文章编号:1006-2475(2022)03-0001-06

基于生成对抗网络的社交机器人检测

李阳阳[1],杨英光[2]

(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. 中国科学技术大学网络空间安全学院,安徽 合肥 230026)

摘要:推特作为一个有着上亿活跃用户的社交媒体,有近 15% 的机器账户通过自动化程序被控制,其中一些机器账号为传播恶意信息的恶意账号。虽然研究者开发了大量复杂的机器账号检测方法,但这些方法都需要有关机器账号的先验知识,并且泛化性不高。为了解决这些问题,提出使用生成对抗网络中的判别器来进行机器账号检测,使得只需要真实账号的示例即可得到良好的检测模型,并在一个流行数据集做实验,AUC 达到了 94% 的分类效果。

关键词:社交机器人; 生成对抗网络; 机器账号检测
中图分类号:TP391　　**文献标志码**:A　　**DOI**: 10.3969/j.issn.1006-2475.2022.03.001

Social Bots Detection Based on Generative Adversarial Networks

LI Yang-yang[1], YANG Ying-guang[2]

(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 Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Twitter is a social media with hundreds of millions of active users. Nearly 15% of bot accounts are controlled by automated programs. Some of these bot accounts are malicious account that spread malicious information. Although researchers have developed a large number of sophisticated bot account detection methods, they all require prior knowledge of bot accounts which are lack of generalization. In order to solve these problems, this paper proposes to use the discriminator from generative adversarial network for bot account detection. This makes it possible to obtain a good detection model with the examples of real accounts. Experiments on a popular dataset show that the AUC achieves 94% classification effect.

Key words: social bots; generative adversarial networks; bot account detection

0 引　言

推特(Twitter)作为一个在线社交网络,拥有上亿数量的活跃用户。用户可以通过推文(tweet)、提及(mention)、转发(retweet)等动作与其他用户进行交互,通过这些交互能够连接全球各地的人,彼此间相互影响。新闻、想法都可以通过这些连接进行传播。对于所有注册的用户,都可以通过网页或者应用编程接口(API)访问推特平台提供的服务。这种方式使得人们能够编写软件来完成控制账户自动进行发送推文或者转发推文等动作[1],这些账户被称作机器账号又被称作社交机器人。随着机器账号的发展,机器账号的活动已经在包括政治[2-4]、健康[5-8]和商业[9]等多个领域中被报道,2017 年有项研究[10]估计活跃用户中有 9% ~15% 为机器账号。有一部分机器账号为恶意账号,发布恶意及有害信息。越来越多人企图使用机器账号达到影响政治经济、引导对立等目的,人们生活的多个方面都受到机器账号影响甚至威胁。所以对社交媒体中机器账号的检测成为了一个极其重要的课题,不同的机器账号检测技术也得到了长足发展。

令人担心的是,机器账号随着时间不断进化和迭代,不断采用更加复杂的技术,例如改变讨论的话题和推文的文本模式,使得机器账号与真实账号的差异性越来越小[11],更加难以对机器账号进行检测。研究者们不断提出更加复杂的方法来加入这场与机器账号的竞赛之中。在过去几年研究者们已经提出了多种基于机器学习的针对推特平台上的机器账号检测的框架。然而这些检测方法仍然面临着 2 个重要的挑战:被动性和泛化性。现有的检测方案都是被动式检测方案[12],这种方法的检测流程是:先观察机器账号的存在,收集相关数据集进行分析,针对分析的

收稿日期:2021-05-08; **修回日期**:2021-06-09
基金项目:国家自然科学基金资助项目(U20B2053); 海南省重大科技计划项目(ZDKJ2019008)
作者简介:李阳阳(1987—),男,江苏扬州人,高级工程师,博士,研究方向:内容安全,社会信息网络,E-mail: liyangyang@cetc.com.cn; 通信作者:杨英光(1996—),男,硕士研究生,研究方向:社交机器人检测,水军检测,E-mail: dao@mail.ustc.edu。

结果设计检测方案,使用检测方案进行检测,机器账号为了规避检测继续进化。这种被动式检测方案使得检测方案需要机器账号的先验知识,并落后于机器账号的发展,机器账号的可分类模式被发现之前,会有很长的时间潜伏在社交网络平台之中。泛化性是为了能够检测不同训练数据集中的机器账号,这个作为一个对抗性的问题至关重要。因为新型机器账号总是能够通过设计来规避检测[13]。目前已有的基于机器学习的检测方法都是将其视为二分类问题,并从收集到的带有 2 类标签的数据来训练模型。但是这些方法一旦遇到不同于训练集特征的新一代机器帐号时,泛化性不足[14],检测效果大大降低。

为了一定程度上解决上述 2 种问题,本文提出一个假设:假如本文对于数据集中的真实账号足够了解,那么本文完全可以通过从真实账号学习到的潜在表征来区分真实账号与机器帐号。这样本文只需要收集大量的真实账号的数据,学习这些真实账号的潜在模式。这样的好处有:1)将以往的二分类问题转化为单一类别的检测方法[15],不需要依赖机器账号的先验知识;2)能够以主动式[12]的或者说是对抗式的思路分析当前的检测方法是否存在检测弱点,从而能够为检测方法提供进一步的改进思路,提升检测方法的稳定性;3)由于只学习真实账号的潜在特征,只要在学习到的潜在可识别模式上与真实账号有差异的机器帐号出现时,就能轻易地被检测出来,大大提高检测方法的泛化性。

本文提出一个检测方法来对机器账号进行检测。本文检测方法使用计算机图像领域中常用的生成对抗网络[16]来训练检测模型,将随机噪声作为生成器的输入来生成虚假数据,将生成的虚假数据和数据集中的真实账号作为真实数据输入到判别器中,并不断迭代,这样判别器为了能够识别生成器生成的质量越来越好的虚假数据,就必须对输入的真实数据即真实账号充分地学习,学习到潜在的识别模式。这样本文就能拿到训练好的判别器来识别机器账号和真实账号。

通过使用生成对抗网络作为本文的检测方案,从而给对抗性思路提供又一种实现。同时使用生成器不断进行迭代,生成的虚假数据成功逃脱了当前最先进的检测器的检测。本文也将二分类问题变成了单分类问题,用经过多轮训练的判别器进行机器账号检测,泛化性的问题也在一定程度上得到了解决。

本文的主要工作有:

1)提出了一个假设,即如果对于真实账号的数据进行充分学习,本文不需要机器账号的数据也能对其进行分类。

2)第 1 个将生成对抗网络引入到机器账号检测领域中的研究,并提供了一种对抗性思路的实现。

3)通过使用生成对抗网络的判别器进行机器帐号检测,提高了检测的泛化性。

1 相关工作

1.1 有监督方法

目前有大量使用监督机器学习的相关研究,这些方法主要从账户元信息[14]、社交网络拓扑[17]和推文内容[10]上提取大量特征。这些方法都依赖带标注的数据集。Botomete[18-19]方案利用随机森林算法,提取特征训练 7 个不同的分类器,在十倍交叉验证下的 AUC 值为 0.95 AUC。文献[20]中用贝叶斯算法做机器账号检测。文献[21]用卷积神经网络和长短期记忆网络联合抽取文本特征和时间序列信息进行账号检测。文献[22]使用异构图神经网络,基于拓扑中的账户之间总会产生“聚合”的假设,根据其他账户如何与这个账户“聚合”即可进行机器账号的检测。文献[15]将机器账号检测视为单分类问题,选取了几个单分类检测算法对比了不同算法的检测效果。

1.2 无监督方法

也有一些研究使用无监督方法,这些方法对于跨域检测的鲁棒性更好,并且更适合发现机器账号的协同作用。因为单独考虑账号时,账户可能并不会表现出差异,从而被有监督方法忽略。文献[23]通过检测重复出现的嵌入式的 URL 的内容发现机器人群组。文献[24]提取使用马尔可夫集群算法来识别机器账号群组。文献[25]使用聚类方法通过账户特征和使用情况来查找机器人群组。文献[26]分析发现有些账号总是同一时间转发推文,利用这个模式,以 0.94的精度寻找到了机器人群组。

2 数据集与特征提取

本文使用的数据集来自 Cresci 与合作者完成的文献[11]中公布的数据集。该数据集包含了不同类型的机器账号信息,还包括了每一个账号最近发表的推文及推文信息,学者们利用该数据集做了大量研究。

2.1 数据集

该数据集的概况如表 1 所示,总共包含 9386 个账号,3474 个真实账号,其余的机器账号被划分成 4 种不同的类型。数据集中提供了账号发表的推文数据,同时也提供了账号的元数据,如朋友和关注者的数量、是否为默认头像等。每一个账号都经过了人工验证,以确认分类是否正确。

表 1 数据集概述

数据集	描述	账号数	推文数
Genuine accounts(gc)	经验证真人操作的账号	3474	8377522
Social spambots #1(s1)	意大利政治候选人的转推账号	991	1610176
Social spambots #2(s2)	付费应用垃圾邮件发送账号	3457	428542
Social spambots #3(s3)	亚马逊商品广告账号	464	1418626
Traditional spambots #1(t1)	恶意软件垃圾邮件发送账号	1000	145094

2.2　特征提取

为了能够训练出效果良好的分类模型,本文从数据集中抽取了 41 个特征,抽取的特征情况如表 2 所示。

表 2　抽取的特征及解释

特征变量名	描述
screen_name	昵称长度
usr_prfim	是否包含图片
usr_prfbg	是否包含背景图
usr_prfbn	是否有横幅图
usr_twtrt	发推比率
usr_ffrat	追随者与朋友之比
usr_faves	喜欢字段
twt_srcts	source type(st)
twt_src2b	st = bot_thebig2s 数
twt_int30	30 min 间隔发推数与总推文占比
twt_srctw	st = twt_official 数
twt_srcbu	st = bot_notsures 数
twt_srcpp	st = pop_platform 数
twt_srcas	st = pop_assister 数
twt_srcpu	st = pop_notsures 数
twt_srcna	st = NA 数
twt_srcbe	st = bot_enabler 数
twt_wrdsd	推文词数的方差
twt_atsmn	推文的提及平均数
twt_atssd	推文提及数的方差
tweets	推文总数目
usr_frnds	朋友数目
usr_fllws	关注者数目
usr_verif	verified 字段
usr_actyr	账户年龄
usr_allrt	是否都是转推
twt_langs	语言码长度
twt_srces	source 字段长度
twt_rtwts	推文中转推占比
twt_rplys	推文回复占比
twt_int15	15 min 间隔发推数与总推文占比
twt_wrdmn	推文平均词数
twt_bunch	2 s 间隔发推占比
twt_max15	15 s 间隔最大数占比
twt_min15	15 s 间隔最小数占比
twt_max30	30 s 间隔最大数占比
twt_min30	30 s 间隔最小数占比
twt_hshmn	hashtag 的平均数
twt_hshsd	hashtag 的方差
twt_urlmn	推文 url 的平均数
twt_urlsd	推文 url 数的方差

3　方　法

3.1　基线检测方法

本文使用 tweetbotornot2 作为基线方法进行对比。该方法基于监督分类器 xgboost[27],从用户账户的属性、推文统计信息和基于文本的模式抽取了 3 大类特征。该方法的分类效果与国际最流行检测器 Botometer[18-19] 不分伯仲,但 Botometer 未开源,而 tweetbotornot2 已经开源,使用 R 语言实现(tweetbotornot2. mikewk. com),可以部署并进行机器账号检测。本文在对比实验时,使用上述抽取的 41 维特征对 tweetbotornot2 进行训练,所以下述简称为 TW-41。

3.2　提出的方法

在实验中发现,TW-41 对上述数据集中抽取的特征进行训练后,在测试集中得到的评价指标 AUC 能达到 0.98,效果十分不错,但该方法的缺点是跨数据集的泛化性较差。

于是本文考虑能否使用一种对抗性的方法,从随机噪声中产生特征数据集,通过不断迭代,从而规避 TW-41 的检测,另外当前的机器检测方法都是同时依赖于真实账号和机器账号的示例数据集,从示例数据集中学习 2 种类型数据的差异,从而完成机器账号检测的任务。但是由于机器账号是不断进化的,通过示例数据集的学习并进行分类的方案无法应对新的变种机器账号。同时真实账号是易于获得的,本文考虑能否通过单分类方法,即只需要真实账号的数据,对真实账号特征分布充分学习,从而能够对机器账号进行检测。这样也能够大大提高检测方案的泛化性。

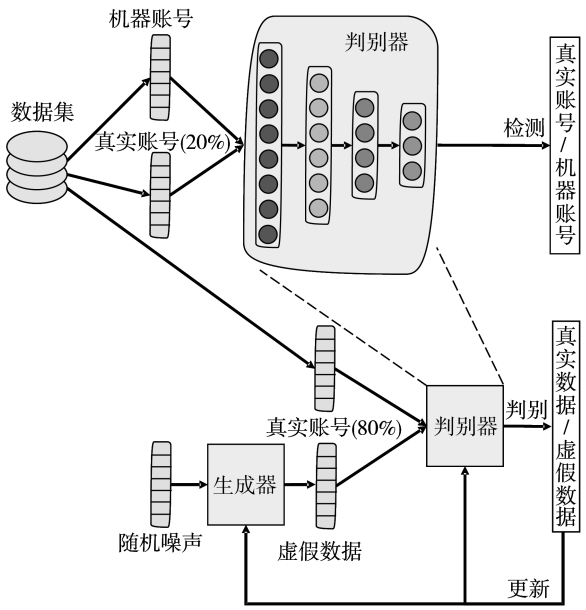

图 1　使用生成对抗网络检测方法的流程

针对上述的思考,本文决定使用生成对抗网络来解决上述 2 个问题。图 1 展示了本文的检测方法流程:将随机噪声 $z \sim p_z(z)$ 作为输入到生成器,让生成器产生虚假数据,生成器 G 使用多层感知机实现,如式(1):

$$G(v;\ \theta^G) = f(z;\theta^f) \tag{1}$$

其中,f 使用多层感知机实现,θ^f 是 f 的参数,变量 z

来自高斯分布,如式(2):

$$p_z(z) = N(z_v^{\mathrm{T}}, \sigma^2 I) \tag{2}$$

其中,$z_v \in \mathbb{R}^{d\times 1}$,$d$ 等于对账号抽取的特征向量维度。为了能够让生成器生成足够接近真实账号的数据欺骗判别器,本文定义生成器的损失函数如式(3):

$$L^G = \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z))) \tag{3}$$

其中,z 为随机噪声,G 为生成器,D 为判别器。最小化 L^G 即可优化生成器。本文使用数据集中只包含真实账号中比例为 80% 的数据提供给使用深度神经网络实现的判别器 D,作为真实数据输入进行学习,如式(4):

$$D(\boldsymbol{x}^i;\theta^D) = \frac{1}{1+\exp(f(\boldsymbol{x}^i;\theta^f))} \tag{4}$$

其中,$\boldsymbol{x}^i$ 是特征向量,θ^f 是判别器 f 的参数。将虚假数据和对应的标签及真实数据和对应的标签输入到判别器,判别器不断进行迭代,损失函数如式(5):

$$L^D = \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(\vec{x}^i)) + \frac{1}{m}\sum_{i=1}^{m}\log(D(x^i)) \tag{5}$$

其中,$\boldsymbol{x}^i$ 是真实数据抽取的特征向量,$\vec{\boldsymbol{x}}^i$ 是生成器生成的虚假特征向量。通过最小化损失函数让生成器 G 产生与真实数据足够相似的数据,来迷惑生成对抗网络中的判别器 D,同时逃避 tweetbotornot2 方法的检测。由于生成器产生的数据不断混淆判别器。判别器必须充分学习真实数据的潜在特征,才能区分是真实数据还是生成器产生的虚假数据。这样对真实数据学习的足够好的判别器能够对真实账号和机器账号进行区分,即使判别器在数据集中从来没见过机器账号。

4 实验与结果

4.1 实验设置

生成对抗网络中的判别器和生成器使用的都是有 3 个隐藏层的深度神经网络。生成器的输入是随机产生的 64 维随机噪声,输出是代表虚假数据的 41 维特征向量。判别器的输入是 41 维特征向量,输出[0,1]的数值代表该特征向量属于真实数据的概率值。3 个隐藏层的神经元数量分别是 1024、512、256,激活函数是能够减少梯度稀疏程度的 LeakyReLU,并添加了批量正则化层。本文使用 Adam 方法来优化网络,并设置批尺寸为 batch_size = 32,学习率大小设置为 lr = 0.0002,可使网络收敛的更加稳定。

4.2 逃脱检测

为了能够评估使用生成对抗网络中的生成器是否能够产生与真实账号高度相似的数据,并逃脱当下最流行检测器的检测。本文选择训练好的 TW-41 作为评估方法,评估生成器生成的虚假数据的效果。生成器的输入是噪声数据,输出是特征向量。判别器不断输出对这些特征向量属于真实数据的概率值,将其与数据的标签作为损失函数的输入,计算损失,通过反向传播优化生成器。本文将迭代次数设置为 500 次,迭代次数与 TW-41 的判断精度的结果如图 2 所示。

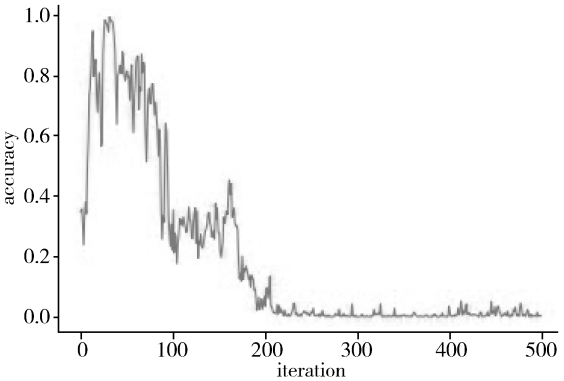

图 2 TW-41 生成对抗网络随迭代次数的检测精度

从图 2 可以看出,随着生成器的迭代,生成器生成的数据从完全随机逐渐在特征上接近真实数据,但一开始产生的数据与 TW-41 用于训练的机器账号相似,被识别出来,所以有了检测正确率的上升。随着迭代,当生成器产生的数据已经与真实账号的分布一致时,TW-41 已经无法识别这是生成的虚假数据,将其识别为真实账号,所以精确度逐渐收敛到 0。最后生成的数据完全逃脱了检测。

4.3 判别器分类效果

本文使用数据集中的真实账号和生成器产生的虚假数据对判别器进行训练。这样如果判别器对真实账号的模式学习的足够好,理论上不需要收集机器账号就能够对机器账号进行检测,一定程度上能够解决检测方法跨数据集的泛化性问题。为了验证该想法,本文将随机抽取真实账号的 80% 作为训练数据,并将剩余的 20% 的真实账号和数量大致相等的 Social spambots #2(s2)的 20% 作为测试数据。设置生成对抗网络的迭代次数(epoch)分别为 250、500、750、1000。由于是单分类问题,并且判别器输出是概率值,所以本文选择 AUC 作为评价指标。不同迭代次数的 AUC 值如表 3 所示。

表 3 GAN 迭代次数及其判别器对测试数据集的 AUC 值

迭代次数	AUC 值
250	0.948
500	0.952
750	0.922
1000	0.926

从表 3 可以看出,训练得到的判别器其 AUC 值均能达到 90% 以上,得到了很好的分类效果。

本文用真实账号与 4 种机器账号数据中的一种进行组合成训练集,参数设置如表 4 所示。使用组合后的训练集中 80% 的数据对 TW-41 方法进行训练,

仅使用训练集中 80% 的真实账号对生成对抗网络中的判别器进行训练。挑选出其他机器账号数据的 20% 与剩余的 20% 的真实账号数据组成测试集。用 TW-41 和生成对抗网络中 GAN-41 的判别器分别在测试集上进行测试,使用 AUC 作为评价指标。结果如表 5 所示。

表 4　TW-41 与 GAN 对比实验的训练集和测试集参数设置

训练集	测试集
gc(80%) + s1(80%)	gc(20%) + s2(20%)
gc(80%) + s2(80%)	gc(20%) + s2(20%)
gc(80%) + s3(80%)	gc(20%) + s2(20%)
gc(80%) + t1(80%)	gc(20%) + s2(20%)

表 5　TW-41 与 GAN-41 的 AUC 值对比

训练集/测试集	TW-41	GAN-41
gc + s1/gc + s2	0.8917	**0.948**
gc + s2/gc + s2	**0.9993**	0.948
gc + s3/gc + s2	0.6360	**0.948**
gc + t1/gc + s2	0.9456	**0.948**

从表 5 可以看出 TW-41 只有在使用与训练集相同分布的测试集上测试时能够达到非常不错的效果,AUC 指标超过 99%。但当测试集中的机器人数据没有在训练集中出现时,TW-41 的效果有些下降,此时 GAN 的判别器识别机器人的效果要更加优秀,同时也说明了本文检测方案的泛化性比 TW-41 等基于传统机器学习的检测方法更强。

为了能够进一步查看不同类型的特征对于检测效果的影响,本文仅使用账户元数据中 13 个特征数据训练 TW-13 检测器,以及仅仅使用剩下的 28 个从推文中抽取的内容特征训练 TW-28 检测器分别与对应特征维度的生成对抗网络进行对比重复上述实验,结果如表 6 和表 7 所示。从结果中能够得出本文的检测方法在特征维度不同时,泛化性依然强于 tweet-botornot2。

表 6　TW-13 与 GAN-13 的 AUC 值对比

训练集/测试集	TW-13	GAN-13
gc + s1/gc + s2	0.9667	**0.9926**
gc + s2/gc + s2	**0.9999**	0.9926
gc + s3/gc + s2	0.6307	**0.9926**
gc + t1/gc + s2	0.9480	**0.9926**

表 7　TW-28 与 GAN-28 的 AUC 值对比

训练集/测试集	TW-28	GAN-28
gc + s1/gc + s2	0.7215	**0.9736**
gc + s2/gc + s2	**0.9970**	0.9736
gc + s3/gc + s2	0.7262	**0.9736**
gc + t1/gc + s2	0.9247	**0.9736**

5　讨　论

通过上述实验的结果可以发现只需要数据集中的真实账号和将随机噪声输入到生成器产生的生成数据来训练判别器,判别器就能够以 AUC 值超过 94% 的高精确度来检测机器账号。本文分析原因是由于判别器对真实账号抽取的特征学习到了良好的可识别模式,使其达到了良好的分类效果。并且还可以发现,当生成器不断迭代时,TW-41 对生成的数据检测的精度呈现先上升后下降的趋势,其原因是由于生成器生成数据的质量不断上升,当其迭代到 50 次左右时,生成的数据与真实数据接近,但仍然存在可以被 TW-41 识别的特征,TW-41 便能够以很高的精度检测出生成器生成的与真实数据相似的虚假数据。当迭代到更高次数时,生成的数据与真实账号更加相似,生成的数据质量越来越好,使得 TW-41 的检测精度不断下降直到接近 0,逃脱了检测,可见即使是当前检测精度极高的先进检测方法,仍然存在着检测弱点。

通过上述研究可以知道本文不需要提供机器账号的数据就能训练出能够检测机器账号的模型,也将二分类问题转化为单一分类问题,并大大提高了检测方案的泛化性,使得本方法相较其他方法更能够检测最新一代的机器账号。

6　结束语

本文提出了使用生成对抗网络来进行机器账号检测。使用真实账号作为真实数据训练判别器,使用训练好的判别器就可以对机器账号进行检测,本文检测方法能够实现 AUC 值超过 94% 的高精确度,并且不需要任何机器帐号的先验知识。同时本文也用时下最先进的 TW-41 检测器,评判生成器生成的虚假数据的攻击效果。

本文研究了如果有模型能够对真实账号的特征学习到良好的可识别模式,那么不需要对机器账号有先验知识就可以达到很好的分类效果。

目前本文方法只在一种数据集中进行了训练,这种方式得到的模型的稳定性还不够好,不能够在所有的数据集组合作为测试集时,AUC 值都达到 90% 以上。未来笔者将收集更多数据集,用不同数据集中的真实账号训练同一个判别器,达到更好的分类效果,进一步借鉴不同检测方案中特征抽取的方式,用不同方式训练出多个判别器。多个判别器能够在不同的特征空间中学习到可分类模式,这样由不同分类模式的判别器组成的集成分类器就能对单一账号进行打分,得出一个鲁棒性更高的检测器。同时笔者也将分析生成器生成的虚假数据和真实数据的相似性和差异性,分析导致虚假数据逃脱 TW-41 检测的原因,找出 TW-41 等基于传统机器学习方法实际存在的弱点。

参考文献：

[1] CHU Z, GIANVECCHIO S, WANG H N, et al. Who is tweeting on twitter: Human, bot, or cyborg? [C]// Proceedings of the 26th Annual Computer Security Applications Conference. 2010:21-30.

[2] BESSI A, FERRARA E. Social bots distort the 2016 U. S. presidential election online discussion[J]. First Monday, 2016,21(11): DOI: https://doi. org/10. 5210/fm. v21i11. 7090.

[3] FERRARA E. Disinformation and social bot operations in the run up to the 2017 french presidential election[J]. First Monday, 2017,22(8): DOI: https://doi. org/10. 5210/fm. v22i8. 8005.

[4] STELLA M, FERRARA E, DE DOMENICO 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.

[5] ALLEM J P, FERRARA E. Could social bots pose a threat to public health? [J]. American Journal of Public Health, 2018,108(8):1005-1006.

[6] ALLEM J P, FERRARA E, UPPU S P, et al. E-cigarette surveillance with social media data: Social bots, emerging topics, and trends[J]. JMIR Public Health and Surveillance, 2017,3(4):e98.

[7] BRONIATOWSKI D A, JAMISON A M, QI S, et al. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate[J]. American Journal of Public Health, 2018,108(10):1378-1384.

[8] DEB A, MAJMUNDAR A, SEO S, et al. Social bots for online public health interventions[C]// 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2018:1-4.

[9] 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, 2019, 13(2):1-27.

[10] VAROL O, FERRARA E, DAVIS C A, et al. Online human-bot interactions: Detection, estimation, and characterization[J]. Social and Information Networks, arxiv preprint arXiv:1703. 03107, 2017.

[11] 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. 2017:963-972.

[12] CRESCI S, PETROCCHI M, SPOGNARDI A, et al. From reaction to proaction: Unexplored ways to the detection of evolving spambots[C]// Companion of the Web Conference 2018. 2018:1469-1470.

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

[14] 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.

[15] RODRIGUEZ-RUIZ J, MATA-SANCHEZ J I, MONROY R, et al. A one-class classification approach for bot detection on twitter[J]. Computers & Security, 2020, 91: 101715. 1-101715. 14.

[16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020,63(11):139-144.

[17] LOYOLA-GONZALEZ O, MONROY R, RODRIGUEZ J, et al. Contrast pattern-based classification for bot detection on twitter[J]. IEEE Access, 2019,7:45800-45817.

[18] 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. 2016:273-274.

[19] SAYYADIHARIKANDEH M, VAROL O, YANG K C, et al. Detection of novel social bots by ensembles of specialized classifiers[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020:2725-2732.

[20] KANTEPE M, GANIZ M C. Preprocessing framework for Twitter bot detection[C]// 2017 International Conference on Computer Science and Engineering. 2017:630-634.

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

[22] LIU Z, CHEN C, YANG X, et al. Heterogeneous graph neural networks for malicious account detection[J]. Machine Learning, arXiv preprint arXiv:2002. 12307, 2018.

[23] CHEN Z, SUBRAMANIAN D. An unsupervised approach to detect spam campaigns that use botnets on Twitter[J]. Social and Information Networks, arXiv preprint arXiv: 1804. 05232, 2018.

[24] AHMED F, ABULAISH M. A generic statistical approach for spam detection in online social networks[J]. Computer Communications, 2013,36(10-11):1120-1129.

[25] MILLER Z, DICKINSON B, DEITRICK W, et al. Twitter spammer detection using data stream clustering[J]. Information Sciences, 2014,260:64-73.

[26] CHAVOSHI N, HAMOONI H, MUEEN A. Identifying correlated bots in Twitter[C]// International Conference on Social Informatics. 2016:14-21.

[27] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.