

Cerberus: 基于深度学习的跨网站社交机器人检测系统

汤家伟¹, 刘育杉¹, 高敏¹, 宫庆媛², 王新¹, 陈阳¹

(1. 复旦大学计算机科学技术学院, 上海 200438;

2. 复旦大学智能复杂体系基础理论与关键技术实验室, 上海 200438)

摘要: 社交网站吸引了数十亿用户, 影响着人们的生活方式。社交网站作为开放平台, 注册加入的门槛较低, 社交机器人能够轻易地注册, 并进行舆论导向控制、不实信息传播等有害活动, 以谋取利益。单一社交网站的机器人检测系统往往需要依赖用户的历史行为数据进行分析。因此, 社交机器人在被识别出之前往往已经成功实施了恶意攻击。为尽早地识别出社交机器人, 提出了跨网站社交机器人检测系统 Cerberus。Cerberus 可以解决用户早期在单个社交网站上数据不充足导致的用户识别“冷启动”的问题。Cerberus 使用用户在 Medium 网站上的个人信息和历史活动信息, 对用户链接在 Twitter 上的账号是否为社交机器人账号进行预测。结果表明, 该系统的 AUC 值可达 0.7552, 具有良好的识别准确性。

关键词: 在线社交网络; 社交机器人检测; 跨网站链接; 深度学习; 冷启动用户

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202436

Cerberus: cross-site social bot detection system based on deep learning

TANG Jiawei¹, LIU Yushan¹, GAO Min¹, GONG Qingyuan², WANG Xin¹, CHEN Yang¹

1. School of Computer Science, Fudan University, Shanghai 200438, China

2. Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200438, China

Abstract: Social networking sites have attracted billions of users and influence people's lifestyles. However, as open platform with low requirements for registration and joining, it is inevitable that social bots are able to easily register and do harmful things such as controlling public opinions and spreading inaccurate information for profit. Nevertheless, single-site social bot detection systems often rely on historical behavioral data to identify bots, and the detection occurred after the social bots have implemented their attacks. To identify social bots as early as possible, this paper proposed Cerberus, a cross-site social bot detection system. Cerberus can solve the cold-start problem of user identification caused by insufficient user data on a single platform at an early stage. Cerberus used personal information and historical activity on the Medium website of users to make prediction about whether a user's account on Twitter was a social bot. The results from our experiments show that the AUC score of Cerberus can reach 0.7522, which has good recognition accuracy.

Key words: online social network, social bot detection, cross-site linking, deep learning, cold-start user

收稿日期: 2024-08-07; 修回日期: 2024-09-29

通信作者: 陈阳, chenyang@fudan.edu.cn

基金项目: 国家重点研发计划 (No.2022YFB3102901); 国家自然科学基金项目 (No.62072115, No.62102094); 上海市科技创新行动计划 (No.22510713600)

Foundation Items: National Key R&D Program of China (No.2022YFB3102901), The National Natural Science Foundation of China (No.62072115, No.62102094), Shanghai Science and Technology Innovation Action Plan (No.22510713600)

0 引言

随着互联网的发展,人们逐渐花费大量时间在社交网站上。Facebook^[1]和 Twitter^[2]是非常重要的媒体工具,至今已经吸引了数十亿用户。这些社交网站吸引了拥有不同背景、不同兴趣和职业的用户,并为用户进行社交互动、信息分享和内容创造提供了平台。

由于社交网站的开放性,社交机器人很容易加入其中。这些社交机器人会进行控制社交网站中的舆论导向、左右金融决策、传播不实信息^[3]等活动,并影响对社交网络的相关研究^[4-5],因此设计一个检测系统对用户的身份进行预测并加以标注,使社交网站重点监管社交机器人用户,可以很大程度上提高普通用户的体验。

目前,社交机器人的识别主要依赖于监督式机器学习方法^[6-9]。这类方法通过提取用户在活动时产生的各种信息与其自身设置的描述性信息来识别社交机器人。然而,这些方法在面对信息不完整的用户时,往往难以预测其是否为社交机器人。因此,当这类算法成功识别社交机器人时,该社交机器人可能已经对社交网站上的其他用户造成了一定的负面影响。此外,用户在注册初期,由于活动较少,缺乏足够的用户历史活动数据,存在“冷启动”问题,因此,这类方法难以对用户身份进行预测。Venkatadri等^[10]的研究表明,同一用户在不同社交网站上的部分账号属性具有一致性。De Meo等^[11]发现,用户在多个社交网站中的某些行为具有一定的关联性,这一发现可以通过跨网站数据扩充用户账号特征,提升用户账号特征的互补性与多样性。根据跨网站用户账号的关联性和其特征的互补性与多样性,本文利用跨网站链接功能,通过链接用户在其他社交网站上的账号,使用其在该网站上的个人信息和活动数据完善用户在目标社交网站上的身份信息。通过跨社交网站扩充用户账号特征,较早地对用户的身份进行预测,解决用户身份识别的“冷启动”问题,减少社交机器人对社交网络 and 用户造成的影响。

本文提出了一个基于深度学习的跨网站社交机器人检测系统 Cerberus。Cerberus 通过结合用户在两个社交网站产生的描述性特征与文本内容来预测该用户是否为社交机器人。对于文本内容,本文使用深度学习来提取相关特征,将其与用户的描述性

特征同时输入监督式学习模型中进行二分类,以预测该用户是否为社交机器人。Cerberus 利用 Medium-Twitter 跨社交网站的真实世界数据开展实验。本文的主要贡献如下。

(1) 本文首次提出利用跨网站用户行为数据进行在线社交机器人检测。基于用户在不同社交网站上账号属性的一致性,利用同一用户在其他社交网站上的行为数据来预测其在目标社交网站上是否为社交机器人。

(2) 本文提出了一个基于深度学习的跨网站社交机器人检测系统 Cerberus。具体而言, Cerberus 可以通过用户在一个社交网站上的数据帮助与其关联的社交网站预测该用户是否为社交机器人,以解决“冷启动”问题。

(3) 对主流社交网站 Medium 与 Twitter 进行案例研究,在大规模真实用户数据上对 Cerberus 的检测效果进行了评估。与已有检测系统对比, Cerberus 达到了最好的检测效果。

1 相关工作

1.1 社交机器人检测

目前,已有大量工作针对社交机器人的账号和行为特征,提出了不同的社交机器人检测算法。主要有提取用户的行为规则、基于特征的监督式学习的方法,从用户的社交关系图进行图分析、使用随机游走或分割策略进行识别的图方法,以及基于用户社交关系图、尝试用半监督式学习的方法。

部分研究根据发现的社交网站中的用户行为进行社交机器人检测。Fanzil等^[12]根据用户的行为特征,包括其简介、时序特征、活动行为和文本内容等,利用深度学习来检测社交机器人。张艳梅等^[13]与夏崇欢等^[14]基于粉丝数、关注数与发布的微博数等数据,识别微博平台中的“网络水军”与恶意用户。Arin等^[15]结合长短期记忆网络捕捉人类和机器人的复杂社交媒体活动,以识别社交机器人。Liu等^[16]在使用元数据、文本内容和网络结构等特征的基础上,通过引入针对不同社区的专家混合层,更准确地识别社交机器人。

部分研究使用图分析的方法进行社交机器人检测。社交网站中,用户间进行社交行为从而产生社交关系,用户间因评论、点赞、主页访问等互动形成用户之间的交互关系,进而构成社交关系图。程晓涛等^[17]认为“网络水军”策略不断演变,虽

然“网络水军”可以变更自身的微博内容与转发行为,但无法改变与网站中正常用户的连结关系,形成的结构图具有一定的稳定性。Li等^[18]将机器学习与图分析方法相结合以识别社交机器人。Liu等^[19]首先将社交图分为多个子图,然后设计了社区游走规则并生成每个账号的表示向量,最后使用表示向量对模型进行训练和识别社交机器人。Mendoza等^[20]在图分析方法的基础上引入了半监督学习并取得了较好的结果。Feng等^[21]强调注重用户关系及其影响力的异质性,通过构建用户形成的异质图的拓扑结构来识别社交机器人。

对于社交机器人的检测也有一些特殊的方法。例如著名的“蜜罐陷阱”^[22],检测者通过构建与普通用户不进行社交的账号作为“蜜罐陷阱”,诱使社交机器人关注,以诱捕社交机器人。夏崇欢等^[14]在构建数据集时使用“蜜罐陷阱”来诱捕社交机器人。

虽然上述研究均取得了较好的结果,但都是从研究对象所在的社交网站获取数据并识别社交机器人,而这往往需要社交机器人活动一段时间并产生相应的数据。也就是说,在用户新注册后的一段时间内,上述研究均不能较好地预测该用户是否为社交机器人。

1.2 跨网站用户行为分析

随着多种不同类型社交网站的流行,以及这些社交网站之间进行合作,用户有更多的机会在多个社交网站上注册账号,同时使用多项服务。“跨网站链接”功能在此背景下应运而生,并广受用户的喜爱,以Foursquare为例,有研究表明其中57.10%的用户使用了该功能^[23]。当一个用户拥有多个社交网站账号,用户通过使用跨网站链接功能,可以同时多个网站发布相同的帖子。同时,用户也可以在每个社交网站单独进行社交活动。同一用户在不同社交网站的活动具有相似性也有差异性。

Liu等^[24]基于社交关系结构分析了用户在多个社交网站上行为记录之间的相似性。Goga等^[25]研究了用户在不同的社交网站上语言风格之间以及时间空间信息之间的相似性,并且发现利用用户在不同网站上发布信息的时空轨迹,最能准确识别出同一用户。

Sang等^[26]研究了用户在Twitter上的历史行为和Youtube上的视频观看兴趣之间的关系,通过将用户的历史行为按照时间线分段,分别研究用户

在同一时间段内,在Twitter上发布的话题以及在Youtube上的视频观看兴趣,提出了一种利用用户在Twitter上的历史行为数据,对Youtube“冷启动”场景下进行视频推荐的系统。Kong等^[27]研究了用户在Foursquare和Twitter上行为之间的关联。Venkatadri等^[10]的研究表明,同一用户在不同社交网站上的账号属性具有较高的一致性。

De Meo等^[11]发现,用户在多个社交网站产生的行为数据具有广泛的实际应用。例如,可以通过聚合用户在不同社交网站上的个人信息,建立一个更全面的用户在线档案,利用跨社交网站的信息解决在单一社交网站中用户标签识别“冷启动”问题。可以根据用户在某一社交网站的行为信息,优化另一个社交网站中对用户的推荐内容。

鉴于新兴社交网站不断涌现,用户普遍使用多个社交网站,多数新兴社交网站中会持续存在较多处于“冷启动”阶段的用户,此时,社交机器人还未对社交网站实施恶意攻击,实现对社交机器人的早期检测将有助于维护新兴社交网站的正常运营。对社交机器人检测的研究需要从单个社交网站拓展到跨社交网站中。在“冷启动”阶段,用户的行为数据、社交关系等都比较有限,这给准确预测用户是否为社交机器人带来了很大难度。据调研,有的社交网站超过一半的用户使用了跨网站链接功能,在新兴社交网站上链接了他们的其他社交网站账号。Gong等^[28]提出了一种基于跨网站数据预测用户是否会在另一社交平台成为有影响力的用户的方法,并基于Medium与Twitter两个热门社交网站验证了该解决方案的有效性。如果可以通过跨社交网站丰富用户的行为数据,尽早地对用户的身份进行识别,则可以大大减小社交机器人对社交网站造成的影响。

2 研究背景与数据获取

本节首先介绍Medium与Twitter两个社交网站,及其用户产生数据的行为。然后,对抓取模块进行详细的描述,包括用户是否为社交机器人的标签获取方式。

2.1 Medium和Twitter

截至2018年, Twitter共有3.36亿个活跃用户,每天会新增约3.4亿条推文。同时,该社交网站也存在大量社交机器人。有研究显示大约有8.5%的账号为机器人账号, Kong等^[27]估计Twitter活跃用

户中有9%~15%为社交机器人。Medium是由Twitter创始人之一——埃文·威廉姆斯开发的、专注于用户阅读与写作的社交网站。该社交网站创始的初衷是鼓励用户创建超过Twitter文本字符上限的文本,即提供给用户一个发表长文的平台。该社交网站可以直接使用Twitter账号登录,并给用户推荐其Twitter好友的Medium账号,因此很多Twitter的用户会同时使用Medium。截至2019年,Medium每月的活跃用户达到0.85亿~1亿个,每月有近1亿个读者访问Medium页面,Medium的用户和Twitter的用户有很大的重复性,并且可以通过用户的Medium账号关联至其Twitter账号。

2.2 Medium和Twitter的社交机器人

由于Twitter与Medium的开放性,两个社交网站的用户中存在许多社交机器人。根据社交机器人的目的,可以将其分为有害的恶性机器人与有益的良性机器人。恶性机器人的主要目的为欺诈或欺骗用户,传播垃圾信息。恶性机器人往往被用于发布广告,宣传具有特定舆论导向的观点和意见,操纵金融趋势,传播垃圾邮件、恶意软件和有害链接,影响搜索引擎的结果进而使特定的链接首先显示,提供一个购买社交媒体粉丝的市场,制造虚假新闻来源等。在社交网站中,恶性机器人给用户和社会环境带来了负面影响。良性机器人的主要活动为向用户提供有益信息或娱乐内容等,如Facebook上为抑郁症患者提供帮助的社交机器人^[29],Github上为新人能够快速入门提供帮助的机器人^[30],Twitter上会定时发布敲钟推文的大本钟机器人账号等。本文将不区分社交机器人是否带有恶意,所有社交机器人都是检测对象。

2.3 Botometer机器人检测工具

Botometer是由美国印第安纳大学OSoMe团队通过一系列工作创建的Twitter机器人检测工具^[31-34]。该检测工具基于用户资料、朋友关系和社交网络等超过1 000个特征,使用了随机森林分类器算法来检测用户是否为社交机器人,通过分析以下6类特征对用户是否为社交机器人进行评断:网络(度分布、聚类系数等)、用户账号信息、好友(关注数量、粉丝数量等)、时间(推文频率等)、推文内容(自然语言处理特征等)和情绪特征。目前,已有多项研究使用该检测工具的检测结果作为预测社交机器人的重要依据。

通过OSoMe团队提供的应用程序编程接口,人们可以通过用户名或用户编号对Twitter用户是否为社交机器人进行预测。Botometer会自动抓取Twitter用户的个人账号信息、历史发表推文、转发和引用等信息对用户身份进行检测,返回用户在多种方面可能为社交机器人的分数。

2.4 数据获取与账号标签

本节主要描述了如何从Medium的创始人埃文·威廉姆斯开始,获取Medium与Twitter的用户数据以及检测用户是否为社交机器人的过程。数据获取流程如图1所示。

为获得比较全面的用户列表,首先需要爬取Medium用户的ID,由埃文·威廉姆斯开始,通过他的关注用户列表进行广度优先搜索,爬取用户ID,共爬取了2 500万个用户ID。然后随机选择了35万个用户,爬取他们的Medium账号信息。通过对Medium爬虫,可以爬取用户公布的账号资料信息、关注用户列表、粉丝用户列表,以及用户的历

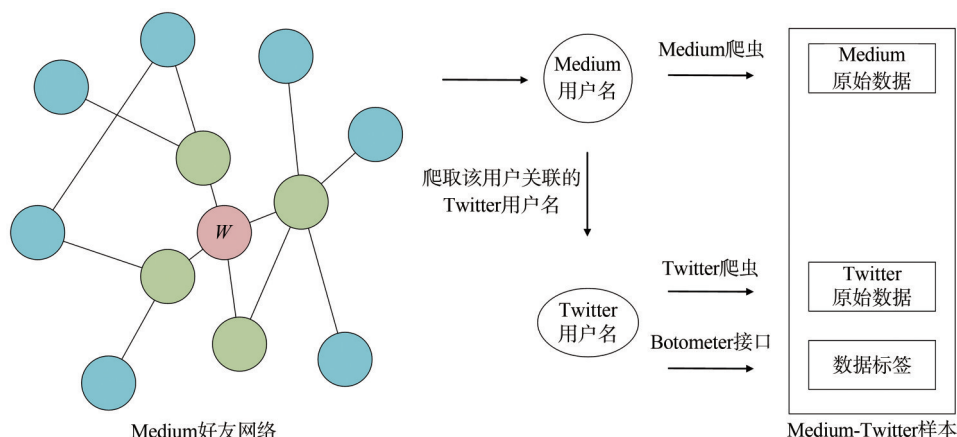


图1 Medium-Twitter用户数据获取流程

史点赞、历史高亮、历史评论和历史发布4类历史活动信息。筛选出无上述4类历史活动的Medium用户,共194 839个用户,其中有124 083个用户链接了Twitter账号。

根据Medium用户链接的Twitter账号,通过Twitter应用程序编程接口对用户信息爬取,并使用Botometer应用程序编程接口得到用户是否为机器人的可能性分数(0~5分),分数越大说明该用户越有可能为社交机器人。本文将分数大于4分的用户账号标记为社交机器人账号,反之则标记为普通用户账号。去掉无效的或者链接有问题的用户账号后,共有91 713对可用的Medium-Twitter账号。

其中,被Botometer判断为机器人的Twitter账号有4 129个,构成了实验数据集的正样本。负样本为另外随机选择的4 129个合法Twitter账号,并爬取了这些账号所链接的Medium账号数据,构成实验用的训练集与测试集。

3 系统设计与构建

本节提出了跨网站社交机器人检测系统Cer-

berus, Cerberus使用用户在Medium与Twitter的特征识别该用户在Twitter是否为社交机器人。

3.1 系统概述

Cerberus系统设计如图2所示。整个系统由自然语言处理模块、描述性特征处理模块和决策模块3个部分组成。自然语言处理模块负责处理用户在Medium产生的文本内容并输出自然语言处理特征。该模块输入的文本共有4类,分别为历史发布、历史点赞、历史高亮和历史评论。其中历史发布为用户自己发表的文章,历史点赞为用户点赞过的文章,历史高亮为用户高亮过的文章,历史评论为用户回复过的文章。描述性特征处理模块负责数据统计与处理,将用户简介中的内容特征化,并统计上述4类文本的数据,最后输出描述性特征。决策模块从自然语言处理模块与描述性特征处理模块获得其输出的特征,并将这些特征输入监督式学习模型作二分类,来预测该用户是否为社交机器人。

3.2 自然语言处理模块

用户在Medium上会对其阅读过的文章进行互动,如点赞、高亮或评论。同样地,用户也可能会

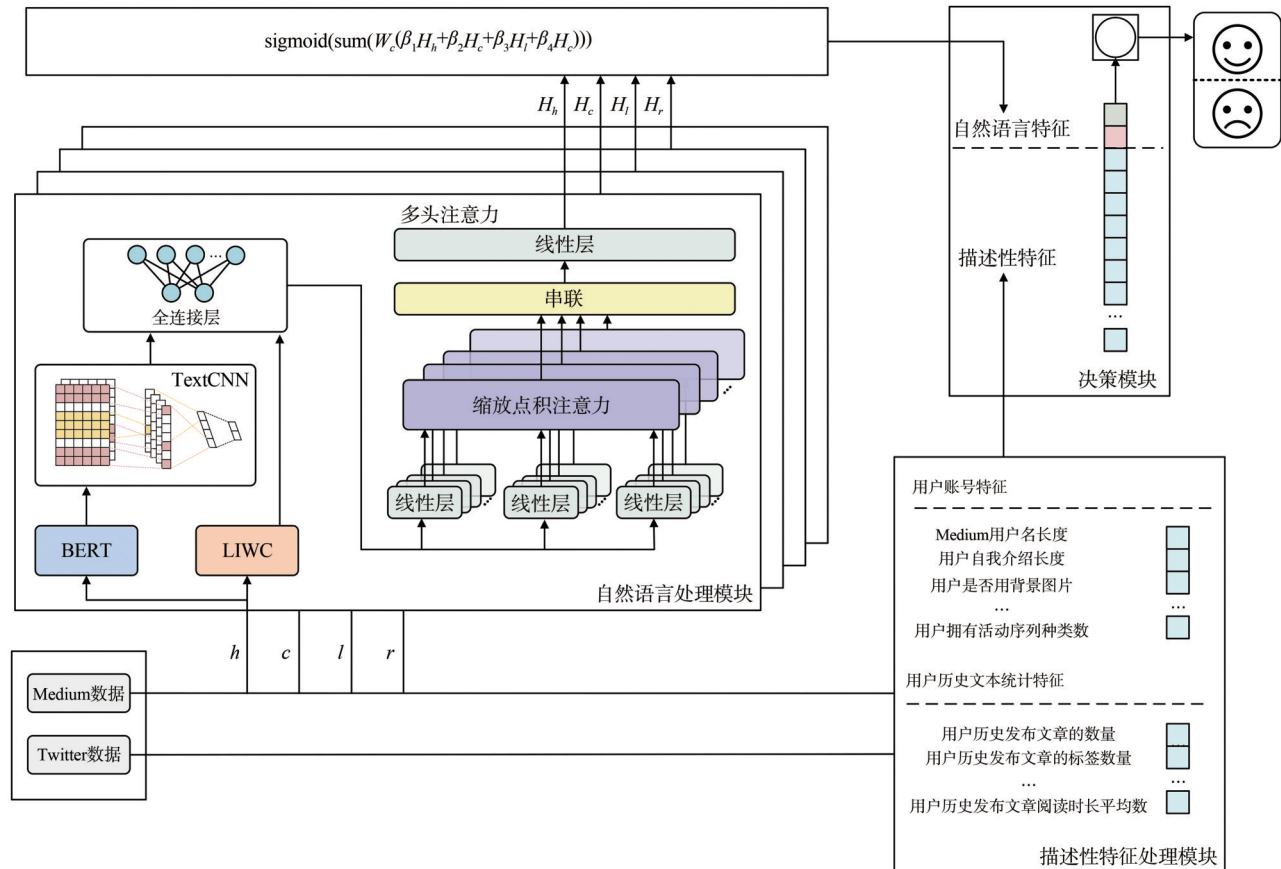


图2 Cerberus系统设计

发布自己的文章, 这些文章可以在该用户的历史发布中找到。自然语言处理模块学习用户产生的不同类别的文本语言特征, 利用该特征对用户的身份加以分析。

自然语言处理模块对历史发布、历史高亮和历史点赞取其标题, 对历史评论取其文本内容, 分别得到某一用户 $\{h_1, h_2, \dots\}$ 、 $\{c_1, c_2, \dots\}$ 、 $\{l_1, l_2, \dots\}$ 、 $\{r_1, r_2, \dots\}$ 4 种类型的文本内容。对于每一类文本内容, Cerberus 会考虑其在语言研究与词汇计数 (linguistic inquiry and word count, LIWC) 方面的指标, 帮助用户了解文本作者的思维、情感、个性和与人交流的方式, 随后通过多标签二值化函数将其编码。LIWC 的输出将与文本卷积神经网络 (text convolutional neural network, TextCNN) [35] 的输出一起被送入多头注意力层, 从而获取该类文本内容的特征矩阵。最后, 该模块综合各类文本内容输出的矩阵输出该用户在自然语言处理模块的特征。

下文均以历史高亮类文本为例, $h_m = [h_m^1, h_m^2, \dots, h_m^q, \dots, h_m^Q]$, 其中, h_m 为第 m 个历史高亮的文章标题, h_m^q 为该标题中第 q 个单词, Q 为该标题的所有单词数, h_m^Q 为该标题中最后一个单词。为了将文本编码为词嵌入向量, Cerberus 引入了基于变换器的双向编码器表示 (bidirectional encoder representations from transformer, BERT) 预先训练的模型, 使文本的稀疏词序列转换为稠密的词嵌入向量。将 h_m^q 输入 BERT 可以得到该单词对应的词向量 e_m^q 。将所有的 h_m^q 通过该词嵌入层处理后可以得到对应的 $e_m = [e_m^1, e_m^2, \dots, e_m^q, \dots, e_m^Q]$ 。之后将其输入卷积神经网络层 [36]。卷积神经网络能有效地捕获局部信息, 很好地提取文本的潜在特征。它以 e_m 作为输入, 输出提示文本 h_m 的潜在特征 r_m , 最终对每一条文本输出一个一维向量。同时, 使用 LIWC 处理 h_m , 并使用多标签二值化函数将其编码为描述性特征 s_m 。随后, 将卷积神经网络提取的文本潜在特征 r_m 和描述性特征 s_m 合并, 并将其输入一个全连接神经网络, 输出记为 n_m 。

受 Transformer [37] 的启发, 本文引入了多头注意力机制, 该注意力机制可以帮助 Cerberus 更进一步地将注意力集中在预测社交网站用户是否为社交机器人的关键信息上, 并减少对无关信息的关注。多头注意力机制由多个独立的缩放点积注意力组

成, 每个缩放点积注意力都被称作一个“头”。如式 (1)、式 (2) 所示, 将 n_m 作为每一个“头”的输入, 从而进一步提取关键信息。其中 Q_i, K_i, V_i 均为将 n_m 映射到不同空间的线性变换矩阵, 分别代表查询、键与值。

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) \quad (1)$$

$$Q_i = n_m W_i^Q, K_i = n_m W_i^K, V_i = n_m W_i^V \quad (2)$$

在式 (1) 中使用到的缩放点积注意力如式 (3) 所示。该机制为了保证模型在不同参数与不同数据规模下的鲁棒性, 引入了向量 d_k 用于缩放与保持模型稳定。

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (3)$$

使用多个缩放点积注意力可以让每个“头”将注意力放在不同的重点上, 捕捉更多对结果产生影响的部分。多头注意力机制将每个“头”的输出合并, 并与线性变换矩阵 W^O 相乘, 后取其结果并输出 t_m , 如式 (4) 所示。

$$t_m = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_n) W^O \quad (4)$$

随后, 将 t_m 累加得到表示该用户历史高亮文本特征的矩阵 H_h , 如式 (5) 所示, 其中 M 为历史高亮文本标题总数。

$$H_h = \sum_{m=1}^M t_m \quad (5)$$

该用户的 4 类文本经过上述处理后可以分别得到 H_h, H_c, H_l, H_r , 分别表示历史高亮、历史点赞、历史发布与历史评论文本对应的输出。对每个类别的矩阵赋以相应的权重 $\beta_1, \beta_2, \beta_3, \beta_4$ 后, 得到本模块输出的自然语言特征 f_{NLP} , 如式 (6) 所示。

$$f_{\text{NLP}} = \text{sigmoid}\left(\text{sum}\left(W_c(\beta_1 H_h + \beta_2 H_c + \beta_3 H_l + \beta_4 H_r)\right)\right) \quad (6)$$

其中, W_c 表示线性变换矩阵。

3.3 描述性特征处理模块

描述性特征处理模块主要负责提取用户在两个社交网站简介中的字段和前文提到的 4 种文本内容的统计特征, 分别在后文被称为账号描述性特征与文本统计特征, 并将其处理为描述性特征矩阵供决策模块使用。

账号描述性特征主要来源于从用户的个人资料中提取的数值型特征。社交机器人和普通用户在一些统计信息上存在显著区别。由图 3 可以看出, 相

对于普通用户, 社交机器人的粉丝数、关注数以及好友数往往较少, 且由于社交机器人对自我账号设置的不重视, 其自我介绍的长度往往较短并且在用户昵称中包含较多的数字。此外, 社交机器人账号涉及的活动行为种类普遍较少, 有862个社交机器人只有一种活动行为, 230个普通用户只有一种活动行为。有4种活动行为的社交机器人只有53个, 而普通用户则有640个。这个数据的差异说明了这些社交机器人往往不能使用Medium提供的所有服务功能, 而是仅使用部分功能进行活动。

与偏向于文本内容的自然语言处理模块不同, 描述性特征处理模块主要关注历史发布、历史高亮、历史点赞和历史评论4类文本的统计特征。

如图4所示, 社交机器人和普通用户在分享文章方面同样有显著的差异, 社交机器人往往发布较多的文章, 较少高亮、回复、点赞文章, 相较于普通用户, 有更多的社交机器人从未使用过高亮和回复功能, 和之前的数据显示结果一致, 社交机器人往往不能利用Medium提供的所有服务功能。同时由图5可以看出社交机器人发布的文章的获赞数和阅读时长低于普通用户发布的文章, 而社交机器人发布的文章包含的标签数多于普通用户账号发表的文章。这可能是社交机器人所发表的文章质量较差导致其获赞数较少和被阅读时长较短, 但是他们希望通过标签吸引更多的阅读用户, 因此标签数较多。同时, 社交机器人点赞的文章包含的标签数

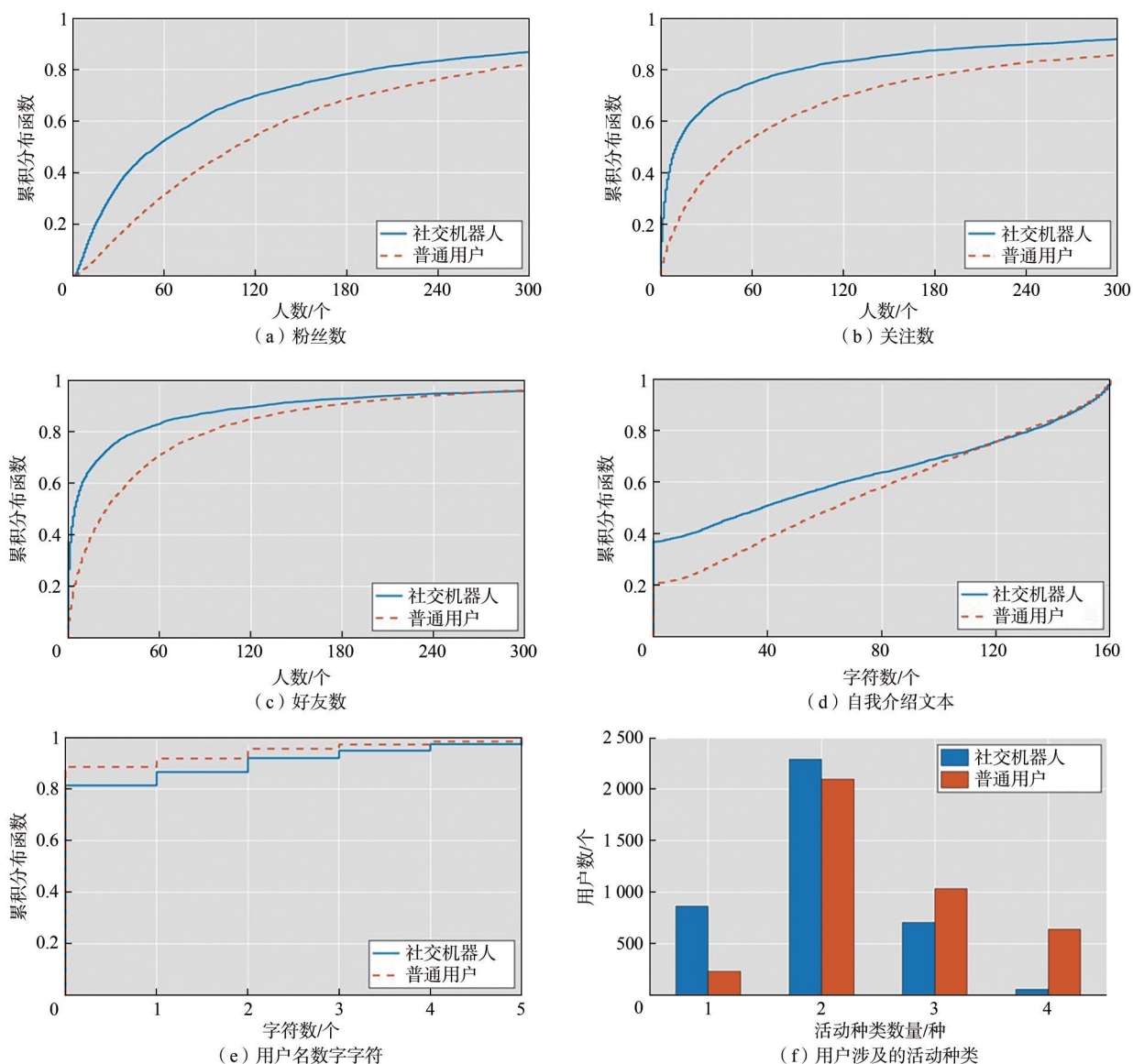


图3 社交机器人和普通用户在账号描述特征方面的差异

多于普通用户点赞文章包含的标签数。这可能是由于普通用户点赞文章往往带有目的, 所关注的主题

更专一, 因此标签数较少。

该模块对用户提取的所有特征见表1, 主要分为

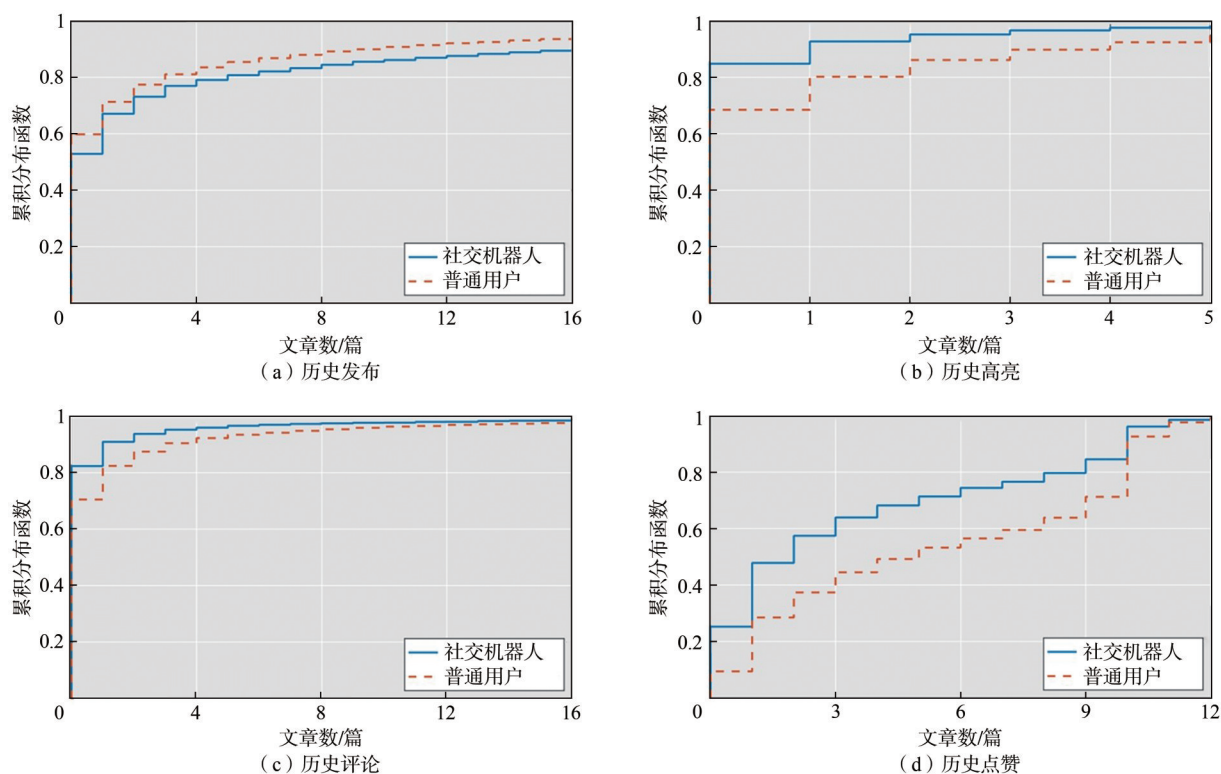


图4 社交机器人和普通用户在分享文章方面的差异

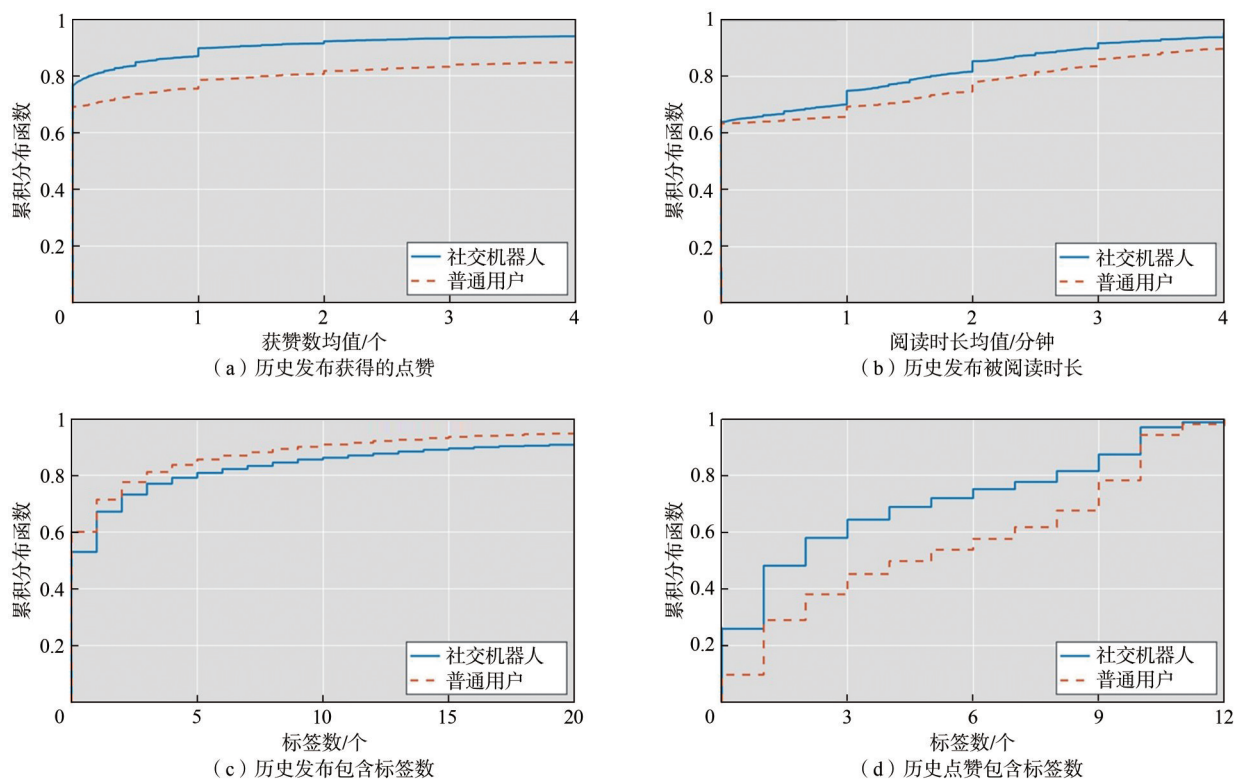


图5 社交机器人和普通用户在发布文章与阅读文章时行为方面的差异

以下两个部分。第一个部分为账号描述性特征，从用户在Medium与Twitter中的简介中提取。第二个部分为文本统计特征，通过在用户历史的活动记录及分享文本行为中提取。

3.4 决策模块

Cerberus将自然语言处理模块与描述性特征处理模块输出的特征进行合并，然后将其输入决策模块，以预测用户是否为社交机器人。决策模块主要使用监督式学习算法作二分类。笔者使用的模型包括XGBoost^[38]分类器、CatBoost^[39]分类器、随机森林^[40]分类器、多层感知机（multilayer perceptron, MLP）分类器和LightGBM^[41]分类器。本文针对每个模型使用网格搜索确定其超参数。在对每个模型进行超参数验证后，选取性能最优的模型作为Cerberus决策模块的分类器。至此，Cerberus能够较准确地预测一个用户是否为社交机器人。

4 实验结果

本文实现了Cerberus，并使用前文收集到的数据评估了其性能。

4.1 实验设置

数据集中共有8 258位用户，数据集中正负样本比为1:1。在保证分割后正负样本比例不变的情况下，将数据集以7:3的比例划分为训练集和测试集。对每个分类器，在测试集上，对其网格搜索中的每种参数组合都使用5折交叉验证以获得超参数。

为实现自然语言处理模块，本文选择了被广泛使用的机器学习框架PyTorch。Cerberus在BERT中使用了预训练好的模型Bert-base-uncased^[42]。在

TextCNN中，采用了4组卷积核进行卷积操作，每组卷积核包含100个，尺寸分别为3、4、5。对于多头注意力机制，其“头”数被设置为8。在训练过程中，笔者使用了Adam^[43]作为优化器，其参数Learning_rate被设置为0.0003。决策模块使用的监督式学习模型来源于Python的开源库：CatBoost、LightGBM、Scikit-learn^[44]和XGBoost。

本文将标记为社交机器人的样本记为阳性，将标记为人类的样本记为阴性。TP和FN代表标记为社交机器人，但被模型分别预测为社交机器人和人类的数量，FP和TN代表标记为人类，但被模型分别预测为社交机器人和人类的数量。实验使用下述指标来评估模型性能。

(1) 精准率

精准率代表在被预测为社交机器人的样本中实际为社交机器人的比例，如式（7）所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

(2) 召回率

召回率代表在实际为社交机器人的样本中被预测为社交机器人的比例，如式（8）所示。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

(3) F1分数

F1分数是精准率与召回率的调和平均数，如式（9）所示。

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

(4) 受试者工作特征曲线下面积（area under curve, AUC）

AUC代表模型对随机选取的社交机器人样本

表1 描述性特征

特征类型	特征名称
账号描述性特征	Medium用户名/Twitter用户名/Medium昵称长度
	用户自我介绍长度
	Medium用户名/Twitter用户名/Medium昵称中数字数量
	用户是否有背景图片
	用户是否有Facebook账号
	用户粉丝数/关注人数
	用户好友(同时为关注者和粉丝)数
文本统计特征	用户涉及的活动种类数
	用户历史发布、历史高亮、历史点赞和历史评论类型文章的数量
	用户历史发布、历史点赞和历史评论类型文章的标签数量
	用户历史发布、历史高亮、历史点赞和历史评论类型文章的阅读时长平均数/方差
	用户历史发布、历史点赞和历史评论类型文章的点赞数量平均数/方差
	用户历史发布、历史高亮、历史点赞和历史评论类型文章的回复数量平均数/方差
	用户历史发布、历史高亮、历史点赞和历史评论类型文章的发文间隔平均数/方差

的预测打分比随机选取的普通用户高的概率。

4.2 决策模块算法间的比较

Cerberus 中的决策模块可以支持各种二分类器算法, 实验分别尝试了 CatBoost 算法、XGBoost 算法、LightGBM 算法、随机森林算法和 MLP 算法, 各算法的超参数均由网格搜索获得。各模型的参数与实验结果均在表 2 中列出。从实验结果可以看到 CatBoost 算法在 F1-score 与 AUC 两项指标上均取得了最好的成绩, 分别为 0.6833 与 0.7552, 本文将选取 CatBoost 算法作为 Cerberus 在决策模块的分类器。

表 2 Cerberus 系统实验结果

算法	参数设置	Precision	Recall	F1-score	AUC
CatBoost	depth=5	0.6981	0.6692	0.6833	0.7552
	iterations=100 learning_rate=0.1				
XGBoost	learning_rate=0.1 n_estimators=50	0.6991	0.6667	0.6825	0.7549
LightGBM	learning_rate=0.005 random_state=500	0.6843	0.6451	0.6641	0.7337
	reg_alpha=1.2 reg_lambda=0.4				
随机森林	criterion="entropy" max_depth=2	0.6756	0.6078	0.6399	0.7151
	min_samples_leaf=1 min_samples_split=2 n_estimators=300 random_state=6				
MLP	hidden_layer_sizes=[100]*6 learning_rate="adaptive" learning_rate_init=0.0001 max_iter=1 000	0.6419	0.4934	0.5579	0.6452

4.3 特征重要性分析

为探索自然语言处理模块与描述性特征处理模块产出的特征在预测某一用户是否为社交机器人时起到的作用, 采用卡方统计^[45], 对特征的重要性进行定量分析。表 3 展示了卡方值前 10 的特征及其卡方值, 可以看到两模块所产出的特征均可以在预测中提供较关键的信息。

4.4 与其他研究成果的比较

本文将 Cerberus 与如下社交机器人检测系统作对比。

USMB^[46]: 通过分析用户发表的内容, 提取与自然语言处理相关的特征, 如发布内容中不同词性的单词的数量、大小写字母的数量与发布的内容中是否包含链接或标签等特征。基于这些特征, 使用决策树预测某一内容是否由社交机器人发布。通过

表 3 卡方统计特征重要性

排序	特征名称	χ^2 值
1	用户涉及的活动种类数	577.761
2	自然语言处理模块将用户预测为社交机器人的概率	526.521
3	自然语言处理模块将用户预测为普通用户的概率	526.521
4	用户点赞文章数	421.220
5	用户点赞文章包含的平均标签数	405.354
6	用户高亮文章发文间隔方差	275.893
7	用户点赞文章平均阅读时长	210.391
8	用户链接 Twitter 名长度	163.605
9	用户链接 Twitter 名中数字个数	123.910
10	用户名长度	116.154

对每位用户所有发文的特征进行均值计算得到该用户的整体特征, 并据此判断该用户是否为社交机器人。

Stweeler^[47]: 提取分析用户的账号特征, 如账号的关注者数量、粉丝数量、发文习惯、平均发文间隔等特征, 使用随机森林算法对提取的特征进行分析, 判断用户身份。

比较结果呈现在表 4 中, 基于 F1-score 和 AUC 的评估指标, Cerberus 系统的性能相较于其他系统表现最优。说明用户的账户信息特征对用户的身份识别有较大的影响, 仅通过用户生成文本很难对用户身份进行准确判断。

表 4 Cerberus 和其他研究成果的对比实验

社交机器人检测系统	Precision	Recall	F1-score	AUC
Cerberus	0.698	0.669	0.683	0.755
USMB	0.542	0.697	0.610	0.581
Stweeler	0.633	0.567	0.594	0.693

5 结束语

本文针对在线社交网络的社交机器人检测问题, 提出并实现了一个跨网站社交机器人检测系统 Cerberus, 旨在精确识别 Medium-Twitter 上的社交机器人。本系统考虑用户在两个社交网站产生的描述性特征与文本内容, 利用深度学习实现社交机器人的有效检测。实验结果表明, 本文所设计的系统在 Twitter-Medium 跨网站真实数据集上取得更高的检测性能, 其中 AUC 值达到了 0.7552。本文提出的跨网站社交机器人检测系统, 通过有效地利用跨网站数据, 为新兴社交网站上社交机器人检测提供了一种有效的解决方案。

本文的研究还存在一些局限性。首先,本文实验过程中使用的社交网站组合比较单一,仅以Medium-Twitter为研究对象,无法评估Cerberus的泛用性。笔者会在将来的工作中将Cerberus应用到更多研究对象,以确定其泛化能力。其次,本文在两个社交网站的特征选择上较有限,仅考虑了文本内容与描述性特征。引入图像、社交关系与节点的中心性^[48]等多模态数据,将有助于进一步丰富用户特征。再次,数据集的样本标签依赖Botometer的检测结果,但本文未对Botometer的检测结果进行校验,在人力资源允许的情况下,可以随机筛选部分Botometer的检测结果进行校验,以验证其检测结果的准确性。最后,本文使用的数据集相对较小,可能存在对不同类型的社交机器人覆盖度不足的问题,增加样本数量,可能会提升Cerberus在真实应用场景下的性能。

参考文献:

- [1] WILSON C, BOE B, SALA A, et al. User interactions in social networks and their implications[C]//Proceedings of the 4th ACM European Conference on Computer Systems. New York: ACM, 2009: 205-218.
- [2] GABIELKOV M, RAO A, LEGOUT A. Studying social networks at scale: macroscopic anatomy of the twitter social graph[EB]. arXiv preprint, 2014, arXiv: 1404.1355.
- [3] 陈苑文, 王晓, 李灵犀, 等. 基于社交媒体数据增强的交通态势感知研究及进展[J]. 智能科学与技术学报, 2022, 4(1): 1-13.
CHEN Y W, WANG X, LI L X, et al. Traffic situational awareness research and development enhanced by social media data: the state of the art and prospects[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(1): 1-13.
- [4] 孟雪, 杨若楠, 李睿琪. 中国现实题材电视剧的海外传播效果研究: 以YouTube平台为例[J]. 智能科学与技术学报, 2023, 5(3): 343-351.
MENG X, YANG R N, LI R Q. Research on the overseas communication effect of Chinese realistic theme TV series: take YouTube platform as an example[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(3): 343-351.
- [5] 陈妍, 罗雪琴, 梁伟, 等. 基于情感信息融合注意力机制的抑郁症识别[J]. 智能科学与技术学报, 2022, 4(4): 600-609.
CHEN Y, LUO X Q, LIANG W, et al. Depression recognition based on emotional information fused with attentional mechanism[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(4): 600-609.
- [6] ZHENG H Z, XUE M H, LU H, et al. Smoke screener or straight shooter: detecting elite sybil attacks in user-review social networks[C]//Proceedings of the 2018 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2018.
- [7] COSTA H, BENEVENUTO F, MERSCHMANN L H C. Detecting tip Spam in location-based social networks[C]//Proceedings of the 28th Annual ACM Symposium on Applied Computing. New York: ACM, 2013: 724-729.
- [8] LI H Y, CHEN Z Y, MUKHERJEE A, et al. Analyzing and detecting opinion Spam on a large-scale dataset via temporal and spatial patterns[C]//Proceedings of the International AAAI Conference on Web and Social Media. Buffalo: AAAI Press, 2021, 9(1): 634-637.
- [9] RAHMAN M, CARBUNAR B, BALLESTEROS J, et al. Turning the tide: curbing deceptive Yelp behaviors[C]//Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2014: 244-252.
- [10] VENKATADRI G, GOGA O, ZHONG C T, et al. Strengthening weak identities through inter-domain trust transfer[C]//Proceedings of the 25th International Conference on World Wide Web. Switzerland: International World Wide Web Conferences Steering Committee, 2016: 1249-1259.
- [11] DE MEO P, FERRARA E, ABEL F, et al. Analyzing user behavior across social sharing environments[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 5(1): 1-31.
- [12] FAZIL M, SAH A K, ABULAISH M. DeepSBD: a deep neural network model with attention mechanism for SocialBot detection[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4211-4223.
- [13] 张艳梅, 黄莹莹, 甘世杰, 等. 基于贝叶斯模型的微博网络水军识别算法研究[J]. 通信学报, 2017, 38(1): 44-53.
ZHANG Y M, HUANG Y Y, GAN S J, et al. Weibo spammers' identification algorithm based on Bayesian model[J]. Journal on Communications, 2017, 38(1): 44-53.
- [14] 夏崇欢, 李华康, 孙国梓. 基于行为特征分析的微博恶意用户识别[J]. 计算机科学, 2018, 45(12): 111-116.
XIA C H, LI H K, SUN G Z. Microblogging malicious user identification based on behavior characteristic analysis[J]. Computer Science, 2018, 45(12): 111-116.
- [15] ARIN E, KUTLU M. Deep learning based social bot detection on twitter[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1763-1772.
- [16] LIU Y H, TAN Z X, WANG H, et al. BotMoE: Twitter bot detection with community-aware mixtures of modal-specific experts[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2023: 485-495.
- [17] 程晓涛, 刘彩霞, 刘树新. 基于关系图特征的微博水军发现方法[J]. 自动化学报, 2015, 41(9): 1533-1541.
CHENG X T, LIU C X, LIU S X. Graph-based features for identifying spammers in microblog networks[J]. Acta Automatica Sinica, 2015, 41(9): 1533-1541.
- [18] LI S D, ZHAO C Y, LI Q, et al. BotFinder: a novel framework for social bots detection in online social networks based on graph embedding and community detection[J]. World Wide Web, 2023, 26(4): 1793-1809.
- [19] LIU F, YANG C F, LI Z Y, et al. Accou2vec: a social bot detection model based on community walk[J]. IEEE Transactions on Dependable and Secure Computing, 2023: 1-17.
- [20] MENDOZA M, TESCONI M, CRESCI S. Bots in social and interaction networks: detection and impact estimation[J]. ACM Transactions on Information Systems (TOIS), 2020, 39(1): 1-32.
- [21] FENG S B, TAN Z X, LI R, et al. Heterogeneity-aware Twitter bot detection with relational graph transformers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2022, 36(4): 3977-3985.
- [22] 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. New York: ACM, 2010: 435-442.
- [23] GONG Q Y, CHEN Y, HU J Y, et al. Understanding cross-site linking in

- online social networks[J]. *ACM Transactions on the Web*, 2018, 12(4): 1-29.
- [24] LIU S Y, WANG S H, ZHU F D, et al. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling[C]//*Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. New York: ACM, 2014: 51-62.
- [25] GOGA O, LEI H, PARTHASARATHI S H K, et al. Exploiting innocuous activity for correlating users across sites[C]//*Proceedings of the 22nd International Conference on World Wide Web*. New York: ACM, 2013: 447-458.
- [26] SANG J T, YAN M, XU C S. Understanding dynamic cross-OSN associations for cold-start recommendation[J]. *IEEE Transactions on Multimedia*, 2018, 20(12): 3439-3451.
- [27] KONG X N, ZHANG J W, YU P S. Inferring anchor links across multiple heterogeneous social networks[C]//*Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. New York: ACM, 2013: 179-188.
- [28] GONG Q Y, CHEN Y, HE X L, et al. Cross-site prediction on social influence for cold-start users in online social networks[J]. *ACM Transactions on the Web*, 2021, 15(2): 1-23.
- [29] KIM T, RUENSUK M, HONG H. In helping a vulnerable bot, you help yourself: designing a social bot as a care-receiver to promote mental health and reduce stigma[C]//*Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2020: 1-13.
- [30] HE H, SU H N, XIAO W X, et al. GFI-bot: automated good first issue recommendation on GitHub[C]//*Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York: ACM, 2022: 1751-1755.
- [31] 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 - WWW '16 Companion*. New York: ACM, 2016: 273-274.
- [32] 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*. New York: ACM, 2020: 2725-2732.
- [33] VAROL O, FERRARA E, DAVIS C, et al. Online human-bot interactions: detection, estimation, and characterization[C]//*Proceedings of the International AAAI Conference on Web and Social Media*, Buffalo: AAAI Press, 2017.
- [34] 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*, Vancouver: AAAI Press, 2020.
- [35] CHEN Y. Convolutional neural network for sentence classification[D]. Waterloo: University of Waterloo, 2015.
- [36] LI A, QIN Z, LIU R S, et al. Spam review detection with graph convolutional networks[C]//*Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York: ACM, 2019: 2703-2711.
- [37] VASWANI A. Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017.
- [38] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016: 785-794.
- [39] PROKHORENKOVA L, GUSEV G, VOROBIEV A, et al. CatBoost: unbiased boosting with categorical features[J]. *Advances in Neural Information Processing Systems*, 2018, 31.
- [40] SVETNIK V, LIAW A, TONG C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. *Journal of Chemical Information and Computer Sciences*, 2003, 43(6): 1947-1958.
- [41] KE G, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[J]. *Advances in Neural Information Processing Systems*, 2017, 30.
- [42] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv: 1810.04805.
- [43] KINGMA D P, BA J, HAMMAD M M. Adam: a method for stochastic optimization[EB]. arXiv preprint, 2014, arXiv: 1412.6980.
- [44] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: machine learning in python[J]. *Journal of Machine Learning Research*, 2011, 12: 2825-2830.
- [45] WELCH B L. On the comparison of several mean values: an alternative approach[J]. *Biometrika*, 1951, 38(3/4): 330-336.
- [46] FERREIRA DOS SANTOS E, CARVALHO D, RUBACK L, et al. Uncovering social media bots: a transparency-focused approach[C]//*Proceedings of the Companion Proceedings of the 2019 World Wide Web Conference*. New York: ACM, 2019: 545-552.
- [47] GONG Q Y, LIU Y S, ZHANG J Y, et al. Detecting malicious accounts in online developer communities using deep learning[C]//*Proceedings of the IEEE Transactions on Knowledge and Data Engineering*. Piscataway: IEEE Press, 2023: 10633-10649.
- [48] 王明宇, 宫庆媛, 瞿晶晶, 等. 基于机器学习的GitHub企业影响力分析与预测[J]. *智能科学与技术学报*, 2023, 5(3): 330-342.
- WANG M Y, GONG Q Y, QU J J, et al. Analysis and prediction of GitHub company influence based on machine learning[J]. *Chinese Journal of Intelligent Science and Technology*, 2023, 5(3): 330-342.

[作者简介]



汤家伟 (1997-), 男, 复旦大学计算机科学技术学院科研助理, 主要研究方向为社交网络、机器学习、用户行为大数据分析。



刘育杉 (1999-), 女, 复旦大学计算机科学技术学院硕士生, 主要研究方向为在线社交网络用户数据挖掘。



高敏 (1997-), 女, 复旦大学计算机科学技术学院博士生, 主要研究方向为社交网络用户行为建模、图深度学习、网络科学等。



宫庆媛（1991- ），女，博士，复旦大学智能复杂体系基础理论与关键技术实验室青年副研究员，主要研究方向为在线社交网络用户行为大数据。



陈阳（1981- ），男，博士，复旦大学计算机科学技术学院教授、博士生导师，上海市智能信息处理重点实验室副主任，主要研究方向为社会计算、计算机网络、大规模用户行为数据挖掘等。



王新（1973- ），男，博士，复旦大学计算机科学技术学院党委书记、教授、博士生导师，主要研究方向为新一代互联网体系结构、无线与移动网络、数据中心网络、社交网络、网络编码的应用。