软件学报ISSN 1000-9825, CODEN RUXUEW E-mail: jos@iscas.ac.cn

*Journal of Software*, [doi: 10.13328/j.cnki.jos.000000] http://www.jos.org.cn

©中国科学院软件研究所版权所有. Tel: +86-10-62562563

基于机器学习的网络水军检测综述[[1]](#footnote-0)\*

XXX12

1(北京交通大学 计算机与信息技术学院, 北京 100044)

2(XXX北京市重点实验室, 北京 100044)

通讯作者: XXX, E-mail: XXX@bjtu.edu.cn

摘 要: 随着互联网的发展,用户评论对信息获取、商品购买等方面产生了前所未有的影响.例如,在社交网络中,用户评论可以快速呈现某个话题成为讨论的焦点,它可以促进电子商务中商品的销售,也可以影响书籍、电影或专辑的评分.在这些网络应用和服务中，“网络水军”是一种网络可疑行为,可以在网络空间中产生不正常的、破坏性的甚至是非法的行为,误导公众的认知,给网民和社会带来不良影响.因此，如何检测和打击虚假网络水军已成为一个亟待解决的问题,引起了信息技术和社会学研究者的关注.本文主要从信息技术的角度对其重新进行研究,总结了基于机器学习的网络水军检测的最新研究成果,分析了网络水军的特征,对基于机器学习的检测方法和评价标准进行了总结和分类.与以往的综述不同,文中还讨论了网络水军检测未来的发展方向,如:社交网络分析、多模态数据分析、数据隐私保护等.

关键词: 机器学习;社交网络;水军检测

中图法分类号: TP311

中文引用格式: XXX.基于机器学习的网络水军检测综述.软件学报,2021,32(7). http://www.jos.org.cn/1000-9825/0000.htm

英文引用格式: XXX. Survey on Network Astroturfing Detection based on Machine Learning. Ruan Jian Xue Bao/Journal of Software, 2023 (in Chinese).http://www.jos.org.cn/1000-9825/0000.htm

Survey on Network Astroturfing Detection based on Machine Learning

XXX12

1(Beijing Jiaotong University, Beijing 100044, China)

2(Beijing Key Laboratory of XXX, Beijing 100044, China)

**Abstract**: With the rapid development of the Internet, user reviews have gained unprecedented influence on information acquisition and product purchasing decisions. For instance, in social networks, user reviews can quickly generate discussions around a particular topic, drive e-commerce sales, and impact the ratings of books, movies, or albums. Within these online platforms and services, the presence of ‘Astroturfing’ poses a significant concern. Such spammers engage in suspicious activities that can range from generating fake content to engaging in harmful or even illegal behaviors in cyberspace. These actions can mislead the public and have adverse effects on internet users and society as a whole. Therefore, it has become imperative to address the issue of detecting and combating fake online spammers, attracting the attention of researchers in the fields of information technology and sociology. This paper focuses on investigating this issue from an information technology perspective. It provides a comprehensive overview of the latest research findings on detecting network spammers using machine learning techniques. The paper analyzes the characteristics of network spammers and categorizes various detection methods and evaluation criteria based on machine learning. In addition to reviewing existing approaches, this paper also explores future directions in online spammer detection, such as the application of social network analysis, analysis of multi-modal data, and ensuring data privacy protection. By addressing these challenges and exploring new avenues, we can further enhance our ability to identify and combat online spammers effectively.

**Key words**: machine learning; social network; astroturfing detection

# 引言

近年来,随着互联网用户数量的快速增长,社交媒体平台成为了人们交流、表达观点和获取信息的重要渠道.网络水军是指一群人或一组组织利用互联网和社交媒体平台,通过大量虚假账号和冒充真实用户的方式,以操纵舆论、制造谣言、刷粉丝、刷点击量、干扰讨论或破坏特定目标的形象为目的的行为.他们通过发布大量虚假信息、无意义的评论、刷屏、攻击其他用户等手段,试图影响舆论倾向、制造舆论压力或干扰正常的社交媒体活动.网络水军的存在对信息传播的公正性、社交媒体的信誉和用户体验带来了负面影响. 2015年10月,亚马逊起诉1114名网络水军成员,指控他们在亚马逊网站上对商品和服务提供虚假评论,违反了美国法律[1]. 2016年4月,美国一家科技社交网站宣布,Twitter上的网络水军成为美国总统选举期间的秘密武器[2]. 2022年,江苏省公安机关破获了一起“医药”系列自媒体账号敲诈勒索案.某网络水军团伙打着“医药观察”、“医药联盟”、“医药学术”等名义开办多个自媒体账号,杜撰医药企业负面文章并通过其自媒体账号进行炒作,以付费删帖相要挟,对10余个省市30余家医药企业实施敲诈勒索,涉案金额高达400余万元.由此可见,网络水军会在网络空间中引发不正常、破坏性甚至违法的行为,误导公众认知,给社会和网民带来负面影响.如何有效地检测和对抗网络水军已成为一个迫切需要解决的问题.

本文将从技术的角度而非社会学的角度研究网络水军,工作重点是针对基于机器学习的网络水军检测进展进行梳理、归纳、分析及讨论.本文的组织结构如下:第2节首先对网络水军的定义进行归纳,第3节介绍基于机器学习的网络水军检测方法,包括基于监督学习、基于半监督学习和基于无监督学习的检测方法,第4节介绍网络水军检测模型的评价指标,最后,在第五节中进行总结并展望了未来的发展方向.

# 网络水军定义

对基于机器学习的网络水军检测方法进行研究概述,首先要明确网络水军的定义并探明网络水军的特征.本小节主要从网络水军的定义、网络水军与其他网络可疑行为的对比以及网络水军的特征进行介绍.

## 网络水军的解释性定义

网络水军是一种利用互联网和社交媒体平台的非法行为,其特点是使用大量虚假账号和伪装成真实用户的方式,以达到操纵舆论、传播谣言、增加粉丝数量、提升点击量、干扰讨论或损害特定目标形象的目的.若以一种严格且准确的方式来定义网络水军是非常困难的,因此,本文对于网络水军提供几个解释性定义如下:

(1) 冒充真实用户:为了更加隐蔽和逼真,网络水军会冒充真实用户的身份,模仿他们的言论风格和行为习惯,以融入社交媒体平台的用户群体中[3].

(2) 大量虚假账号:网络水军通常通过创建大量虚假账号来增加其影响力,这些账号可能使用虚假的个人资料和头像,并且它们之间可能存在关联或互相支持.

(3) 相互支持和组织化:网络水军往往会互相支持和组织化,形成一个庞大的网络.他们可能会互相转发、点赞、评论,以提升自己和其他水军的可信度和影响力[4].

(4) 大规模发布内容:网络水军会发布大量内容,包括帖子、评论、推文等,以扩大其影响范围和传播力度.这些内容可能是重复的、低质量的或与特定议题有关的[5].

(5) 频繁活动和快速响应:网络水军会频繁活跃于社交媒体平台,并快速响应特定事件或话题,以便在关注度高时传播他们的信息.

需要注意的是,上述的网络水军特征并非一定都存在于每个网络水军的行为中,而且网络水军的行为也在不断演变和改变策略,以适应社交媒体平台的防控措施.因此,准确识别网络水军行为需要综合考虑多个方面的因素.

## 网络水军与其他网络可疑行为对比

在网络可疑的行为类别中,网络水军与其他已知的可疑行为(如垃圾邮件[6]、虚假评论[7]、社交网络垃圾信息[8-9]、恶意链接[10])相比既有相似之处,也有区别.我们在表1中对它们进行了比较.

表**1** 网络可疑行为对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 网络可疑行为 | 应用平台 | 攻击方式 | 任务周期 | 攻击规模 |
| 垃圾邮件 | 电子邮件、SMS | 机器攻击 | 无 | 巨大 |
| 虚假评论 | 电商平台 | 机器攻击、人工攻击 | 无 | 小 |
| 垃圾内容(社交网络) | 社交网络 | 机器攻击、人工攻击 | 无 | 小 |
| 恶意链接 | 搜索引擎 | 机器攻击 | 长时间 | 巨大 |
| 网络水军 | 电商平台、社交网络 | 机器攻击、人工攻击 | 短时间 | 巨大 |

如表1所示,我们选取了常见网络可疑行为的四个方面进行对比.在常见网络可疑行为中,网络水军主要出现在B2C(business-to-consumer)、C2C(customer-to-customer)的电商平台(如淘宝、亚马逊、京东)和社交网络平台(如抖音、tiktok、快手、微博、Facebook、Twitter)中.相比之下,垃圾邮件只出现在电子邮件和短消息业务(SMS)中,而恶意链接主要存在于搜索引擎当中.在攻击方式上,网络水军与大多数网络可疑行为一致,都是通过机器攻击或人工攻击的方式实现的,可以快速达到执行要求的目的.在任务周期方面,网络水军通过在几小时或几天内快速完成攻击任务,而恶意链接则需要更长的时间才能够实现.

# 基于机器学习的网络水军检测方法

为了更全面地概括基于机器学习的网络水军检测方法,本节将网络水军检测方法归纳为基于监督学习、半监督学习、无监督学习和其他检测方法四种.

## 基于监督学习的网络水军检测方法

基于监督学习的方法可以实现对网络水军的检测,其基本机制是将网络水军检测任务视为分类任务.Chen等[11]通过对一个网络论坛的调查发现,网络水军具有一些共同特征,如:网络水军言论发布的比例、网络水军发布者ID、首次发帖、回复评论、发布时间和推文活跃度等.通过研究得出了不同网络水军之间的关系,并构建了一个带有径向基函数的支持向量机分类器实现了对网络水军的检测. Lee等人[12]通过对网络水军的全面分析,利用用户画像特征、内容特征和社交网络特征训练了一个基于随机森林的分类器,实现了正常用户和网络水军的判别.Dong等人[13]基于自编码器和随机森林的特性,采用随机决策树模型指导全局参数学习过程,提出了一种可以训练的端到端的网络水军检测模型.在亚马逊评论数据集上广泛的实验表明,该模型能够有效的实现网络水军的检测.

## 基于半监督学习的网络水军检测方法

与基于监督学习的方法相比,基于半监督的检测方法通过将未标记数据与少量标记数据相结合,可以大大提高检测模型的准确性.Li等人[14]提出了一种两视图的半监督方法,利用大量可用的无标签评论来确保基于协作训练算法框架建立的网络水军评论检测.针对网络水军检测方法缺乏真实信息这一问题,Aghakhani等人[15]提出了FakeGAN系统,首次增强并采用生成对抗网络(GANs)进行文本分类任务,特别是检测网络水军评论.实验结果表明GANs可以有效地用于文本分类任务,具体来说,FakeGAN可以有效地检测网络水军评论.

## 基于无监督学习的网络水军检测方法

Lau等人[16]提出了一种新的文本挖掘模型并集成到语义语言模型中,用于网络水军发表的不真实评论. 根据从amazon.com上收集的真实数据集对模型进行评估,该模型在水军检测方面优于其他著名的基线模型.Liu等人[17]提出了一种检测方法DetectVC,该方法融合了用户关注行为图中的结构信息和从关注市场中收集的先验知识,在大规模微博数据集上的实验结果表明,DetectVC能够同时检测微博用户及其粉丝的注水行为. Zhang等人[18]首次提出了一种基于循环神经网络的识别网络水军虚假评论的检测方法(DRI-RCNN).该方法基于评论的虚假上下文和真实上下文特性以及词嵌入,利用循环神经网络向量来表示评论中的每个单词进而检测虚假评论.

综上所述,我们在表2中比较了不同文献中的使用的检测方法、提取的特征以及使用的基础模型.

表**2** 基于机器学习的网络水军检测模型对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 文献列表 | 监督学习 | 半监督学习 | 无监督学习 | 特征 | 基础模型 |
| 文献[11] | √ | — | — | 行为特征 | K-nearest |
| 文献[12] | √ | — | — | 行为、结构特征 | Random forest |
| 文献[13] | √ | — | — | 行为、结构特征 | Decision forest |
| 文献[19] | √ | — | — | 结构特征 | SVM |
| 文献[15] | — | √ | — | 内容特征 | GANs |
| 文献[20] | — | √ | — | 内容特征 | PU-learning |
| 文献[21] | — | √ | — | 行为特征 | C4.5 |
| 文献[16] | — | — | √ | 内容特征 | SLM |
| 文献[17] | — | — | √ | 结构特征 | DetectVC |
| 文献[22] | — | — | √ | 内容特征 | LDA |
| 文献[23] | — | — | √ | 内容特征 | MF |
| 文献[24] | — | — | √ | 结构特征 | Markov |
| 文献[25] | — | — | √ | 结构特征 | RCNN |

# 评价指标

有许多评价指标可用于评估网络水军检测模型的性能,包括:准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1值(F1 score)等,这些指标通常用于分类模型的评估.

## 精确率、召回率和F1值

准确率(Precision)、召回率(Recall)和F1值(F1 score)是常用的分类指标.精确率(Precision)是指所有预测为正的样本中实际为正样本的概率,召回率(Recall)是指在实际为正的样本中被预测为正样本的概率.F1值(F1 score)是精确率(Precision)和召回率(Recall)的加权平均值. 精确率(Precision)、召回率(Recall)和F1值(F1 score)公式如下:

(1)

(2)

(3)

其中TP为真阳性,即:该样本实际为阳性而被分类为阳性;FP表示假阳性,即:该样本实际为阴性而被分类为阳性;FN表示假阴性,即:该样本实际为阳性而被分类为阴性.在水军检测研究领域,有许多许多方法使用精确率、召回率或F1值作为评估指标来评估模型的性能,如:文献[13,16,18,19,22,26-30,32].

## 准确率和错误率

准确率(Accuracy)评估的是所有样本中被正确分类的样本的比例,而错误率(Error Rate, ER)则是衡量错误分类的样本占所有样本的比例,二者均是最常用的评估指标.准确率(Accuracy)和错误率(ER)公式如下:

(4)

(5)

文献[33-34]使用错误率(ER)作为实验评价指标,而文献[12-13,15,18,30,36-37]则使用准确率(Accuracy)作为评估他们构建的分类器的指标.

# 总结与展望

近年来,随着互联网的快速发展和社交媒体的普及,网络水军问题变得更加突出,对舆论和社会产生了负面影响.在网络水军检测研究领域,基于机器学习的方法是最常用和有效的方式之一,这些方法利用大数据和高级算法来识别网络水军的特征和模式,从而实现准确的检测和分类.本文针对基于机器学习的网络水军检测进行概述,首先介绍了网络水军的定义并于其他网络可疑行为进行了对比,然后分别介绍了基于监督学习、半监督学习和无监督学习的水军检测方法,进而通过梳理文献,对文献中评价检测模型的评估指标进行了总结.未来,水军检测领域将面临新的挑战和机遇.随着人工智能和机器学习的快速发展,我们可以期待更先进的水军检测算法和技术的出现.

(1) 强化学习和深度学习的应用:随着强化学习和深度学习在各个领域的成功应用,将其引入水军检测领域具有巨大的潜力.通过让算法自动学习和适应水军的新策略和行为模式,可以提高检测的准确性和鲁棒性.

(2) 多模态数据分析:随着社交媒体平台上内容的多样化,水军行为也趋向于使用图片、视频等多种形式的内容.未来的研究将探索如何结合多模态数据进行水军检测,实现对其行为的更全面的理解和识别.

(3) 社交网络分析:水军通常以网络结构组织,并通过社交网络平台进行传播和操纵.未来的研究将重点关注社交网络分析,以识别水军的网络结构和行为模式,从而提高检测的效果.

(4) 数据隐私保护:随着对用户数据隐私的重视,未来的水军检测研究将更加关注如何在保护用户隐私的前提下进行有效的检测.研究人员将探索采用差分隐私和加密技术等方法来保护用户数据的隐私性.

# 致 谢

感谢《计算机工程前沿(2023春)》课程的负责人XXX老师和本学期的全体授课教师.通过这门课,让我学习了许多关于计算机的前沿知识,包括但不限于:边缘计算、光纤传感器、射频指纹、嵌入式系统、移动自组织网络、片上系统、量子网络编码、数字孪生技术、计算机视觉中的阴影去除技术等.特别是在学习与本人研究方向相近的课程时(例如边缘计算在自动驾驶、智能交通场景的应用,自动驾驶场景下基于数字孪生摄像头的场景生成技术等),让我有了新的灵感和突破点.另外,还要特别感谢XXX老师和XX老师传授的论文写作经验,让我在梳理文献和构建知识框架等技能上有了提高.希望在未来的科研工作中,能够灵活运用本门课程所学的知识,更好的提升自己的科研能力.

References:

1. Choo E, Yu T, Chi M. Detecting opinion spammer groups through community discovery and sentiment analysis[C]//Data and Applications Security and Privacy XXIX: 29th Annual IFIP WG 11.3 Working Conference, DBSec 2015, Fairfax, VA, USA, July 13-15, 2015, Proceedings 29. Springer International Publishing, 2015: 170-187.
2. Heydari A, ali Tavakoli M, Salim N, et al. Detection of review spam: A survey[J]. Expert Systems with Applications, 2015, 42(7): 3634-3642.
3. Akoglu L, Chandy R, Faloutsos C. Opinion fraud detection in online reviews by network effects[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2013, 7(1): 2-11.
4. Carter G R. Toxic Sludge is Good For You! Lies, Damn Lies, and the Public Relations Industry[J]. Journalism & Mass Communication Educator, 1997, 52(1): 87.
5. McNutt J G. Researching advocacy groups: Internet sources for research about public interest groups and social movement organizations[J]. Journal of Policy Practice, 2010, 9(3-4): 308-312.
6. Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter[C]//Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). 2010, 6(2010): 12.
7. Jiang M, Beutel A, Cui P, et al. A general suspiciousness metric for dense blocks in multimodal data[C]//2015 IEEE International Conference on Data Mining. IEEE, 2015: 781-786.
8. Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots+ machine learning[C]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010: 435-442.
9. Hu X, Tang J, Gao H, et al. Social spammer detection with sentiment information[C]//2014 IEEE international conference on data mining. IEEE, 2014: 180-189.
10. Ghosh S, Viswanath B, Kooti F, et al. Understanding and combating link farming in the twitter social network[C]//Proceedings of the 21st international conference on World Wide Web. 2012: 61-70.
11. Song J, Lee S, Kim J. Crowdtarget: Target-based detection of crowdturfing in online social networks[C]//Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015: 793-804.
12. Lee K, Webb S, Ge H. Characterizing and automatically detecting crowdturfing in Fiverr and Twitter[J]. Social Network Analysis and Mining, 2015, 5: 1-16.
13. Dong M, Yao L, Wang X, et al. Opinion fraud detection via neural autoencoder decision forest[J]. Pattern Recognition Letters, 2020, 132: 21-29.
14. Li F H, Huang M, Yang Y, et al. Learning to identify review spam[C]//Twenty-second international joint conference on artificial intelligence. 2011.
15. Aghakhani H, Machiry A, Nilizadeh S, et al. Detecting deceptive reviews using generative adversarial networks[C]//2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 89-95.
16. Lau R Y K, Liao S Y, Kwok R C W, et al. Text mining and probabilistic language modeling for online review spam detection[J]. ACM Transactions on Management Information Systems (TMIS), 2012, 2(4): 1-30.
17. Liu Y, Liu Y, Zhang M, et al. Pay Me and I'll Follow You: Detection of Crowdturfing Following Activities in Microblog Environment[C]//IJCAI. 2016, 16: 3789-3796.
18. Zhang W, Du Y, Yoshida T, et al. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network[J]. Information Processing & Management, 2018, 54(4): 576-592.
19. Chen Y R, Chen H H. Opinion spam detection in web forum: a real case study[C]//Proceedings of the 24th International Conference on World Wide Web. 2015: 173-183.
20. Liu B, Dai Y, Li X, et al. Building text classifiers using positive and unlabeled examples[C]//Third IEEE international conference on data mining. IEEE, 2003: 179-186.
21. Xu A, Feng X, Tian Y. Revealing, characterizing, and detecting crowdsourcing spammers: A case study in community Q&A[C]//2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015: 2533-2541.
22. Dong L, Ji S, Zhang C, et al. An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews[J]. Expert Systems with Applications, 2018, 114: 210-223.
23. Ma H, Zhao W, Tan Q, et al. Orthogonal nonnegative matrix tri-factorization for semi-supervised document co-clustering[C]//Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part II 14. Springer Berlin Heidelberg, 2010: 189-200.
24. Fakhraei S, Foulds J, Shashanka M, et al. Collective spammer detection in evolving multi-relational social networks[C]//Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. 2015: 1769-1778.
25. Zhang W, Du Y, Yoshida T, et al. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network[J]. Information Processing & Management, 2018, 54(4): 576-592.
26. Hu X, Tang J, Zhang Y, et al. Social spammer detection in microblogging[C]//Twenty-third international joint conference on artificial intelligence. 2013.
27. Hu X, Tang J, Liu H. Leveraging knowledge across media for spammer detection in microblogging[C]//Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014: 547-556.
28. Hu X, Tang J, Liu H. Online social spammer detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2014, 28(1).
29. Sedhai S, Sun A. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research[C]//Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. 2015: 223-232.
30. Liu Y, Pang B. A unified framework for detecting author spamicity by modeling review deviation[J]. Expert Systems with Applications, 2018, 112: 148-155.
31. Liu Y, Pang B, Wang X. Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph[J]. Neurocomputing, 2019, 366: 276-283.
32. You L, Peng Q, Xiong Z, et al. Integrating aspect analysis and local outlier factor for intelligent review spam detection[J]. Future Generation Computer Systems, 2020, 102: 163-172.
33. Sun H, Morales A, Yan X. Synthetic review spamming and defense[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 1088-1096.
34. Wang G, Wang T, Zheng H, et al. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers[C]//23rd {USENIX} security symposium ({USENIX} security 14). 2014: 239-254.
35. Yang X, Yang Q, Wilson C. Penny for your thoughts: Searching for the 50 cent party on sina weibo[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2015, 9(1): 694-697.
36. Li L, Qin B, Ren W, et al. Document representation and feature combination for deceptive spam review detection[J]. Neurocomputing, 2017, 254: 33-41.
37. Dhingra K, Yadav S K. Spam analysis of big reviews dataset using Fuzzy Ranking Evaluation Algorithm and Hadoop[J]. International journal of machine learning and cybernetics, 2019, 10: 2143-2162.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | XXX(199X－),男,博士生,主要研究领域为机器学习模型的可靠性与鲁棒性、大数据场景下的数据安全与隐私保护. |  |  |  |

1. \* 基金项目: 国家自然科学基金(61906090, U20B2064)

   收稿时间: 2023-06-10; 修改时间: 2023-06-24; 采用时间: 2023-06-24 [↑](#footnote-ref-0)