综　述

doi: 10.3969/j.issn.1673-5692.2021.03.001

# 社交网络机器账号检测综述

李阳阳[1]，曹银浩[2]，杨英光[3]，金　昊[1]，杨阳朝[4]，石　珺[4]，李志鹏[4]

（1. 中国电子科技集团公司电子科学研究院社会安全风险感知与防控大数据应用国家工程实验室，北京　100041；
2. 北京航空航天大学网络空间安全学院，北京　100191；
3. 中国科学技术大学网络空间安全学院，合肥　230026；
4. 深圳市网联安瑞网络科技有限公司，深圳　518042）

**摘　要：** 随着移动互联网的大面积普及，社交网络用户数量在这些年也呈指数级增长，比如国外的推特和国内的微博等。与此同时社交网络中的机器账号也在大幅增长，这些机器账号不仅散布广告和低俗信息，甚至会模仿正常用户发言来操控舆论，挑拨对立，影响用户间正常的交流和社交网络氛围。因此机器账号检测应运而生，需要检测社交平台中的机器账号数量来避免正常用户被误导，并呈现出真实的舆论环境。文中介绍了这些年主流的机器账号检测方案：众包检测平台，基于机器学习的方案，基于深度学习的方案，基于社交关系图的方案和主动式检测方案等。并大体介绍了用于机器账号检测的各项算法技术，总结了各项技术的优缺点。最后本文总结了当前机器账号检测中存在的一些问题和难点，展望了相关研究的未来发展方向。

**关键词：** 社交机器人；社交网络；机器账号；机器学习；深度学习
**中图分类号：** TP391　　**文献标志码：** A　　**文章编号：** 1673-5692(2021)03-209-11

# A Survey of Social Bot Detection

LI Yang-yang[1], CAO Yin-hao[2], YANG Ying-guang[3], JIN Hao[1], YANG Yang-zhao[4], SHI Jun[4], LI Zhi-peng[4]

(1. National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC), China Academy of Electronics and Information Technology, Beijing 100041, China);
2. School of Cyberscience and Technology, Beihang University, Beijing 100191, China;
3. School of Cybersecurity, University of Science and Technology of China, Hefei 230026, China;
4. Shenzhen Cyber Aray Network Technology Co., Ltd., Shenzhen 518042, China)

**Abstract:** With the widespread popularity of mobile Internet, the number of social network users have also increased exponentially in recent years, such as Twitter and Weibo. At the same time, the number of social bots in social networks has also increased significantly. These social bots not only spread advertisements and vulgar information, but even imitate normal users to manipulate public opinion, provoke opposition, and affect normal communication between normal users and the atmosphere of social network. Therefore, machine account detection came into being. It is necessary to detect social bots in social platforms to avoid misleading normal users and present a real public opinion environment. This article introduces the mainstream social bot detection solutions over the years: crowdsourcing detection platforms, machine learning-based solutions, deep learning-based solutions and social graph-based solutions. It also introduces various algorithm technologies for social bot detection, and summarizes the advantages and disadvantages of each technology. Finally, this article summarizes some problems and difficulties in the cur-

**收稿日期：** 2021-02-01　　**修订日期：** 2021-03-05
**基金项目：** 国家自然科学基金项目(U20B2053)；海南省重大科技计划项目(ZDKJ2019008)

rent machine account detection, and looks forward to the future development direction of related research.

**Key words**: social bot; social network; sybil; machine learning; deep learning

## 0 引 言

近年来,随着互联网相关技术的飞速发展,尤其是移动互联网的普及,每个人都拥有自己的上网设备,使得人们加入社交网络几乎没有门槛和成本。过去几年社交网络用户数量也呈指数级增长,几乎每个人都有了自己的社交网络账号,全球数十亿人在使用各种社交网络[1]。随着人们对于社交网络的依赖日益加深,社交网络甚至成为很多人获取社会新闻的第一渠道,这也就使社交网络的重要性不断上升。但是在社交网络海量的用户和数据中,出现了很多机器账号,又被称作社交机器人。这些账号并非真人控制,大多由程序自动控制。有一部分机器账号为恶意账号,发布很多无用或者有害的信息,为幕后主使谋取利益。机器账号占比也非常惊人,据有关报道,国外社交网络巨头推特中,机器账号的推文占据了全部的大约 32%[2]。美国前总统奥巴马的关注者中,有将近 30% 的账号为机器账号,其当初的竞争者米特罗姆尼的关注者中,也有 20% 的用户可能是机器账号[3]。相比于 2012 年左右的数据,如今的机器账号占比可能更加疯狂。不过这是考虑到所有账户的情况,就活跃用户而言,2017 年有项研究[4]估计活跃用户中有 9% ~15% 为机器账号。

机器账号不仅在数量上呈现发展趋势,技术上也处在不断迭代的过程中。第一代机器账号大约 2011 年前出现在社交网络中,彼时这些账号只有很少的社交链接,自动化特征也较为明显。其后的第二代机器账号变得更加可信并开始流行,这些机器账号拥有大量的社交链接,也不再重复的发送相同内容,但是仍然可以通过特征工程,使用机器学习等方法精确的检测出来。随着技术的发展,2016 年以来已经发现了第三代机器账号,这些账号由于人为操作和自动化的混合程度加深,甚至从其他真实账号盗取信息,利用人工智能技术生成高可信的文本或图片,其行为更像真实人类账号,使得机器账号更加难以被检测识别。

虽然社交网络中存在着有益的机器账号,但一些恶意机器账号的出现无疑对我们的社交网络环境产生了影响,机器账号技术的进步无疑对网络安全产生了威胁,越来越多人试图利用社交网络机器账号达到不良目的[4],影响政治经济、引导对立等。这些账号在常见的微博、豆瓣等平台散布低俗赌博等广告信息,诱导网络用户点击广告或者钓鱼网站链接,以此牟利。在国外的社交网络中,许多机器账号还会用来影响政治活动[5-6]。在美国的选举活动之中,有许多机器账号在网络中发布大量的政治观点和看法,借此来影响舆论,并且影响正常用户即选民的看法。此外还有许多的机器账号被用来进行市场营销,发布相关产品的广告或软文,增加其曝光度,从而制造流行趋势。社交机器人账号在网络中的这些行为,影响了社交网络中信息的真实性。还有研究[7]表明,现在社交网络中许多用户会不小心泄露自己的隐私,比如姓名、年龄、住址、学校公司等信息。并且警惕性不高,给了不法分子可乘之机,他们会利用机器账号来进行社会工程学攻击,获取到正常用户的身份信息,进行诈骗或者其他操作,使用户的财产隐私安全受到威胁。

为了应对恶意社交网络机器账号对社会稳定、金融安全、个人隐私等领域的现实威胁,社交网络机器账号检测技术成为迫切需要发展的一项技术。发展针对模仿程度更高、机器行为更为隐匿的第三代机器账号的检测识别技术尤为必要。

## 1 研究现状

自从社交网络上机器账号泛滥以来,就有许多针对机器账号检测的研究,随着人工智能的发展,机器账号隐藏和检测的研究都在加速进行,自相关研究开展以来,相关方法可以分为以下几类[7]:众包社交机器账号检测平台、基于传统机器学习的检测技术、基于深度学习的检测技术、基于社交网络图的检测技术和主动式机器账号检测技术。

### 1.1 众包社交机器账号检测平台

文献[8]提出了众包社交机器账号检测平台,认为机器账号检测对于人类而言是一项较为简单的技术,因此创建了一个在线图灵检测平台,通过雇佣大量工作者和专家对脸书和人人网中的账号资料进行测试,向多个工作者提供相同的账号资料,将多数人的意见作为最终判定。

其具体流程如图 1 所示,首先在社交网络中将用户举报和异常行为的可疑用户进行筛选,筛选出可疑用户。同样地,对互联网中的众包工作者也先进行筛选,利用已确认的数据进行测试筛选,筛选掉一部分准确率极低的工作者,其余分为一般和高准确率的工作者。将可疑用户信息传给一般工作者进行判断,然后由高准确率工作者进行进一步的判断和确认,由两部分的判断结果共同决定可疑用户是否为机器账号。在使用过程中,检测平台的误报率接近于 0,可以保证非常高的检测正确率。

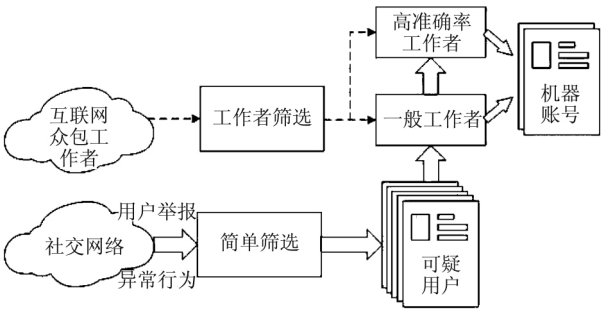



图 1　众包社交机器账号检测平台

然而,其缺点也非常明显。作者声称,如果在社交网络的早期就进行这种工作会有较好的效果,但其成本对于已具规模的社交网络平台而言几乎是不现实的。如今各主流社交平台用户数目在过去几年内都经历了爆发增长,例如 2019 年推特月活跃用户数已经达到 3. 36 亿,相比于 2012 年翻了 2. 5 倍[9],相比之下这种成本高昂、效率低下的服务就显得并不适用,每天海量的用户和数据使这样的方案只能停留在理论和实验过程中,而无法真正地投入实际应用。

## 1. 2　基于机器学习的检测技术

目前主流的检测技术是基于机器学习的机器账号检测技术,也是最为常见的。

基于机器学习的机器账号检测技术其实质是将这个问题看作一个二分类问题,在对账号提取出所需要的特征后,利用分类算法对数据进行分析,训练出检测模型,再利用模型对所需要分类的账号进行数据分析,并将其分类。其主要流程如图 2 所示。

### 1. 2. 1　数据获取

首先需要获取用户数据,在推特上可以直接利用其接口来获取用户的个人信息,包括用户名,粉丝数,注册时间,个人描述等,还可以获取用户公开的推文信息,训练模型所需要的特征就可以从这里提

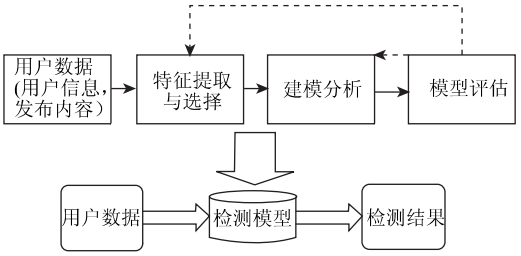



图 2　基于机器学习检测技术流程

取。问题在于训练数据的标注,目前大多数数据集是通过观察一部分相同目的的集群账号是否符合机器账号的标准。这些账号通常会共同发布相似的内容,达到目的。例如 cresci-2017[10] 数据集中的机器账号会为了竞选活动、商品营销、应用宣传等发布相似的内容带有共同的主题标签。而 pronbots[11] 数据集中的机器账号发布的内容中会带有诈骗网站的地址。与前面不同的是,caverlee[12] 数据集,其研究人员利用基于蜜罐的技术,进行了长达七个月的实验,他们用 60 个推特账号作为实验的蜜罐,实际上这些蜜罐不会参与正常的社交活动,只会互相发送@ 消息,同时只关注蜜罐账号。作者认为这些蜜罐账号不会吸引到正常用户来关注或者发送消息,因此将所有吸引到的账户都归为机器账号。而且即使有误分类,其错误率也和人共分类类似。经过去除重复和失效账号,七个月的实验吸引到了 22 223 个机器账号。

### 1. 2. 2　特征选择

不同的研究利用不同的用户特征来进行建模分析,以推特目前应用最广的 Botometer[13] (原名 Botornot) 检测平台为例,Botomter 将用户特征了分为了六大类[14],如表 1 所示。包括用户信息特征、网络特征、朋友特征( 这里面的“好友”并非关注者,而是其推文中转发提及以及被转发提及的用户) 、推文特征、情绪特征、时序特征等。表格中只是大概介绍了其主要特征,实际上 Botometer 在研究中总共提取了 1 000 多个具体特征,务求详尽。

Botometer[13] 是一个在线机器账号检测平台,在 2014 年推出,可以对提供的推特账号进行打分,分数越高,则这个账号是一个机器账号的概率也就越高。当用户提供一个账号昵称或者 ID 时,系统会获取这个账号的公开资料和数百条公开推文,以及这个账号的提及信息。会提取上述六个方面的总共 1 150 多个具体特征,然利用其已经训练好的检测模型对这个账号进行打分,平台也只是会提供分数,

表 1 Botometer 所采用特征类型

| 特征类型 | 详细信息 |
| --- | --- |
| 用户信息 | 用户个人信息: 昵称/ID、背景/头像、简介长度、粉丝数、关注数、推文/转推/回复/提及数、账户年龄等 |
| 网络特征 | 转发/提及所组成的网络特征: 节点/边的数量、网络密度、图的强度、聚类系数等 |
| 朋友特征 | 其互动好友的分布特征: 常用语言数量、账户年龄分布、关注/粉丝数量分布、推文数量分布、简介长度分布等 |
| 推文特征 | 推文的特征: 推文中词性( part-of-speech) 标签的频率和比例、单词计数和单词熵的分布特征等。词性标签: 动词、名词、形容词、情态助动词、前位限定词、感叹词、副词、代词和疑问词等 |
| 情绪特征 | 单个与整体推文所体现的情绪态度: 单个推文的正面/负面符号数、幸福指数、极化分数、唤醒分数、正负分值比等以及总体推文的幸福指数、幸福指数标准差等 |
| 时序特征 | 推文时序、连续推文/转发的时间差、连续提及的时间差等 |

并不会给出账号是否为机器账号的判断。这是第一个公开的推特机器账号检测接口,目的就是为了提高公众对于这些机器账号的认识。这个系统虽然提取的特征范围广,但是其每项特征都较为简单,并没有进行深入分析,例如推文中的情绪特征,只是由表面的单词特性和表情符号来确定,并没有进行深入挖掘。

还有很多机器学习方法利用了其他不同的特征来研究账号检测技术。文献 [3] 在研究中除了传统的用户特征之外,还将用户推文的情感特征作为一大部分加入了特征分析之中,比如推文的情感和账号的整体情感等。作者最终发现: 在情感方面机器账号的变化比人类要少得多,而且在表达情绪时,人类会倾向于表达更为强烈的情感,也更可能会与推特普遍观点相悖。这是情感特征在机器账号检测中的一个典型案例,表明情感特征的确是研究应当努力的一个方向。

文献 [15] 利用推特的 API 创建了一个社交机器账号,用于观察研究社交机器账号的行为特征,在观察阶段结束时,这个账号已经收获了 100 多位关注者,还获得了许多真实的互动信息,比如评论转发点赞等。通过对这个社交机器账号行为的观察,研究者决定选取七个特征来训练模型,包括@ 平均数、主题标签平均数、链接数、转发总数、原创推文总数、发推频率和发推的平台数等,以此来检测机器账号。文献 [16] 发现当前的大多数研究专注于模型的准确率,由于如今依旧是正常用户远多于机器账号,所以将机器账号识别为正常用户的代价并不高,作者就提出了 BoostOR 模型,使用的特征有: 转发所占比例、推文的平均长度、推文中 URL 链接比例、推文间的时间间隔等。引入了 Adaboost 的部分方法,最终目的是提高模型的召回率和 F1 值,能够识别出更多机器账号,此模型的 F1 值在两套数据集上都是最高的。文献 [17] 利用 N - grams 来对机器账号进行检测,利用用户的推文内容,对推文进行语义分析来判断推文作者是否为机器账号。

这些研究大都专注于用户的某一项特征,但是其研究较为深入,对这一方面的挖掘较为深入,也取得了不错的效果。

### 1.2.3 分类算法

机器学习算法可大致分为三类: 有监督、无监督和半监督学习[18]。监督学习[19]主要是通过带标签的数据样本训练得到最优模型,将模型的输出与标签做对比,如果效果不佳,则需要重新训练。模型训练完成后再通过这个模型对未知的样本数据进行预测分析。可用于机器账号检测监督学习算法有: 随机森林( Random Forest) [13]、贝叶斯算法( Bayes´ theorem) [20]、支持向量机算法( Support Vector Machine) [21]、逻辑回归算法( Logistic Regression) [21]等。而无监督学习与监督学习最大的不同就是: 无监督学习使用的训练数据是不带有标签的,也就是未经标注的数据,直接对数据建模,主要针对的是先验知识不足,人工标记较为困难的数据。常常被用于聚类问题,由于效果不易评估,很少被用于机器账号的检测。而第三种半监督学习则是综合了监督学习和半监督学习的特点,训练样本中一部分带有标签,另一部分不带。半监督学习相比于无监督学习可以提高模型准确性,减少人工标注的成本,并且可以利用无标签数据提高模型的泛化能力。主要可以应用于半监督分类和半监督聚类。

因为机器账号检测问题的目的明确,训练模型的效果容易评估,因此当前大部分用于机器账号检测的算法都是监督学习。不过利用无监督学习进行机器账号检测研究也在不断增多,文献 [22] 中机器账号检测技术就是使用的非监督学习,其研究人员认为普通的人类账号不可能长时间地保持高度同步,因此,高度同步的账号很可能是机器账号。他们

开发了一个相关性检测器 DeBot 来识别社交网络中的相关用户账号,首先收集账号的时间序列作为输入,将其进行聚类匹配,相似程度极高的账号可能为一批机器账号。DeBot 不需要带有标签的数据,而是将账号聚类成相关的集合,数据集中的效果要比前文的 Botomter 更好,而且这个过程也是接近实时的,每天可以以 94% 的准确率检测数千个机器账号,在 2016 年一年的时间里积累了 50 多万个独立的机器账号。使用非监督学习的好处是显而易见的,没有了固定数据集的束缚,模型可以使用大量的实时用户数据来训练模型,并且数据集的规模也可以更大。

而且在更常使用的监督算法之中,随机森林算法是运用最为广泛的。随机森林算法实质上是一个包含多个决策树的分类器,其输出的分类结果由决策树输出的分类结果决定。随机森林最早是从文献[23]提出的随机决策森林(Random decision forests)发展而来,之后由文献[24]提出随机森林的算法,并注册商标。随机森林是将集成算法中 bagging 算法与决策树学习相结合。决策树由于易于实现并且可解释性强,常用于各种机器学习的任务,但是决策树过深时容易发生过拟合,使其在训练样本中效果很好,但是可能不能很好的预测实际数据。而随机森林可以很好的解决这一问题。随机森林训练算法把 bagging 的技术应用到决策树学习中。给定训练集 $X = x_1, x_2, \cdots, x_n$ 和目标 $Y = y_1, y_2, \cdots, y_n$,bagging 方法重复多次从训练集中有放回地采样,然后在这些样本上训练树模型。重复的次数是自由参数,可以通过训练找到最优值。在这个 bagging 的通用方案之上,随机森林在学习的每次分裂过程中会选择随机的特征子集,这样可以降低决策树之间的相关性,多个相关性不高的决策树就可以降低分类器的过拟合性。并且随机森林实现简单,训练速度很快,应用范围很广泛。前面提到的 Botometer[13] 方案就是利用的随机森林算法,提取的特征用于训练七个不同的分类器,其十倍交叉验证的性能为 0.95AUC,体现了随机森林在这方面的卓越性能。文献[20]进行的实验验证了三个分类器中效果最好的也是随机森林算法。文献[25]提出了一套较为简单的模型,简化了特征工程,重点放在了提高机器账号检测方案的扩展性和通用性上,实现了可实时获取推特数据并进行检测的检测模型,其关键在于数据集的选择,实验发现,将所有的数据用来训练模型结果并不好,选择其中一部分才有最佳的效果,其中实验用到的算法也是随机森林算法。

除了随机森林算法,贝叶斯算法也是常用于机器账号检测的算法之一。朴素贝叶斯算法在 20 世纪 50 年代就已经开始了广泛研究,并且在 60 年代就已经引入到了文本信息检索之中[26],至今一直被广泛用于文本识别分类之中。文献[27]专门用朴素贝叶斯算法做机器账号检测,但是在对比多个分类器的实验中,朴素贝叶斯总是不能取得最好的成绩。文献[20]对比了随机森林,朴素贝叶斯和误差降低剪枝(REP)决策树三种算法,其中随机森林效果最好,朴素贝叶斯分类效果稍差。文献[21]用了四种分类算法进行实验,分别是: 逻辑回归、多项式朴素贝叶斯、SVM 支持向量机、梯度提升树,其中梯度提升树效果最好,SVM 效果次之,朴素贝叶斯效果排名第三。可看出朴素贝叶斯分类器效果不是最好,也不会太差。而由决策树发展而来的分类算法效果一直不错。

### 1.3　基于深度学习的检测技术

随着深度学习的火热发展,最近已经有越来越多的研究将其运用到机器账号检测过程中。深度学习算是机器学习的一个分支,深度学习以人工神经网络为基础架构,对数据进行表征学习[28]。与传统的机器学习不同的是,深度学习对数据的要求更多,需要更多的数据和时间来训练模型,同时深度学习可以利用无监督或者半监督的特征学习以及用分层的特征提取算法来代替人工获取特征[29],可以大大节省时间并发现一些隐藏特征。

长短期记忆(long short-term memory,LSTM)是一种时间循环神经网络,最早发表于 1997 年[30],是专门设计出来解决一般的循环神经网络[31](recurrent neural network,RNN)存在的长期依赖问题。适用于处理和预测时间序列里间隔和延迟较长的事件,如今经常作为大型深度神经网络的一部分参与构造。机器账号检测的研究者也将 LSTM 用到了相关实验和项目之中[32-33]。

文献[32]将卷积神经网络(convolutional neural network,CNN)和 LSTM 网络用到机器账号检测之中,其模型如图 3 所示。CNN 网络用于提取推特文本内容的特征及其关系,第二层将推特元数据视为时间信息,并使用该时间信息作为 LSTM 的输入来提取用户社交活动时间特征。图中的 Rtu 即为用户的推文在几天内的被转发信息,MENu,COMu,URLu,HTu 分别为几天内用户被提及、用户的推文

中被评论、推文中链接和推文中主题标签的信息特征。而 Tu 则为 CNN 网络提取的文本内容特征，最后在融合特征层，将前面的内容特征和元数据特征融合来检测机器账号，最终得出检测结果。

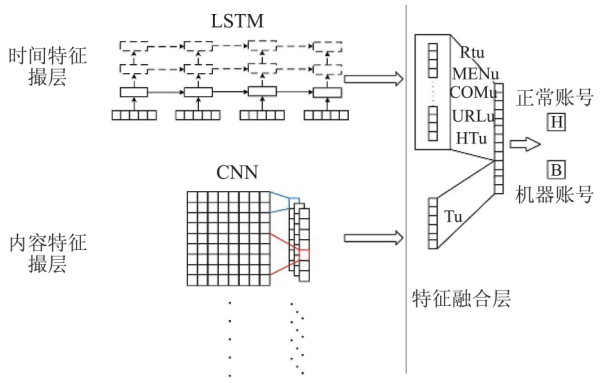



图 3 基于 CNN 和 LSTM 的 DeBD 结构

文献 [3] 利用推特内容和元数据得到的模型可以在推文级别进行机器账号的检测，从用户元数据中提取上下文特征，并将其作为辅助输入提供给处理推文文本的 LSTM 网络，其模型仅需要一条推文就可以来判别是否为机器账号。类似的还有文献 [4] 中的模型，如图 4 所示，其作者使用 BiLSTM 算法来进行机器账号的检测，BiLSTM 是一种使用双向 LSTM 的算法，两个 LSTM 方向相反，图中的 $LSTM_L$ 代表前向 LSTM，$LSTM_R$ 代表后向 LSTM，共同组成 BiLSTM 网络。其模型使用了推文的上下文作为输入，经过词嵌入后进入 BiLSTM 网络，最后前向 LSTM 和后向 LSTM 的输出进行拼接，再经过归一化函数之后进行分类，得到我们需要的检测结果。此模型仅使用推文内容作为输入，没有使用其他的特征，这种方法的好处就在于节省了大量的特征提取的工作时间，不需要手工的特征和先验知识，可以提高工作效率，也更便于部署到批量检测的场景中。

文献 [5] 将异构图神经网络应用到了恶意账户的检测之中，其核心是账户之间总会产生“聚合”，分辨一个账户是正常账户还是恶意账户的关键就是同一个拓扑中的其他账户如何如这个账户“聚合”，可分为“设备聚合”和“活动聚合”，主要应用场景是国内的支付平台支付宝，但是其思想也可以应用到社交网络平台中。文献 [6] 提出了一个两阶段的，基于图的机器账号检测系统，该系统利用了监督学习和无监督学习，使用 SOM，第 1 阶段在最大化良性集群与疏远恶意机器账号之间建立了折衷方案，使最后的结果避免了高 FP 和 FN 值。作者

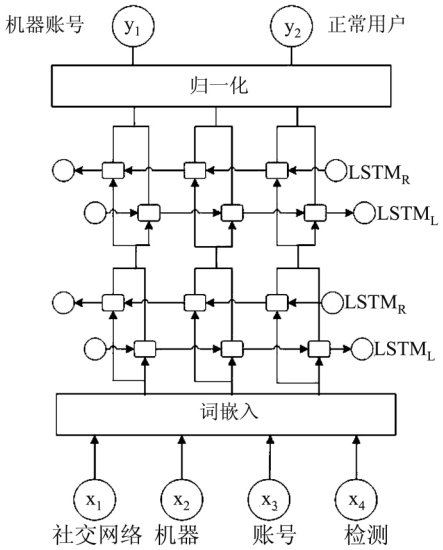



图 4 基于 BiLSTM 的检测模型

提出的另一个模型 [7] 利用 HAT 进行增量学习实时处理数据。虽然该模型收敛所需的时间更长，但在最终模型中却表现出出众的分类性能，适用于基于流的检测系统。

同样地，基于深度学习的检测技术也有其缺点，当数据集不够大的时候，神经网络的效果往往不好而且容易产生过拟合现象。

表 2 总结了六种主流机器账号检测模型，包括其所用特征，基础算法和其主要的特点。

## 1.4 基于社交关系图的检测技术

基于社交关系图的检测技术的主要依据是社交网络中用户之间所形成的社交网络图，社交网络图可以用于理解和分析社交网络平台上用户之间的关系。因此基于社交关系图的检测技术重点关注于用户之间的关系，毕竟在社交网络中，不会有账号孤立存在，彼此之间都是有联系的，正常用户和机器账号的社交关系图往往有很大区别。比如正常用户的好友中会有很大一部分来自于现实中的好友，彼此相互关注，互动很多。而机器账号则不会有这样的特征，机器账号的互关好友就会少很多，这在社交关系图上会很明显，其评论和点赞也比较少，大部分是发送推文或者转发来扩大影响力。而且正常用户和机器账号的好友中，机器账号的所占比例也会不同。因此正常用户的社交关系图的结构与机器账号的图结构会有显著区别，基于社交关系图的检测方案正是利用这种区别，加上用户的网络特征来进行识别和检测。

表 2　主流检测模型原理和特点

| 模型 | 使用特征 | 基础算法 | 特点 |
|---|---|---|---|
| Botometer [13] | 综合特征 | 随机森林 | 特征多样,模型成熟,算法简单 |
| Debot [22] | 账号间相关性 | pairwise | 检测速度快,实时,适应性高 |
| Sentibot [3] | 情感特征 | Adaboost/gradient boosting | 引入情感特征, |
| BoostOR [16] | 推文特征 | Adaboost | 提高 F1 值,减少机器账号被错误分类 |
| 模型 [33] | 用户/推文特征 | LSTM | 只用一条推文即可检测,方便快捷 |
| MLPGD [1] | 整体推文特征 | 神经网络 | 典型的神经网络,准确度高 |

SybilRank [38] 代表了该框架的一个示例: 对方可能控制多个社交机器账号( 在这种情况下通常称为 sybils) 冒充不同的身份并发起攻击或渗透。提议的检测 sybil 账号的策略通常依赖于检查社交图的结构。例如,SybilRank 假定 sybil 账号只显示少量指向合法用户的链接,而不是主要连接到其他 sybil,因为它们需要大量的社交关系才能显示出可信赖的状态。利用此特征来识别密集的相互联系的社交机器账号。

文献 [39] 设计了基于随机游走的检测模型 SybilWalk,在无向社交图上进行随机游走。简易的网络示例如图 5 所示,在社交关系图之外创建两个节点 $l_b$ 和 $l_s$ 代表绝对的正常节点和机器节点,将社交关系图中标签为正常节点的与 $l_b$ 相连,标签为机器节点的与 $l_s$ 相连,$l_b$ 标记分数为 0,$l_s$ 标记分数为 1,每个节点的得分为其随机游走到 $l_b$ 之前到达 $l_s$ 节点的概率,将节点的初始分数设置为 0.5,可知每个节点得分与其邻居节点得分相关,经过足够多轮迭代之后的得分作为最终分数。作者认为在社交关系图中,正常用户内部和机器账号内部的连接都较为紧密,而机器账号与正常账号之间的连接较少,因此正常账号随机游走到 $l_s$ 的概率较小,而机器账号随机游走到 $l_s$ 的概率更大,所以一个节点的得分较高时,说明这个账号是机器账号的概率也较高,所以可以将这个分数作为此节点是机器账号的概率。

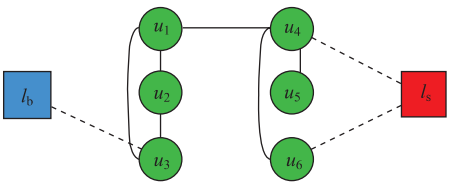
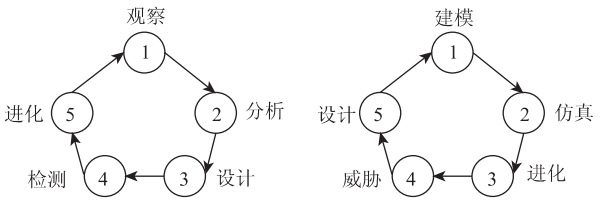



图 5　SybilWalk 标签增强型社交网络示例

此项研究解决了先前的随机游走模型的对于噪声敏感和在弱同构网络图中准确率不高的问题,而且其可扩展性也很高。此外还有 GANG [40]、SybilSCAR [41]、SybilFuse [42] 等研究都是基于社交关系图所做的机器账号检测方法。

## 1.5　主动式机器账号检测技术

如图 6( a) 所示,文献 [43] 指出现有的检测方案都是被动式检测方案,其检测流程是: 先观察到机器账号的存在、收集相关数据集进行分析、针对分析的结果设计检测方案、使用检测方案进行检测、机器账号继续进化,然后进入下一轮检测的拉锯战中。为了避免检测方案在机器账号进化时失效,提出了一种能够提前发现检测模型弱点,从而及时改进的主动式检测方案。其主要流程如图 6( b) 所示,先对机器账号进行行为建模、仿真模拟机器账号行为、进化产生新的机器账号、评估进化后的机器账号是否存在其他检测维度、设计检测方案。




(a) 被动式检测方案　　(b) 主动式检测方案

图 6　两种检测方案

文献 [44] 给出了一种主动式机器账号检测方案的实现。该方案对社交网络中账号的动作按照时间线进行提取并建模,将账号的不同动作如: 发推、回复、转推等按照时间先后顺序建模成序列。因为真实账号通常在行为模式上表现出高度的不一致性,而同一组受控机器账号,却会表现出高度的同质性,从而可以用字符串分析的方式将机器账号和真实账号进行区分。作者使用了遗传算法对以行动序列为表征的机器账号模拟进化,结果证实经过 2 000 多轮迭代,进化后的机器账号逃脱了字符串分析方式的检测。这促使作者继续评估进化后的机器账号

与真实账号间是否存在同种建模方式下的其他检测维度。最终,提出了一种基于香农信息熵测度账号行动序列混乱程度的方法,改进了对演化后机器账号的识别。

## 2 实验数据集

现有的常用数据集整理如表 3。其中账号数目与原本数目有所差异是将数据集中无效的账号去除掉所造成的。Stefano Cresci 团队和文献 [25] 的研究人员都收集了很多数据集,对社交网络机器账号的研究有很大的帮助。

表 3 常用数据集简要总结

| 数据集 | 年份 | 机器账号 | 正常用户 |
|---|---|---|---|
| Caverlee [9] | 2011 | 15 483 | 14 833 |
| Gilani [45] | 2017 | 1 039 | 1 365 |
| Carol [4] | 2017 | 826 | 1 747 |
| Cresci2017 [10] [46] | 2017 | 7 049 | 2 764 |
| Midterm-2018 [25] | 2018 | 41 395 | 7 790 |
| Cresci-stock [47-48] | 2018 | 6 907 | 5 992 |
| Vendor-purchased [49] | 2019 | 1 024 | 0 |
| Verified [25] | 2019 | 0 | 1 964 |
| Pronbots [49] | 2019 | 17 880 | 0 |
| Political-bots [49] | 2019 | 62 | 0 |
| Celebrity [49] | 2019 | 0 | 2 113 |
| Cresci-rtbust [50] | 2019 | 332 | 322 |
| Botwiki [25] | 2019 | 697 | 0 |
| Botometer-feedback [49] | 2019 | 139 | 375 |

## 3 总结与展望

(1) 加强对机器账号情感特征的分析

毫无疑问普通用户与机器账号之间最大的不同在于推文中所暗含的情绪因素。前文中提到的诸多方案中,只有很少的研究将情感特征纳入实验分析之中,大部分的机器学习检测方案重点依旧是账号的属性,比如:关注者数目、推文数量、注册时间地点等。没能将最大的不同—情感特征纳入研究。一小部分关注到情感特征的研究多是根据在推文中的表情符号来进行分析,没能分析推文内容包含的情感。如今机器账号的发展很快,有很多已经可以在发布推文时加入表情符号来伪装成正常用户,这对于机器账号的识别更增加了难度。因此,未来的研究需要分析推文中包含的情感因素来更好地检测机器账号。

(2) 提升检测模型的通用性和泛化能力

如今的机器账号类别很多,根据其不同的目的和行为方式可以分为很多种 [51]。恶意账号:为了盈利而发布大量的恶意链接,诱使人们点击,从而造成人们的财产损失或者隐私信息泄露。这也是当前的研究最为关注的。水军账号:主要是为了营销活动或者政治活动造势,因为其目的隐晦,一般会伪装得与正常用户很相似,并且本来就存在很多人类操控的水军账号,因此很难识别,对正常用户的误导会很严重。僵尸账号:这些账号通常是灰色产业链的一部分,比如明星账号的僵尸粉。还有些账号会进行大量的重复操作从而达到推广的目的,比如经常会有账号带着涨粉广告的头像进行批量的关注转发操作。除了这几种之外还有一些机器账号,对网络环境不会造成负面影响,比如单纯播报天气的机器账号,并没有恶意。

但是如今的大多数检测方案结构较为单一,大部分方案只能对某一类机器账号进行识别,比如恶意账号或者僵尸账号。而无法对其他类别的机器账号进行很好的检测识别。而且基于机器学习的方案对数据集依赖严重,新产生的机器账号可能不符合其模型范例而无法被识别。这些模型重新训练需要花费的时间也较长,也就是旧时的检测方案不能很好的识别新产生的机器账号,无法随时间进化应变。

(3) 增加机器账号群体行为模式的考量

很多研究只针对于单一机器账号进行检测。然而当前的社交网络中的单一机器账号所展现出的特征和行为模式越发与人类账号相似。所以仅仅针对单一机器账号进行检测进行的研究,其应用前景越来越小。但由于大量的机器账户本身由一小部分实体或账户进行控制。所以受控于相同实体或账户的大量机器账号,总会在行为模式或其他特征中存在相似性。以一组机器账号的为目标,对一组机器账号在某个行为特征的维度上进行建模,以行为特征的相似性为切入点,从而将整组机器账号识别并将其与人类账号进行区分将会是未来一个可以研究的方向。

(4) 发展主动式对抗性机器账号检测方案

技术发展日新月异,很多技术不仅可以用来检测机器账号,也被拿去发展更新的机器账号。如今已经出现了一些半社交机器人 [52],也就是其行动不完全由程序自动进行,而是由人类激活程序之后交

给机器账号自动进行,这使得其推文时间更加多变,不确定性增加。如今的机器账号也越来越智能,对正常用户的模仿更加深入,检测也愈发困难。当前的检测技术的研究和开发基本都是使用这样一套流程: 1) 从社交网络上发现了一种新型机器账号。2) 从社交网络上收集新型账号产生的相关数据,建立数据集。3) 对数据集进行分析、建模、开发出一种检测技术。4) 使用检测技术发现更多的同类型机器账号。对于这样一套流程,我们可以发现,这种方式产生的检测技术始终是后知后觉的,并且落后于机器账号的发展。

因此未来的机器账号检测不仅需要更加智能,多挖掘出表面之下的账号特征,如推文情感分析等。检测方案也需要更加综合,例如可以将机器学习与社交关系图结合起来,共同分析账号的特征和社交网络图,并且在部分关节引入人工判断机制,毕竟人类本身更能识别出机器账号的不同。为了进一步提升检测技术的鲁棒性和检测能力,甚至需要更深一步对机器账号的下一步可能的更新方向进行分析,从分析的结果得到可以用来检测新型机器账号的特征维度[43]。以对抗性的思路来产生更加强大、泛化性更高、甚至有预防能力的检测技术[44]。

同时相关方案也需要得到社交网络平台的支持,由平台本身进行检测识别,无疑是最为方便快捷的,同时平台也可以更好地监督新加入的账号,监督新注册用户的异常流量,检测到机器账号则进行公示清号或者贴上机器账号的标签,可以提高识别效率,这对于平台本身而言也是在维护自身氛围和信誉。也可以降低用户的使用成本。

近几年图神经网络发展迅速,在图像处理、用户推荐、生物化学等各个领域都有着不错的应用。也有一些研究人员将其应用到了机器账号检测之中[35],由于社交网络中,用户之间的关系就是图,未来的研究也需要在这方面进一步发展,尤其是异构图神经网络,还有很大的发展潜力。

## 4　结　语

机器账号和对应的检测技术都在不断发展,就如同军备竞赛一般,双方势不两立,却也在共同进步,促进着技术的进步。我们可能没办法彻底清除机器账号,但是需要不断地努力让机器账号的负面影响降到最小,并且发挥其正面作用,这也是诸多研究的意义所在。

## 参考文献:


[1] ALARIFI A, ALSALEH M, AL-SALMAN A M. Twitter turing test: Identifying social machines [J]. Information Sciences, 2016, 372: 332-346.

[2] ABOKHODAIR N, YOO D, McDonald D W. Dissecting a social botnet: Growth, content and influence in Twitter [C] //Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. Hong Kong: Association for Computing Machinery, 2015: 839-851.

[3] DICKERSON J P, KAGAN V, SUBRAHMANIAN V S. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? [C] //2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). Beijing: IEEE Press, 2014: 620-627.

[4] VAROL O, FERRARA E, DAVIS C A, et al. Online human-bot interactions: Detection, estimation, and characterization [C] //Eleventh international AAAI conference on web and social media. Montreal: AAAI, 2017: 1-10.

[5] FREITAS C, BENEVENUTO F, GHOSH S, et al. Reverse engineering socialbot infiltration strategies in twitter [C] //2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Paris: IEEE Press, 2015: 25-32.

[6] SHAFAHI M, KEMPERS L, AFSARMANESH H. Phishing through social bots on Twitter [C] //2016 IEEE International Conference on Big Data (Big Data). Washington: IEEE Press, 2016: 3703-3712.

[7] FERRARA E, VAROL O, DAVIS C, et al. The rise of social bots [J]. Communications of the ACM, 2016, 59 (7): 96-104.

[8] WANG G, MOHANLAL M, WILSON C, et al. Social turing tests: Crowdsourcing sybil detection [J]. arXiv preprint arXiv: 1205.3856, 2012.

[9] Twitter Inc. Twitter Q1-2019-Shareholder-Letter. [EB/OL]. [2020-01-05]. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Shareholder-Letter.pdf.

[10] CRESCI S, DI PIETRO R, PETROCCHI M, et al. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race [C] //Proceedings of the 26th international conference on world wide web companion. Perth: International World Wide Web Conferences Steering Committee, 2017: 963-972.

[11] YANG K C, HUI P M, MENCZER F. Bot electioneering volume: Visualizing social bot activity during elections [C] //Companion Proceedings of The 2019 World Wide

Web Conference. San Francisco: ACM, 2019: 214-217.

[12] LEE K, EOFF B D, CAVERLEE J. Seven months with the devils: A long-term study of content polluters on twitter [C] //Fifth international AAAI Conference on Weblogs and Social Media. Barcelona: 2011: 1-9.

[13] DAVIS C A, VAROL O, FERRARA E, et al. BotOrNot: A System to Evaluate Social Bots [C] //Proceedings of the 25th International Conference Companion on World Wide Web. Montreal: International World Wide Web Conferences Steering Committee, 2016: 273-274.

[14] VAROL O, DAVIS C A, MENCZER F, et al. Feature engineering for social bot detection [J]. Feature engineering for machine learning and data analytics, 2018: 311

[15] DEWANGAN M, KAUSHAL R. SocialBot: Behavioral Analysis and Detection [C] //2016 International Symposium on Security in Computing and Communication. Jaipur: [s. n.], 2016: 450-460.

[16] MORSTATTER F, WU L, NAZER T, et al. A new approach to bot detection: Striking the balance between precision and recall [J]. San Francisco: IEEE Press, 2016: 533-540.

[17] PIZARRO J. Using N-grams to detect Bots on Twitter [C] //Use permitted under Creative Commons License Attribution 4.0 International( CC BY 4.0) CLEF.. Thessalonik: [s. n.], 2019: 1-10.

[18] ALPAYDIN E. Introduction to Machine Learning [M]. Cambridge: MIT press, 2020.

[19] KOTSIANTIS S B, ZAHARAKIS I, PINTELAS P. Supervised machine learning: A review of classification techniques [J]. Emerging Artificial Intelligence Applications in Computer Engineering, 2007, 160( 1): 3-24.

[20] FAZIL M, ABULAISH M. Identifying active, reactive, and inactive targets of socialbots in Twitter [C] //Proceedings of the International Conference on Web Intelligence. Leipzig: Association for Computing Machinery, 2017: 573-580.

[21] KANTEPE M, GANIZ M C. Preprocessing framework for Twitter bot detection [C] //2017 International Conference on Computer Science and Engineering ( UBMK). Antalya: IEEE Press, 2017: 630-634.

[22] CHAVOSHI N, HAMOONI H, MUEEN A. Debot: Twitter bot detection via warped correlation [C] //2016 IEEE 16th International Conference on Data Mining. Barcelona: IEEE Press, 2016: 817-822.

[23] HO T K. Random decision forests [C] //Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal: IEEE Press, 1995: 278-282.

[24] BREIMAN L. Random forests [J]. Machine learning, 2001, 45( 1): 5-32.

[25] YANG K C, VAROL O, HUI P M, et al. Scalable and generalizable social bot detection through data selection [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34( 1): 1096-1103.

[26] RUSSELL S, NORVIG P. Artificial intelligence: A modern approach [M]. Englewood: Prentice-Hall, Inc., 2002.

[27] GAMALLO P, ALMATARNEH S. Naive-Bayesian Classification for Bot Detection in Twitter [C] //Use permitted under Creative Commons License Attribution 4.0 International ( CC BY 4.0) CLEF 2019. Lugano: [s. n.], 2019: 1-9.

[28] DENG L, YU D. Deep learning: Methods and applications [J]. Foundations and Trends in Signal Processing, 2014, 7( 3 - 4): 197-387.

[29] SONG H A, LEE S Y. Hierarchical representation using NMF [C] //International Conference on Neural Information processing. Berlin: Springer, 2013: 466-473.

[30] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9 ( 8): 1735-1780.

[31] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323( 6088): 533-536.

[32] PING H, QIN S. A social bots detection model based on deep learning algorithm [C] //2018 IEEE 18th International Conference on Communication Technology( ICCT). Chongqing: IEEE Press, 2018: 1435-1439..

[33] KUDUGUNTA S, FERRARA E. Deep neural networks for bot detection [J]. Information Sciences, 2018, 467: 312-322.

[34] WEI F, NGUYEN U T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings [C] //2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications ( TPS-ISA). Los Angeles: IEEE Press, 2019: 101-109.

[35] LIU Z, CHEN C, YANG X, et al. Heterogeneous graph neural networks for malicious account detection [C] // Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Torino: Association for Computing Machinery, 2018: 2077-2085.

[36] Abou Daya A, Salahuddin M A, Limam N, et al. A graph-based machine learning approach for bot detection [C] //2019 IFIP/IEEE Symposium on Integrated Network and Service Management ( IM). Arlington: IEEE Press, 2019: 144-152.

[37] ABBAS A D, SALAHUDDIN M A, LIMAM N, et al.

Botchase: Graph-based bot detection using machine learning [J]. IEEE Transactions on Network and Service Management, 2020, 17(1): 15-29.

[38] CAO Q, SIRIVIANOS M, YANG X, et al. Aiding the detection of fake accounts in large scale social online services [C]//Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. San Jose: USENIX Association, 2012: 197-210.

[39] JIA J, WANG B, GONG N Z. Random walk based fake account detection in online social networks [C]//2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Denver: IEEE Press, 2017: 273-284.

[40] WANG B, GONG N Z, FU H. GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs [C]//2017 IEEE International Conference on Data Mining (ICDM). New Orleans: IEEE Press, 2017: 465-474.

[41] WANG B, JIA J, ZHANG L, et al. Structure-based sybil detection in social networks via local rule-based propagation [J]. IEEE Transactions on Network Science and Engineering, 2018, 6(3): 523-537.

[42] GAO P, WANG B, GONG N Z, et al. Sybilfuse: Combining local attributes with global structure to perform robust sybil detection [C]//2018 IEEE Conference on Communications and Network Security (CNS). Beijing: IEEE Press, 2018: 1-9.

[43] CRESCI, S, PETROCCHI, M, SPOGNARDI A, et al. From reaction to proaction: Unexplored ways to the detection of evolving spambots [C]// Companion Proceedings of the the Web Conference. Lyon: International World Wide Web Conferences Steering Committee, 2018: 1469-1470.

[44] CRESCI, S, PETROCCHI, M, SPOGNARDI A, et al.. Better safe than sorry: An adversarial approach to improve social bot detection [C]// Proceedings of the 10th ACM Conference on Web Science. Amsterdam: Association for Computing Machinery, 2018: 47-56.

[45] GILANI Z, FARAHBAKHSH R, TYSON G, et al. Of bots and humans (on twitter) [C]//Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris: Association for Computing Machinery, 2017: 349-354.

[46] CRESCI S, DI PIETRO R, PETROCCHI M, et al. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling [J]. IEEE Transactions on Dependable and Secure Computing, 2017, 15(4): 561-576.

[47] CRESCI S, LILLO F, REGOLI D, et al. $ FAKE: Evidence of spam and bot activity in stock microblogs on Twitter [C]//The 12th International AAAI Conference on Web and Social Media. Stanford: ICWSM, 2018: 580-583.

[48] CRESCI S, LILLO F, REGOLI D, et al. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter [J]. ACM Transactions on the Web (TWEB), 2019, 13(2): 1-27.

[49] YANG K C, VAROL O, DAVIS C A, et al. Arming the public with artificial intelligence to counter social bots [J]. Human Behavior and Emerging Technologies, 2019, 1(1): 48-61.

[50] MAZZA M, CRESCI S, AVVENUTI M, et al. Rtbust: Exploiting temporal patterns for botnet detection on twitter [C]//Proceedings of the 10th ACM Conference on Web Science. Amsterdam: Association for Computing Machinery, 2019: 183-192.

[51] ZHONG L J, YANG W Z, YUAN T T, et al. Survey of abnormal user identification technology in social network [J]. Computer Engineering and Applications, 2018, 54(16): 13-23.

[52] CHU Z, GIANVECCHIO S, WANG H, et al. Who is tweeting on Twitter: human, bot, or cyborg? [C]// The 26th Annual Computer Security Applications Conference. Austin:, Association for Computing Machinery, 2010: 21-30.

## 作者简介

**李阳阳**(1987—),男,江苏人,博士,高级工程师,主要研究方向为内容安全与社会信息网络;

E-mail: liyangyang@ cetc. com. cn

**曹银浩**(1997—),男,河北人,硕士研究生,主要研究方向为内容安全;

**杨英光**(1996—),男,内蒙人,硕士研究生,主要研究方向为社交机器人和水军检测;

**金　昊**(1992—),女,安徽人,博士,工程师,主要研究方向为社交网络分析;

**杨阳朝**(1986—),男,山西人,博士,高级工程师,主要研究方向为人工智能,数据挖掘,社交网络等;

**石　珺**(1987—),女,黑龙江人,博士,高级工程师,主要研究方向为网络空间内容与认识域安全;

**李志鹏**(1989—),男,湖南人,博士,工程师,主要研究方向为数据挖掘与网络安全。