“机器行为模式”的账号,但对于那些在行为层面具备“类人特征”、但在叙事层面存在操纵意图的账号,其识别能力存在不足。相比之下,Layer 2 特征通过引入“权威叙事对抗性”、“政治情绪动员性”、“叙事框架协同性”与“关键节点介入性”等维度,弥补了传统方法在语义与策略层面的缺失。

3.3.2　叙事操纵维度消融实验

在验证政治事件虚假叙事操纵特征整体有效性的基础上,本文进一步通过消融实验分析各叙事操纵维度对模型性能的贡献。具体方法为:在完整特征体系(Layer 1 + Layer 2)基础上,分别剔除四个叙事操纵维度,并比较模型性能变化。表 6 中,除首行以外的其余各行表示剔除相应维度后相较完整模型的指标变化。

表 6　叙事维度消融实验

实验	AUC	PR_AUC	F1 值	召回率	精确率
完整特征	0.9532	0.5738	0.5997	0.7222	0.5195
剔除权威叙事对抗性特征	−0.0004	−0.0022	−0.0061	−0.0224	+0.0025
剔除政治情绪动员性特征	+0.0003	−0.0016	−0.0081	−0.0289	+0.0019
剔除叙事框架协同性特征	−0.0011	−0.0031	−0.0193	−0.0452	−0.0013
剔除关键节点介入性特征	−0.0036	−0.0083	−0.0435	−0.0394	−0.0362

表 6 结果显示,剔除任一虚假叙事维度后,模型的 F1 值和召回率总体均出现下降,说明四类政治事件叙

事操纵特征均对目标账号识别具有贡献。其中,“关键节点介入性”维度影响最为显著,其移除后 F1 值下降 4.35%,召回率下降 3.94%,精确率下降 3.62%,同时 PR-AUC 亦出现明显下降。这表明时间维度的异常行为是识别虚假叙事操纵的重要判据。这表明,目标账号在关键时间窗口内的集中介入、发文节奏异常和非常规时间活跃,是模型识别的重要依据。

叙事框架协同性维度的贡献同样较为突出。移除该维度后,召回率下降 4.52%,F1 值下降 1.93%。该结果表明,操纵账号之间在政治议题内容复用度、政治议题聚焦度等方面呈现出较强协同特征,这种“群体一致性”是区分真实用户与操纵账号的重要依据。相比之下,权威叙事对抗性与政治情绪动员性对整体性能的影响幅度相对较小,但其移除同样会导致召回率与 F1 值下降,说明这两个维度在捕捉权威叙事偏离与情绪动员方面仍具有重要作用。特别是在政治安全语境下,虚假叙事往往通过构建与权威叙事的对立关系,并借助情绪强化机制扩大传播效果。

总体来看,消融实验说明,虚假叙事传播账号的识别不能依赖单一维度完成,而需要综合考虑语义、情绪、协同和时间等多类信息。本文所构建的四维政治事件虚假叙事操纵特征体系在结构上具有互补性:权威叙事对抗性刻画内容相对权威叙事参照的偏离,政治情绪动员性刻画事件讨论中的情绪放大,叙事框架协同性刻画跨账号表达呼应与框架趋同,关键节点介入性刻画传播节奏中的策略性介入;四类特征共同提升了模型对目标账号的识别能力。

3.3.3 特征重要性分析

为进一步揭示模型的判别依据,本文基于最优模型对所使用特征进行重要性排序分析。考虑到 Easy-Ensemble 由多组基于随机欠采样构建的子分类器组成,本文递归提取其内部基决策树的特征重要性分数,并在子分类器内部及不同子分类器之间进行平均汇总;同时,在五折交叉验证框架下对各折结果进一步求平均,以形成较为稳健的全局特征重要性排序。结果如图 2 所示。

图 2 的结果显示,Layer 2 政治事件虚假叙事操纵特征整体贡献度显著,其重要性普遍位于全体特征中位数以上,且大多数特征高于平均水平。具体来看,“政治议题聚焦度”在全部特征中位列第 5 位。这一结果表明,虚假叙事传播账号往往围绕单一政治议题持续输出内容,呈现出高度集中的话题分布特征。相比之下,真实用户通常具有更为分散的话题兴趣,其发文内容在主题上更具多样性。在关键节点介入性维度上,“发文突发参数”“深夜发帖率”“时间间隔熵”等特征同样具有较高贡献度,进一步印证了操纵账号在行为节奏上的异常性。此外,“政治情感异常偏差”“情感强度波动”等情绪类特征亦表现出较高重要性,说明情绪动员是虚假叙事传播的重要手段。操纵账号通过强化负面情绪或制造情绪波动,以吸引用户注意力并引导认知判断。

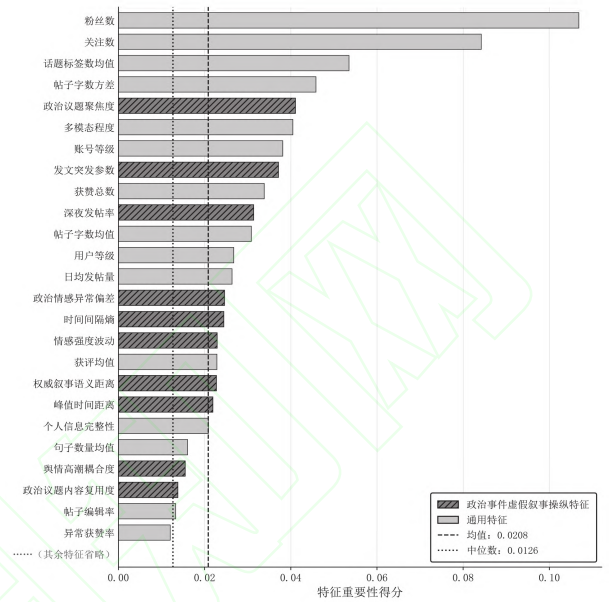


图 2 特征重要性排序图

与此同时,部分通用行为特征如粉丝数量、关注数量、账号等级等仍具有较高重要性,说明通用特征层能够从账号社会关系结构、平台身份属性和一般活动模式等方面提供背景信息,在识别具有异常传播表征的账号时仍具有基础价值。

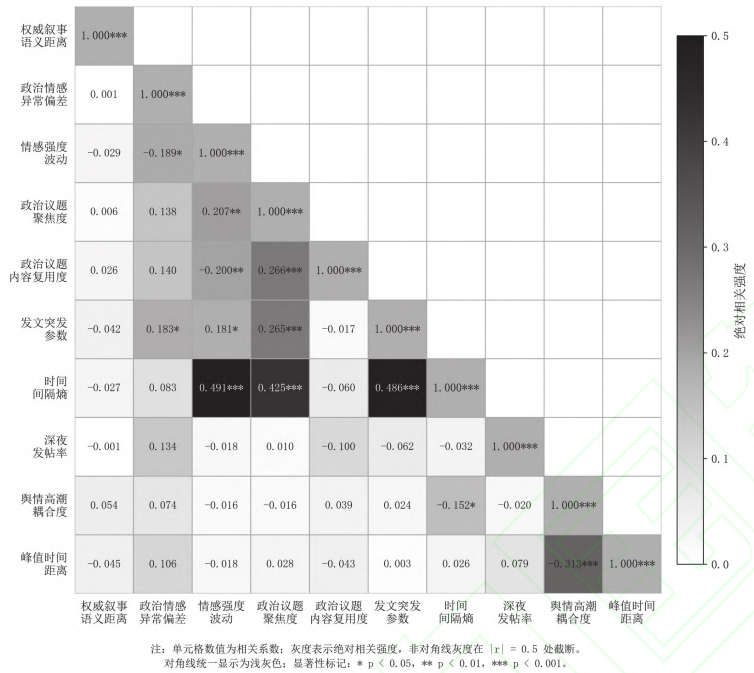
3.4 虚假叙事操纵规律分析

前文已经从模型表现、基线对比、消融实验和特征重要性等方面验证了本文识别方法的有效性。进一步需要回答的是,这些虚假叙事传播账号在具体政治事件中呈现出怎样的传播行为与操纵规律。为此,本文基于政治事件虚假叙事操纵特征层,对目标账号的 Layer 2 特征进行相关性分析,并结合统计显著性结果归纳其主要行为模式。分析结果如图 3 所示。

整体来看,显著相关关系主要集中于叙事框架协同性、关键节点介入性与政治情绪动员性的部分指标之间,而权威叙事对抗性指标“权威叙事语义距离”与其余变量之间未形成显著相关。这一结果表明,虚假叙事传播账号的操纵规律并不主要表现为单一的立场对抗或语义偏离,而更多体现为围绕特定议题,通过内容组织、时间调度与情绪配置等方式实施立体操控。换言之,在政治安全议题传播中,其风险并不完全来源于“说了什么”,还来源于“如何持续说”“何时集中说”以及“以何种情绪方式说”。

首先,虚假叙事传播账号在发布话题和节奏上呈现出“话题聚焦与节奏突发”的特征。实证结果显示,

政治议题聚焦度与发文突发参数之间存在显著正相关关系,并与时间间隔熵呈显著正相关。这表明,目标账号越是围绕单一议题持续发声,其发文行为越表现出更强的突发性和更随机的时间间隔。这说明虚假叙事传播账号在日常阶段保持相对分散的活动,在特定舆情节点迅速转入集中发文状态。由此可见,其传播逻辑是围绕高度集中的叙事主题,在关键时点实施高强度投放,以在短时间内放大议题热度和叙事可见性。



注:单元格数值为相关系数;灰度表示绝对相关强度,非对角线灰度在 $|r| = 0.5$ 处截断。对角线统一显示为浅灰色;显著性标记:* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$。

图 3　虚假叙事操纵特征相关性分析

其次,虚假叙事传播账号在情感偏激和波动上呈现出“强情绪稳定输出与突发释放并存”的特征。一方面,政治情感异常偏差与发文突发参数之间存在显著正相关关系,表明部分账号在短时间内集中发布情绪偏离全网平均水平的内容,具有在关键节点强化情绪表达的倾向;另一方面,政治情感异常偏差与情感强度波动之间呈显著负相关关系,说明这些发布情绪偏激内容的账号,其历史发帖中的情绪波动反而较小,即长期维持相对稳定且强度较高的情绪表达状态。可见,虚假叙事传播账号通过持续维持某一情绪基调,并在特定时点进行集中强化,从而实现对受众认知与情绪的持续影响。

再次,虚假叙事传播账号在内容表达和聚焦上呈现出叙事框架协同传播特征。相关分析表明,政治议题聚焦度与政治议题内容复用度之间存在显著正相关关系。这意味着,当账号越集中于某一主题时,其内容与其他账号帖子之间越呈现较高相似性。该结果说明,虚假叙事传播账号不仅表现在个体层面关注单一议题,更存在跨账号之间的叙事复用、表达借用的现象。通过围绕相同议题持续输出相似内容,操纵主体可以在平台空间中制造多点共振的传播表象,增强特定叙事的存在感与真实感。此外,政治议题内容复用度与情感强度波动之间存在显著负相关关系。即越是依赖内容复用或模板化表达的账号,其历史发帖中的情绪波动越小。也就是说,模板化复制并不只是文本层面的重复,更往往伴随较为稳定的情绪输出框架。这表明虚假叙事传播账号的操纵在相当程度上体现为一种低成本复制、稳定输出与集中投放的运作逻辑。

4　讨　论

基于前文实证结果可以发现,虚假叙事传播账号的异常性并不只体现为账号属性或一般行为模式的偏离,更体现为其在具体政治事件中围绕特定叙事进行持续组织、集中投放和协同扩散的过程。换言之,该类账号的风险并非单纯来源于单条信息的真实性偏差,而是来源于其对议题解释框架、情绪氛围和传播节奏的持续性塑造,从而对舆论场中的认知结构与注意力分配进行系统化的塑造。

本文识别出的“政治议题聚焦—关键节点介入—政治情绪动员—叙事框架协同”特征,可被理解为一种结构化叙事操纵过程。政治议题聚焦体现为目标账号围绕特定议题反复发声,以强化某一问题框架;关键节点介入体现为其在舆情升温或关键节点集中介入,以放大叙事可见度;政治情绪动员体现为通过负向或偏激情绪表达提升传播感染力;叙事框架协同则体现为跨账号之间的表达复用和语义趋同,从而制造多源一致的传播表象。

上述发现对平台治理和风险预警具有一定启示。首先,在账号识别层面,应在传统账号属性、基础行为和网络结构指标之外,进一步纳入面向具体议题的叙事操纵特征。特别是在重大公共事件或政治敏感议题中,平台可重点关注账号的话题聚焦度、发文节奏异常、内容相似性和情绪偏离等指标,以提升对具有异常传播表征目标账号的识别能力。

其次,在识别流程设计上,可采用机器初筛后人工复核的这一效果和效率兼顾的机制。本文模型在正例占比较低的条件下仍保持较高召回率,说明其适合作为前置筛查工具嵌入实际治理流程。在具体应用中,可构建“机器初筛—人工复核”的分级识别机制,由模型在海量用户中快速筛选潜在风险账号,再由专业人员进行语境化研判与风险确认,从而在保证识别效率的同时提升判断准确性。

最后,在认知安全治理层面,应将虚假叙事传播视为一种系统性的叙事干预过程。这意味着治理策略也应从单纯的内容删除或事实纠偏,拓展至权威叙事对抗识别、政治情绪动员监测、叙事框架协同研判和关键节点介入预警。例如,可针对高度聚焦的议题叙事开展多元权威信息补充,削弱单一叙事框架的主导地位;同时,通过识别情绪异常偏离与集中爆发节点,提前介入可能引发舆论极化的传播阶段,从而降低情绪放大效应对公众认知的影响。

5 结论与展望

本文围绕政治安全语境下社交媒体平台虚假叙事传播账号识别问题,构建了“领域数据集构建—分层特征体系设计—不平衡学习模型识别—操纵规律分析”的研究框架,并基于微博平台两个政治安全相关事件开展实证检验。

研究结果表明:第一,面向政治安全议题构建领域专属数据集具有必要性。虚假叙事传播账号的判定依赖事件语境、权威事实参照与账号参与行为,通用社交机器人数据集难以充分反映其叙事传播特征。本文通过“机器初筛—人工精标”的两阶段标注机制,能够提升样本标签的任务适配性与研究针对性。第二,分层特征体系能够有效提升目标账号识别能力。在通用行为特征基础上,引入权威叙事对抗性、政治情绪动员性、叙事框架协同性与关键节点介入性等虚假叙事操纵特征后,模型在召回率(Recall)、F1 值等指标上均有所提升,说明叙事操纵特征能够弥补单纯依赖账号属性、基础行为和一般内容特征的不足。在目标账号占比较低的条件下,不平衡学习模型能够较好适应该类任务,其中 EasyEnsemble 在降低漏检风险与保持综合性能之间取得了较好平衡。第三,虚假叙事传播账号在政治事件中呈现出较明显的策略性操纵特征,主要表现为政治议题聚焦、关键节点介入、政治情绪动员与叙事框架协同。由此可见,政治安全语境下的虚假叙事传播是一种围绕特定议题持续展开的过程性传播干预,其治理重点也应由单条内容纠偏进一步拓展至账号识别与传播过程监测。

需要指出的是,本文仍存在一定局限。其一,研究数据来源于微博平台中的有限事件,样本规模和跨平台适用性仍需进一步拓展;其二,本文基于“机器初筛—人工精标”的标注路径,研究对象对更隐蔽或未进入初筛范围的传播主体覆盖仍有限。未来可在扩大数据规模、拓展跨平台样本和开展模型迁移检验等方面进一步深化研究,以提升虚假叙事传播账号识别方法的泛化能力与现实适用性。

参 考 文 献

[1] 叶进杰, 罗方禄, 史志勇. 数智赋能情报驱动的网络舆情安全态势感知塑造模型研究[J/OL]. 科技情报研究, 1-10[2026-04-27]. https://link.cnki.net/urlid/43.1554.G2.20260409.1402.002.

[2] 程桂龙, 谢 俊. 非传统安全视阈下网络意识形态安全治理[J]. 重庆社会科学, 2020, (4): 121-130.

[3] 修明圆. 网络“认知战”对抗下认知域风险的发生与防范[J]. 情报杂志, 2024, 43(11): 152-157, 77.

[4] 苏志远. 总体国家安全观视野下认知安全的理论阐释[J]. 情报杂志, 2024, 43(2): 41-46, 62.

[5] 汤景泰, 姚 春. 计算宣传与社交媒体平台中的舆论操纵[J]. 探索与争鸣, 2022, (11): 60-67, 178, 181.

[6] 陆 巍, 周睿平, 常一心. 基于社交媒体的俄乌冲突认知对抗: 原因, 特点与启示[J]. 情报杂志, 2023, 42(5): 30-35, 111.

[7] 金 燕, 朱家伟, 朱一梅, 等. 社交媒体中生成式 AI 的舆论操纵特征识别及应对策略研究[J]. 情报理论与实践, 2025, 48(9): 109-119.

[8] Ferrara E, Varol O, Davis C, et al. The rise of social bots[J]. Communications of the ACM, 2016, 59(7): 96-104.

[9] Javed D, Jhanjhi N, Khan N A, et al. Towards the future of bot detection: a comprehensive taxonomical review and challenges on Twitter/X[J]. Computer Networks, 2024, 254: 110808.

[10] Zhou J, Zhang D, Zhu J, et al. Speaking like humans, spreading like machines: a study on opinion manipulation by artificial-intelligence-generated content driving the internet water army on socialmedia[J]. Information, 2025, 16(10): 850.

[11] Guyan Q, Liu Y, Liu J, et al. PEGNN: peripheral-enhanced graph neural network for social bot detection[J]. Expert Systems with Applications, 2025, 278: 127294.

[12] 杜瑞成, 李 霞. 基于关系图保留 Transformer 与语义邻域融合的半监督社交机器人检测[J]. 数据分析与知识发现, 1-17[2026-04-20]. https://link.cnki.net/urlid/10.1478.g2.20260312.1336.002.

[13] Pan J, Liu Y, Liu X, et al. Discriminating bot accounts based solely on temporal features ofmicroblog behavior[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 450: 193-204.

[14] Alarifi A, Alsaleh M, Al-salman A. Twitter Turing test: identifying social machines[J]. Information Sciences, 2016, 372: 332-346.

[15] Bello B S, Heckel R. Analyzing the behaviour of Twitter bots in post Brexit politics[C]//2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2019: 61-66.

[16] 苏妍嫄, 董宵宇, 韩翠娟, 等. 嵌入双通道注意力卷积的联邦学习社交机器人识别模型[J]. 数据分析与知识发现, 2025, 9(11): 12-24.

[17] Liu F, Jiang B, Ma R. Robust social bot detection via multi-relational and noise-aware representationmodeling[J]. Information Processing & Management, 2026, 63(6): 104737.

[18] Zhang S, Chen H, Liu S, et al. UEF-CS: a multimodal framework for class-imbalanced social botdetection[J]. Information Sciences, 2026, 742: 123301.
[19] 许灵毓, 钟义信, 陈志成. 社交机器人对社会舆论的影响因素研究[J]. 智能系统学报, 2024, 19(1): 122-131.
[20] Stella M, Ferrara E, DeDomenico M. Bots increase exposure to negative and inflammatory content in online social systems[J]. Proceedings of the National Academy of Sciences, 2018, 115(49): 12435-12440.
[21] 张梦晗, 陈 泽. 信息迷雾视域下社交机器人对战时宣传的控制及影响[J]. 新闻与传播研究, 2023, 30(6): 86-105.
[22] 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(5): 61-80.
[23] Luo H, Meng X, Zhao Y, et al. Rise of social bots: the impact of social bots on public opinion dynamics in public health emergencies from an information ecology perspective[J]. Telematics and Informatics, 2023, 85: 102051.
[24] 程卿玄, 许 浩, 董 晶, 等. 人智交互中的社交机器人: 评论情感极性对用户参与行为的操纵机制研究[J]. 图书情报工作, 2026, 70(1): 56-68.
[25] 王建明, 郭 鹏. 网络舆情中共情传播的生成逻辑与引导策略[J]. 东岳论丛, 2025, 46(10): 147-155, 192.
[26] 王雪芬, 郭 雷, 常李艳, 等. 双系统理论视角下非理性网络评论形成机理[J]. 科技情报研究, 2025, 7(3): 36-46.
[27] Chen Q, Liu Q. Unveiling emotional contagion in COVID-19 misinformation: computational analysis for public health crisis-surveillance[J]. Health Informatics Journal, 2025, 31(4): 14604582251381175.
[28] 张诗莹, 柯 青. 从情绪场域到话题引流: 社交机器人传播虚假信息的双重机制[J]. 信息资源管理学报, 2026, 16(01): 50-62.
[29] 王晰巍, 赵 阳, 汤 明. 重大突发事件下舆情传播回音室效应中社交机器人交互特征研究[J]. 图书情报工作, 2026, 70(3): 73-87.
[30] Hajli N, Saeed U, Tajvidi M, et al. Social bots and the spread of disinformation in social media: the challenges of artificial intelligence[J]. British Journal of Management, 2022, 33(3): 1238-1253.
[31] Wu Y, Fang Y, Shang S, et al. A novel framework for detecting social bots with deep neural networks and activelearning[J]. Knowledge-Based Systems, 2021, 211: 106525.
[32] Gilani Z, Farahbakhsh R, Tyson G, et al. Of bots and humans (on Twitter)[C]. New York, NY: Association for Computing Machinery, 2017: 349-354.
[32] 张媛媛, 王鑫磊, 杨永清, 等. 社交网络圈群极化: 不同类型意见领袖的异质性分析[J]. 图书情报知识, 2025, 42(2): 145-155.