恶意社交机器人检测方法综述

张 鹏*，秦瑞青，刘润东，兰月新，韦昱妃

（中国人民警察大学 网络舆情治理研究中心，廊坊 065000）

摘要 对 Twitter、Facebook 以及新浪微博等大型在线社交平台上不同类型的社交机器人进行特征分析，围绕社交机器人检测框架，对基于机器学习、深度学习以及其他新兴检测方法的社交机器人检测模型的优缺点和适用性进行总结和分析。研究发现对于不同平台和攻击目标的社交机器人需要提取多种维度的特征并设计相应的检测方法。最后，对如何减少社交机器人的危害以及应对人类与社交机器人共存挑战的措施进行深层次挖掘和分析，并对如何提高识别精度以及热点技术的发展进行了讨论和展望。

关键词 社交机器人；机器学习；深度学习；舆论机器人；舆论操纵；恶意社交机器人

中图分类号 TP391 **文献标志码** A DOI 10.12178/1001-0548.2023229

Review of Detection Methods for Malicious Social Robots

ZHANG Peng*, QIN Ruiqing, LIU Rundong, LAN Yuexin, and WEI Yufei

(Research Center for Network Public Opinion Governance, China People's Police University, Langfang 065000, China)

Abstract The characteristics of different types of social robots on large online social platforms such as Twitter, Facebook and Sina Weibo are reviewed in this paper. Based on the social robot detection framework, the advantages and disadvantages and applicability of social robot detection models based on machine learning, deep learning and other emerging detection methods are summarized and analyzed. It is found that social robots with different platforms and attack targets need to extract multi-dimensional features and design corresponding detection methods. Finally, this paper deeply explores and analyzes how to reduce the harm of social robots and measures to cope with the challenges of coexistence between human and social robots, and discusses and looks forward to how to improve the recognition accuracy and the development of hot technologies.

Key words social robot; machine learning; deep learning; public opinion robot; public opinion manipulation; malicious social robot

在线社交网络的快速发展对社会影响与日俱增，近年来，在 Twitter、Facebook、Ins 和新浪微博等社交媒体平台上出现了大量自动生成内容并与人类互动的社交机器人。社交机器人是一类计算机程序，能够自动注册、创建社交账号、生成推文并组织社交机器人集群统一进行评论、转发、点赞等行为。社交机器人根据其用途和产生影响可分为两类：1）以服务人类、提升人类生活质量为目的，如 2022 冬奥会自动生成赛事新闻文本的机器人，自动播报最新的气象数据、地震信息的社交机器人[1]；2）以恶意散播虚假信息、控制舆论为目标。恶意社交机器人的威胁主要有两方面：1）它们渗入社交网络向目标用户或大众投放垃圾邮件、发布虚假新闻和谣言，使得社交媒体虚假信息泛滥，严重干扰用户的正常生活；2）它们自动生成并发布大量推文改变社交网络原有的信息传播规律，从而在政治事件中影响公众观点进而操纵社交媒体舆论走向，成为扰乱政治选举、操纵舆论走向的工具。因此，恶意社交机器人的检测已经成为保护公民信息安全、维持互联网秩序的一项重要任务。

现有的对于恶意社交机器人的检测研究按照使用方法的不同大体可分为基于机器学习、基于深度学习以及其他新兴的检测方法。社交机器人的检测

收稿日期：2023−09−11；修回日期：2024−06−05

基金项目：教育部人文社会科学研究规划基金（22YJA860012）；警察大学科研重点专项（ZDZX202201）

作者简介：张鹏，博士，副教授，主要从事舆论安全与智能传播方面的研究。

*通信作者 E-mail: zhangpeng@cppu.edu.cn

任务涉及自然语言处理、机器学习、数据挖掘等多个领域[2]，检测方法的核心在于从海量数据中挖掘社交机器人区别于正常用户的行为模式特征，从而实现准确识别和防范。早期的检测研究多使用机器学习算法，侧重于对特征提取的改进，依赖于数据标注的可靠性、特征选取的准确性，可靠的数据集以及特征工程是机器学习检测方案优化的重要前提。如引入情感特征[3]、文本内容特征[4-5]、时序特征[6]、网络特征[7] 等分别训练机器学习模型进行检测，在一定程度上提高了模型准确率，达到了较优的检测水平。然而，随着在线社交平台数据量暴增与社交机器人的不断进化，传统机器学习检测方法已无法满足对于新型社交机器人的检测需求。深度学习方法能够很好地处理海量数据并自动学习特征，在泛化性、适应性和可扩展性等方面优于其他机器学习算法，近年来逐渐成为研究热点。如使用图神经网络提取社交网络图特征[8-9] 进行检测，使用 CNN 和 RNN 等神经网络模型自动提取文本特征[10-13]，在一定程度上提高了检测效率和准确率，避免了手工特征提取。此外还有蜜罐实验[14-18]、运用数字 DNA 表征社交指纹[19-20] 等新兴的检测方法，对于检测与时俱进的新型恶意社交机器人、提升社交机器人检测准确率具有重要意义。

本文首先围绕社交机器人检测框架，对其特征工程进行概述，介绍了社交机器人检测的通用研究思路；接着对国内外近年来基于机器学习和深度学习的社交机器人检测方法研究成果与发展进行回顾与梳理，并对新兴的检测方法与未来研究方向进行总结展望；最后从人类与社交机器人共存的挑战、社交机器人的深层次挖掘、检测方法迁移能力的改进、误检问题的解决这 4 个方面探讨了关于恶意社交机器人检测所面临的现实痛点和技术问题的解决路径，以期为未来研究提供借鉴参考。

1　社交机器人检测概述

1.1　社交机器人特征选择

特征选取是检测研究的关键模块，只有选择恰当的、可量化的、区分度明显的特征作为检测属性，才能为后续提高识别准确率、缩短模型时间开销奠定基础。根据目前已有研究，将检测特征分为用户元数据特征、情感特征、文本内容特征、交互行为特征、用户时序特征和网络特征 6 个类别进行介绍，特征类型及其具体内容如表 1 所示。

表 1　社交机器人特征

特征类型	具体特征
用户元数据特征	用户名/ID、个人资料描述、个人资料图像、位置、验证状态、账户年龄、粉丝数量、关注者数量、发文/转发/回复/提及数
情感特征	推文的正面/负面词语数、情感极化得分、唤醒分数、正负分值比、总体推文的情绪指数、情感得分标准差
文本内容特征	文本长度、语义相似性、情感评分、辱骂/攻击性词语数量、@人员分布、参与话题情况、URL链接数量
交互行为特征	发文/转发/点赞/私信等交互情况、发文频率、点击行为
用户时序特征	发文时序，连续发文/转发的时间差、连续提及的时间差、连续评论时间差
网络特征	节点/边的数量、度分布、连边分析、网络密度、图的强度、聚类系数、社区结构、社区成员、访客进出概率、最短路径分布特征、连通图大小分布特征、稀疏分布、度的相关性、幂率分布特征

1）基于用户元数据特征识别。在早期研究中大多倾向于选择较易获取的账号属性作为检测特征。但随着社交机器人的进化，用户元数据特征不断被伪造，其操纵者可以通过修改定位、补充账号资料、购买假粉丝等行为逃避检测，如果仅用上述特征构造检测模型，则会导致识别准确率过低。

2）基于情感特征识别。社交机器人生成的文本内容语言生硬、情感单一且原创内容较少，这些特征有助于提高模型检测的准确率，因此引入情感特征。文献 [3] 分析印度大选相关数据集发现，用户发布的带有情绪的推文比例与其是否为社交机器人之间存在显著关系，情感特征是识别机器人的关键。

3）基于文本内容特征识别。文献 [4] 研究发现，由于社交机器人背后操纵者目的一致导致其发布的推文内容高度相似，因此需要加入文本相似性特征进行检测。此外，部分社交机器人出于控制政治议题走向的目的大量发送带有攻击、辱骂、煽动性词语的推文[5]，因此可以利用这些恶意、负面的文本内容提取检测特征。

4）基于交互行为特征识别。大部分社交机器人会自动进行发布、评论、转发、点赞等活动，因此研究者提出针对用户的交互行为及反应的行为特点设计检测特征。

5）基于时序特征识别。有研究发现，社交机

器人在社交网络上开展攻击行动时，大多选择在异于大多数当地人的作息时间里进行，并且会经过固定的时间间隔进行评论、点赞、转发等行为[6]。因此研究人员提出提取用户时序特征以及关系演化特征，在一定程度上提升识别准确率。

6）基于网络特征识别。社交平台以用户为节点，由关注、评论、转发等交互行为形成连边，构成社交网络图。社交机器人通常会组织对正常账号进行大规模袭击，并关注大批正常用户以进行伪装，而正常用户并不会回应，因此社交机器人的社交网络会呈现与正常用户差异较大的图结构特征[7]。

1.2 社交机器人检测基本框架

关于社交机器人检测方法的研究有很多，其中机器学习和深度学习是主流方法。此外还有基于 URL 链接和蜜罐技术的检测方法，本文分别对几种常用的模型和新兴的检测方法进行介绍，以期为改进社交机器人检测模型的精度提供参考。基于机器学习和深度学习的社交机器人检测通常包括以下 4 个步骤。

1）数据获取：从 Facebook、Twitter、新浪等在线社交平台上采集数据。基于在线社交网络的数据集可分为两种，一种是根据社交网络结构构建的数据集，称为基于图的数据集；另一种使用社交网络提供的开发 API 接口提取的特征创建的数据集，称为非图数据集。

2）数据预处理：从社交网络中收集的数据必须使用预处理技术进行清洗，然后才能输入到机器学习或深度学习模型。文本预处理通常包括去停用词、分词与词性标记、归一化和噪声去除等步骤。

3）特征工程：特征处理部分包括特征选择和特征提取，是数据降维的一部分。通过筛除区分度不大的特征降低模型复杂度，使模型泛化性增强，减少过拟合。一些常用的特征选择方法包括互信息（Mutual Information）、卡方检验（Chi-squared Test）、粒子搜索优化（Particle Swarm Optimization, PSO）、蚁群优化（Ant Colony Optimization, ACO）、鲸鱼优化算法（Whale Optimization Algorithm, WOA）等。特征提取将已有特征映射到低维空间减少计算成本并提高分类模型准确度。常用的特征提取方法包括主成分分析（Principal Components Analysis, PCA）、词频/逆文档频率 TF-IDF、Word2Vec 以及词袋模型 Bow 等。

4）模型训练与测试：利用选定的特征训练机器学习分类器或神经网络模型，并在测试集上进行测试，计算得到混淆矩阵显示的 TP、TN、FP、FN 等信息，同时根据混淆矩阵对准确率、召回率、精确率、F1 值、ROC 曲线和 AUC 值等性能指标进行评价[21]。

2 基于机器学习的社交机器人检测方法

机器学习模型将社交机器人检测视为二分类问题，通过学习大量社交账号特征来识别社交机器人。根据是否依赖带有标签的数据集可以将算法分为有监督学习、半监督学习、无监督学习以及强化学习 4 种方式[22]。有监督学习方法依赖于带有标签的数据集，侧重于提取有助于区分人类和机器人账户的各种特征。无监督方法使用不带有标签的数据集进行学习，能够在不依赖标记数据的情况下检测社交机器人，因此相对于有监督方法能够避免标注数据集耗费大量人力、人工识别社交机器人账号准确率较差、训练集跨社交网络平台迁移能力差等问题。聚类算法是无监督学习方法的一个典型例子，它通过度量不同账户之间特定属性的相似度距离从而发现社交机器人社群。机器学习在社交机器人检测领域的主要应用是分类算法，即预测社交网络账号是否为社交机器人。分类的过程包括两个阶段：1）使用训练集的数据训练分类器；2）使用训练好的分类器进行预测。

2.1 随机森林

随机森林是一种由多棵决策树组成的并行式集成算法，在众多基于机器学习的检测方法中，随机森林因其易于实现、准确率较高、开销较小等优点成为应用范围最广的分类器。集成学习是指组合多个弱学习器综合各模型表现优异的方面以期得到一个更全面且强大的模型，提升整体模型的泛化能力，主要有两种集成方式：1）Boosting，常见的有 Adaboost、GBDT、XgBoost，它的特点是各弱学习器之间有依赖关系，通过串行方式执行，后面的学习器会根据前一个学习器的误差调整参数；2）Bagging，如随机森林，特点是各弱学习器之间不存在依赖关系，能够并行拟合。随机森林的核心思想是先随机构建若干不同且相互独立的决策树以初始化模型，当输入样本进行分类时，首先训练每个决策树得到预测类别，最后以投票的方式得到组合预测类别作为样本最终的预测结果。

文献 [23] 提出可检测单个 Twitter 账户属性的接口 BotOrNot（也称 Botometer），使用随机森林

方法，从交互行为和发布内容中提取好友关系、社交网络、情感特征共 1 150 种特征来计算一个 Twitter 账户属于社交机器人账号的概率。当上传一个账号昵称或 ID 时，系统会收集该账号的个人资料以及发布的推文等信息，利用训练好的模型对该账号进行打分，准确率可达到 95%。文献 [3] 认为社交机器人与正常用户所发推文的情感特征具有较大区别，因此引入基于情感分析的特征检测 Twitter 社交机器人，使模型的 AUC 值从 0.65 提升至 0.73。文献 [24] 从新浪微博爬取关于 30 116 个用户超 1 600 万条微博的数据，并手动标记垃圾邮件发送者用户数据集，识别准确率高达 99.1%。由此可见随机森林的并行集成模式能够有效提升模型的精度，且泛化性较高，不易陷入过拟合。

随机森林算法的优点包括：适合处理数据量较大的数据集，具有较高的准确性和鲁棒性；不易过拟合，具有高度的泛化能力；能够处理输入变量中的缺失值和异常值；可以并行处理，适用于大规模分布式计算等。缺点有：由于随机森林需要训练多棵决策树，因此在处理大量数据时训练时间较长；由于随机森林由多棵决策树组成，每个决策树的结果又包含若干特征，因此每个特征对最终分类结果的影响可解释性较差；对于样本不均衡的数据集易产生偏差，导致少数类别的分类效果较差。总之，随机森林是一类优秀的分类和回归算法，在处理大量数据、特征选择和数据预处理等方面具有很好的效果，但应用时需要结合具体的数据集进行选择。

2.2　贝叶斯网络

朴素贝叶斯是一种应用贝叶斯定理的分类器，基于所有特征相互独立的假设，在给定一组特征的情况下选择出具有最大概率的类。文献 [25] 基于朴素贝叶斯方法使用基于内容和图的特征识别垃圾邮件机器人，并与决策树、SVM 等模型进行对比实验，准确率高达 91.7%，远高于其他模型。文献 [26] 使用朴素贝叶斯分类器根据用户发推文的时间间隔分布来区分个人账户、托管账户和社交机器人，研究结果表明该特征分布在这 3 类账号之间差异较为明显，具有很好的区分能力。

虽然朴素贝叶斯模型原理简单容易实现，但在实际应用时具有以下局限性：1）朴素贝叶斯模型假设模型特征之间相互独立，但在实际应用时该假设通常不成立，因此当使用特征之间存在依赖关系的数据集做分类时可能会出现错误的分类结果；2）当训练样本数量较少时，会导致模型的分类精度下降，而当样本数量增多时，模型的计算复杂度也会相应增加，因此朴素贝叶斯模型对数据量的要求较高；3）无法有效处理缺失数据，如果数据中存在缺失数据，模型可能无法对其正确分类。总之，朴素贝叶斯是一种较为基础的易于实现的分类方法，在社交机器人检测中通常被用作一类基线模型与其他分类模型作对比。

2.3　支持向量机

支持向量机 (Support Vector Machine, SVM) 采用监督学习的方式，是一种线性判别模型，其基本思想是在特征空间中寻找间隔距离最大的分割超平面以划分数据，多用于解决分类、回归等问题[27]。

当训练样本线性可分时，通过硬间隔最大化学习一个线性分类器，即线性可分 SVM；当训练数据近似线性可分时，引入松弛变量，通过软间隔最大化学习一个线性 SVM；当训练数据不可分时，利用核函数和软间隔最大化的方式，学习非线性 SVM。

文献 [28] 提出了一种基于 SVM 和神经网络的 SVM-NN 算法，该算法能够在使用少量特征的前提下进行检测，最终得到 98% 的准确率。文献 [29] 以提供转发推文服务的机器人为研究对象，使用多种先进算法进行检测，实验结果表明 SVM 准确率最高，优于逻辑回归、决策树等模型。然而由于 SVM 算法在使用二次规划求解支持向量时涉及 n 阶矩阵计算，当矩阵阶数 n 很大时将会消耗大量的内存容量和计算时间，因此 SVM 算法处理大规模训练样本时难度较大，只适用于小样本情况下的检测任务。

2.4　基于其他机器学习方法的社交机器人检测

使用机器学习方法提高模型的检测准确率关键在于对特征的改进，因此许多研究者使用的研究思路是提取新的特征加入模型训练，同时使用多种机器学习方法训练分类器进行比较以得出分类效果最好的模型，并进行消融实验以证明所选特征对于模型检测效果提升的有效性。

文献 [3] 采用包括 SVM、贝叶斯网络、Adaboost、梯度提升算法、随机森林等在内的一系列分类器进行训练，并对每个用户发布的内容使用 LDA 主题模型确定讨论主题，在情感计算的基础上加入情感特征，提高了各分类器的检测准确率。此外，通过比较人类账号和社交机器人在正面情绪、负面情

绪、情绪变化等方面的差别，发现社交机器人发布内容很少会体现情绪的强烈变化，而人类用户的情绪则表现为更加反复无常，当一个用户推文内容的情感得分在 0.5～0.9 时，其更大概率为人类用户而非社交机器人。文献 [30] 基于图像、社交网络邻居节点、账号时序行为等提出 10 种新特征，并训练随机森林、贝叶斯网络、决策树等分类器进行评估。文献 [31] 使用基于随机森林、SVM、逻辑回归方法，基于用户数据和时序特征区分 3 种不同类别的社交机器人，实验结果表明随机森林始终取得最高的分类准确率。

当用于训练的数据没有类别标签时，可以使用无监督学习算法计算不同数据的相似特征将其归到相应类别。随着标注社交机器人数据集的丰富，当数据量较庞大时聚类算法的检测效率通常较低，因此无监督技术一般不单独使用，而是作为特征方法的一种对其他算法作为辅助补充。其中，使用最多的无监督学习算法是聚类方法，通过计算特定属性之间的距离度量样本之间的相似性。文献 [32] 设计了一种结合 SVM 和聚类算法的混合方法来提高垃圾邮件机器人检测的准确率，通过 3 种聚类算法结合决策树、SVM 等机器学习算法实现对垃圾邮件机器人的高效、高精度检测。文献 [33] 使用文本内容、图片、配置文件活动 3 个特征来检测在线社交网络中的异常行为，训练逻辑回归、随机森林、SVM 等分类器并使用主成分分析法根据重要程度对不同特征进行加权。为了研究不同类别账号对议题参与的偏向，使用 LDA 主题模型提取账号发布推文的主题，结果表明恶意社交机器人会针对特定主题展开不同的活动，而正常用户则不带有这种明显的偏向，会参与多个主题相关的不同活动。文献 [34] 利用 4 种分类器：朴素贝叶斯、SVM、随机森林和逻辑回归进行多分类，以检测人类用户、恶意社交机器人以及其他两种用于辅助新闻传播和用于消费的良性社交机器人，评估不同分类器在不同时间窗口下的检测效果，其实验结果为有效区分良性社交机器人和恶意社交机器人的特征作出重要贡献。

综上所述，特征的选择和改进对于提高模型准确率至关重要，提取新的特征加入训练是提升机器学习模型检测准确性的一个关键思路。由于不同机器学习算法在处理具有不同特征组合的数据集时表现优劣各异，以及不同机器学习算法对于数据的假设不同，如线性可分和线性不可分，因此在处理特定类型的数据和特征时各具优势。因此，基于机器学习方法的检测研究，其目标和优化思路为通过特征改进和消融实验以寻求最优的算法和特征组合，以达到最优的检测效果。

3 基于深度学习的社交机器人检测方法

深度神经网络（Deep Neural Networks, DNN）是深度学习领域使用最多、最主要的方法，在许多领域优于机器学习方法。神经网络最初是一个生物学概念，是指由大脑神经元、触点、细胞等组成的网络，用于产生意识以帮助生物思考和行动。后来人工智能研究者受神经网络的启发，发展出了人工神经网络，其主要思想是模拟人的神经元进行信息传递，接收并处理上层神经网络中神经元的信息，再传递给下一层中相邻的神经元，最后以大量神经元为基础相互连接构建多层神经网络。

3.1 图神经网络

传统神经网络无法处理图结构（也称拓扑结构）数据，如社交网络、化学分子结构、知识图谱等。图神经网络（Graph Neural Networks, GNN）根据理论及思想的差异可分为 3 大类：1）游走类模型，如 DeepWalk、Node2Vec、LINE、Stru2Vec；2）消息传递类模型，如 GCN、GAT、GraphSage；3）知识图谱类模型，如 TransE、TransR、TransD。基于随机游走（Random Walk）模型，研究者开发出 SybilGuard[35]、SybilResist[36]、SybilLimit[37]、SybilInfer[38]、SybilRank[39] 等检测社交网络中的异常账号的模型。以上方法对现实社交网络做出许多假设与限制：异常账号与正常用户所处区域差异很小；正常用户所处网络具有快速混合的性质；社交边缘表示强信任关系，攻击者很难与正常用户建立合法联系。但现实社交网络通常无法满足以上假设，因此上述方法在实际应用中通常无法取得理论结果，在大型社交网络上识别效果较差。

图卷积神经网络（Graph Convolutional Networks, GCN）是一种经典的消息传递类图神经网络[40]，借助拉普拉斯矩阵，利用特征分解和傅里叶变换得到易于计算的卷积核在图网络上进行卷积，能够巧妙地从大型图结构数据提取特征，根据图中的节点和连边信息对图中节点特征进行端到端学习得到图的嵌入表示并完成节点分类、图分类、边预测等任务，GCN 原理示意图如图 1 所示。

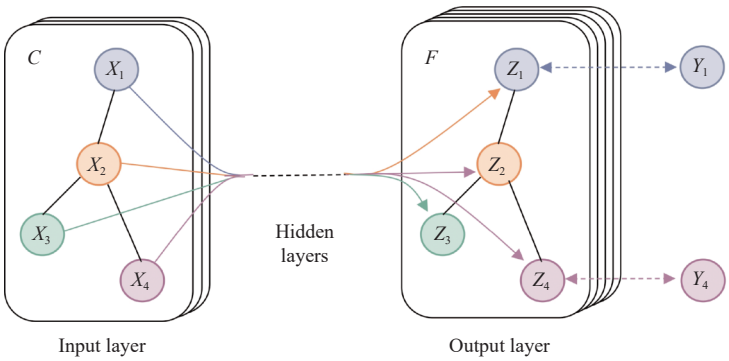

图 1　GCN 结构图

GCN 通过节点和相邻节点的信息来更新该节点的隐层信息，每一层只学习直接和节点相连的节点信息，因此当 GCN 模型设置多个层时，节点就可以间接学习到更外围邻居节点的特征，最后经过全连接层得到节点的预测标签类别。

早期的神经网络用于检测社交机器人时主要应用在文本和图像数据的特征提取方面，不能很好地处理如社交网络这样的图结构数据。并且深度学习更擅长处理大批量数据，利用神经网络能够自动从文本、图像、音频、视频等多种模态数据中提取低维特征。随着图神经网络相关理论的完善，基于 GNN 进行图分类、节点分类的研究逐渐成为热点。文献 [8] 使用 GCN 图卷积神经网络检测垃圾邮件机器人，利用 Twitter 账号的社交网络关系实现对不同账户进行分类，在 Twitter 数据集上取得了 89% 的准确率。GCN 既能考虑节点特征，又能聚合邻居节点的信息，适合处理社交网络数据。文献 [9] 基于情绪多样性特征和社交账号元数据特征提出一种基于图卷积神经网络的 A-GCNII 模型检测社交机器人，在新浪微博数据集上取得了较好的检测效果。

基于图的方法通常假设正常用户不会与社交机器人建立关注关系，然而现实是许多用户通过购买社交机器人账号作为粉丝以提升其社会影响力，另一方面社交机器人也会通过捕获长时间无人使用的真实用户的僵尸账号进行伪装，这使得社交机器人与真实用户之间的界限不再清晰，因此近年来基于图的检测在真实社交网络中的应用逐渐受限。

3.2　生成对抗网络

生成对抗网络（Generative Adversarial Networks, GAN）是一种新兴的神经网络技术，也是一种半监督方法[41]，被广泛应用于计算机图像和自然语言处理等领域。GAN 主要由生成器（Generator）和判别器（Discriminator）两部分构成，能够有效提取数据中的行为特征[42]。其核心思想包括两部分：首先，由生成器通过学习已有数据集，模仿真实数据不断产生新的虚假数据以混淆判别器；同时，判别器必须充分学习真实数据的潜在特征，才能区分真实数据与生成器产生的虚假数据。因此，两者之间相互博弈竞争并不断进化，最终实现在预测时即使判别器在数据集中从来没见过某类机器人账号，对真实数据学习的足够好的判别器依然能够正确区分真实账号和机器人账号。

文献 [43] 使用 GAN 基于文本内容实现社交机器人检测，并将该方法集成为 GANBOT 框架，以捕捉更多社交机器人的行为模式。并在生成器和分类器之间共享一个 LSTM 层，消除了传统 SeqGAN 的收敛限制，在检测准确率上优于现有的上下文 LSTM 方法。文献 [44] 提出训练生成对抗网络 GAN，使用判别器来进行机器账号检测，将二分类问题转化为单分类问题，只需要真实账号的示例即可得到效果良好的检测模型，在一定程度上解决了模型被动性和泛化性的问题。文献 [45] 提出 CS-GAN 模型，结合强化学习、对抗学习以及循环神经网络自动生成内容扩大原始数据集，在监督学习期间提高了检测模型的泛化能力。文献 [46] 构建生成对抗网络 SpamGAN 模型，使用有限的标记数据集检测在线评论中的垃圾邮件，并尝试改善社交机器人检测任务中的普遍存在的数据分布不均衡问题，检测结果优于最先进的有监督和半监督方法。

社交机器人开发技术和检测方法之间存在一种相互对抗、此消彼长的博弈关系，两者在开发和检测技术之间循环演进、螺旋上升。一方面，根据进化论研究可知，社交机器人的开发者会根据现有检测方法所依据的特征改进社交机器人开发技术，更新社交机器人的账户特征，使社交机器人能够尽可

能逃避最新的检测方法，从而使社交机器人不断进化。文献 [47] 研究了社交机器人为避免被 Twitter 发现而采取的不同方法，研究结果显示，社交机器人会通过购买虚假粉丝和发布更多推文等行为提高其账户信誉，缩小与真实用户之间的差异。另一方面，研究社交机器人检测方法的人员也会根据新型社交机器人的特点提取新的检测特征，开发新的检测方法，使检测算法尽可能准确全面地检测社交机器人账户。

3.3 循环神经网络

长短期记忆网络（Long Short-Term Memory, LSTM) 是循环神经网络 RNN 的一种变体，RNN 首先将时序概念引入神经网络设计中，在时序数据处理方面能够表现出较强的性能。但随着时间的推移，RNN 会出现长期记忆能力不足的缺陷，导致梯度消失、梯度爆炸等问题。LSTM 的核心概念在于细胞状态以及“门”结构，通过门控制处理长距离依赖的时间序列数据，细胞状态相当于信息传输的路径，让信息能在序列连中传递下去，其网络结构如图 2 所示。LSTM 神经网络通常与 CNN 卷积神经网络联合对用户行为进行建模，共同搭建检测模型用于社交机器人检测，其中 CNN 用于提取内容特征，LSTM 则用于处理时序相关的信息，如账号的各种点赞、发文、转发等行为。

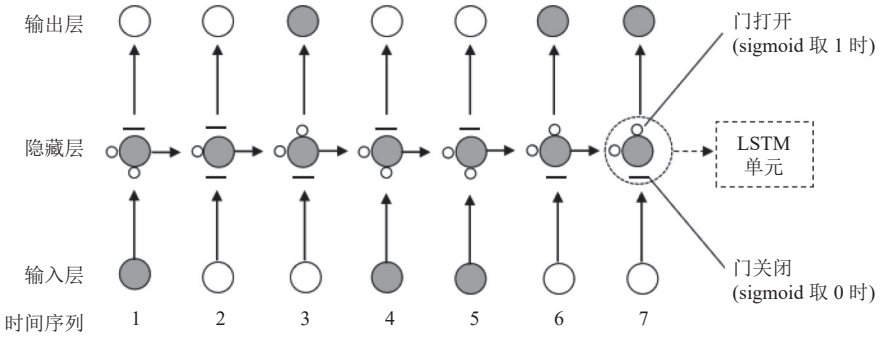

图 2 LSTM 网络结构

社交机器人在设计时会使用一定的调度算法设计账号的活动时间及活动类型，因此其并不随机，而是具有某种分布规律。相比之下，人类账户的活动时间通常随机、没有规律且不可预测。现有的基于深度学习的检测方法虽然克服了对文本内容提取特征的局限，但忽略了对于用户行为的建模。基于此，文献 [10] 使用 BiLSTM、CNN 以及注意力机制构建深度神经网络模型 DeepSBD，对推特账号的昼夜行为以及行为的周期规律提取时序特征，在 5 个基准数据集上进行验证，实验结果表明该模型的检测效果优于最先进的基线方法。文献 [11] 首次将词嵌入方法加入 RNN，使用词嵌入方法代替传统特征工程或复杂的自然语言处理方法编码推文并加入 BiLSTM，不需要事先获取用户的个人资料、社交网络结构或账号的历史行为，有效获取上下文特征，使得检测模型速度更快，更容易实现和部署，能够有效获取上下文信息，提高分类准确度。文献 [12] 提出一种结合文本信息与用户元数据特征的 LSTM 模型，从用户元数据特征中提取辅助特征，输入处理推文文本内容的 LSTM 神经网络，获得超过 96% 的 AUC 得分；并使用过采样技术，从大约 3 000 个社交机器人示例账号生成适合深度神经网络训练的大型标注数据集，从而解决数据分布不均衡问题。考虑到之前的研究大多关注提取文本的内容特征而忽略了社交机器人生成文本的情感属性，文献 [13] 提出一种结合情感特征的恶意社交机器人检测方法，训练带有注意力机制的双向长短期记忆网络 (Bidirectional Long Short-Term Memory model, Bi-LSTM) 对文本内容进行情感计算，分析账号发布文本的情感波动情况并结合元数据特征进行检测。实验结果表明在加入情感特征后，不同机器学习模型 (逻辑回归、决策树、Random Forest、Adaboost、SVM、DNN) 的检测准确率均得到了提高。

综上所述，循环神经网络能够很好地处理复杂文本序列，适合于处理社交媒体上的大量文本数据，能够通过学习和记忆上下文信息有效识别社交机器人账户与正常用户所发布内容间的差异，并结合卷积神经网络和注意力机制等其他技术以提高模型检测效果。

4 其他检测方法

4.1 蜜罐技术

蜜罐（Honeypot）是一种主动防御技术，多用

于网络安全防护，通过模拟一个或多个易受攻击的主机或服务来吸引攻击者，获取流量与样本，从而发现网络威胁、提取威胁特征。文献 [14] 实施了一个蜜罐陷阱，成功检测到数千个社交僵尸。其检测流程大致为：创建一些 Twitter 社交机器人账户，这些账号只会发布一些随机生成的逻辑不通且语序混乱毫无意义的推文，当正常用户看到这些推文时并不会产生兴趣也不会对帖子进行点赞、转发、评论等交互操作，然而这些虚假账号却会吸引盲目、随机关注账号来扩大自己社交圈的社交机器人，以此捕捉社交机器人账号。

文献 [15] 利用蜜罐技术构建了 HoneySpam 2.0 系统跟踪记录社交网络垃圾邮件机器人的行为，其收集的信息可用于检测垃圾邮件机器人和识别垃圾邮件内容。文献 [16] 使用蜜罐配置文件以识别垃圾邮件发送者的行为和特征。文献 [17] 在 Twitter 和 MySpace 网站上配置蜜罐系统来捕获垃圾邮件配置文件特征，最后使用深度学习分类器评估新旧垃圾邮件发送者的特征。文献 [18] 开发出一个基于蜜罐的垃圾邮件发送者检测系统，实现对恶意用户在社交网络上的发帖、私信、关注等交互行为的监控，并利用监测数据提取特征训练机器学习分类器。相比传统蜜罐系统有以下优点：数据集的可移植性、特征属性的跨平台一致性以及更快的检测速度。

4.2　新兴的检测方法

研究发现，经过多代进化的社交机器人账号与真实用户之间的差异变得微乎其微，因此根据已有特征检测单一账户变得愈发困难。相反，新型社交机器人具有群体特征明显而个体特征模糊的特点，因此检测工作应侧重于挖掘社交机器人账户组的集体特征。研究表明社交机器人群体在进行恶意行动时，通常同时、同步地进行评论、发帖、转帖、点赞等操作，因此利用账号系列行为在空间与时间等维度上的相似性检测社交机器人群体是应对社交机器人的行之有效的方法。

基于上述思想，文献 [19] 提出一种称为“社会指纹”的检测技术，利用编码成字符串的数字 DNA 检测第三代垃圾邮件机器人。类似生物的基因信息通过 DNA 碱基对排列结构存储，每一个在线社交网络中的账号也可以使用数字 DNA 记录其在社交平台上的一系列行为，并存储在字符串内。文献 [20] 提出一种利用数字 DNA 编码混合 BERT 模型进行情感分类预训练的混合技术，在数字 DNA 的 B3 内容编码上进行扩展并新提出 B5 类型编码，使模型在垃圾邮件机器人检测任务应用上更具鲁棒性，达到了 85.79% 的准确率。数字 DNA 结构如图 3 所示。

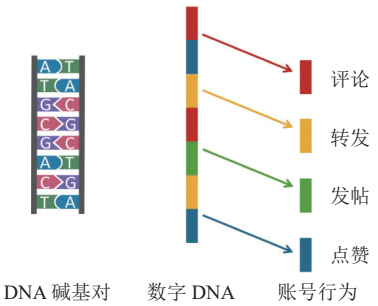

图 3　数字 DNA 结构

该检测方法的核心思想是，首先通过字符串编码技术构建数字 DNA，对在线用户的行为进行编码，通过将账号的各种在线活动映射到一个特定碱基：评论行为-T，转发行为-C，发帖行为-G，点赞行为-A，进而应用生物学领域标准的 DNA 分析技术计算最长公共子序列（Longest Commen Substring, LCS）的相似性来检测具有相似数字 DNA 的账号群体，以此发现社交机器人集群，进而区分 Twitter 上的真实账户和发送垃圾邮件的社交机器人账户。此外，由于不考虑社交网络图的性质，它降低了数据收集过程的成本。

5　结束语

综上，本文围绕社交机器人检测基本框架，详细分析了各部分检测流程，介绍了用于社交机器人检测的主要算法原理和检测步骤，以及国内外针对进化后的新型社交机器人所研究的检测方案。最后，对社交机器人未来发展前景及检测算法的改进方向进行了讨论，阐述了关于使用机器学习、深度学习以及其他新兴方法检测社交机器人的流程、优缺点、识别精度、适用范围等方面的研究现状。下面对社交机器人未来的发展趋势及检测技术的改进方向进行探讨与总结。

1）人类与社交机器人共存的挑战：未来在线社交平台中社交机器人之间的交互将成为常态，人类用户生活在一个充满机器人活动轨迹的世界。因此，提高在线社交网络用户的防范意识是减少社交机器人对舆论危害的关键。正常用户应提高对社交机器人的警惕，意识到在社交网络上分享个人图

片、位置等敏感信息造成隐私泄露的风险。这些个人数据有可能会被社交机器人自动收集并利用以伪装虚假账号，进一步威胁舆论安全。另一方面，受其背后操控者意图影响，社交机器人集群可以参与扩散并放大特定议题，并发布有利于其背后操控者所支持一方的言论，进而干扰真实的舆论环境。

2）社交机器人的深层次挖掘：检测出社交机器人后的逆向推理工作也非常重要，如：社交机器人的攻击目标是什么，它们如何生成推文内容，什么时候采取行动，倾向于参与哪种类型的话题，更容易影响什么样的人群等[48]。这有利于提前防范社交机器人入侵，最大限度减少其危害。此外，还可以对较易受社交机器人影响的群体进行分析，研究哪类人群更易与社交机器人等身份未知的账号进行交互，并提取其社交网络参与度、言语表达习惯[49]等特征，进而发掘此类用户内在的心理因素等特征从而进行针对性的研究[50]。

3）检测方法迁移能力的改进：由于获取数据集的平台不同，不同数据集所包含的特征类型各异。同样的检测方法使用不同的数据集测试得到的结果准确率差异很大，因此研究检测方法跨平台可扩展性的提升对于提高社交机器人检测模型的泛化性和复用性具有重要意义。此外，运用重采样[51]、迁移学习[52]、对抗学习[53]等技术进行数据集扩充和增强以解决检测模型的过拟合问题也是未来的研究趋势之一。

4）社交机器人误检问题的解决：对于社交机器人的误检问题，学界不断改进模型以提高检测准确率。①特征工程的改进，在使用机器学习方法进行检测时，对于特征提取工程的改进一直是研究的重点，通过从用户元数据、推文文本内容、文本情感、社交网络等不同类型的数据中提取特征丰富模型的训练进而提高准确率。②使用图神经网络获取社交网络中用户之间的联系和交互特征，从而挖掘单个账号所不具备的特征。③使用过采样、SMOTE、生成对抗等数据集增强方法改进因数据集中样本数据分布不均衡即社交机器人账号与正常账号数量差距悬殊对模型训练造成的误差，并通过集成学习方法综合多个模型或多个数据子集的预测结果，改善对于数据集中的少数类别（恶意社交机器人账号）检测的准确率，提高模型的鲁棒性和泛化能力。④使用交叉验证等方法对模型进行评估，避免模型在特定数据集上的过拟合现象[30]，从而减少误检率。

参 考 文 献

[1] CRESCI S. A decade of social bot detection[J]. Communications of the ACM, 2020, 63(10): 72-83.

[2] LATAH M. Detection of malicious social bots: A survey and a refined taxonomy[J]. Expert Systems with Applications, 2020, 151: 113383.

[3] DICKERSON J P, KAGAN V, SUBRAHMANIAN V S. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? [C]//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2014: 620-627.

[4] WANG Y H, WU C H, ZHENG K F, et al. Social bot detection using tweets similarity[M]//Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer, 2018.

[5] RAO S, VERMA A K, BHATIA T. A review on social Spam detection: Challenges, open issues, and future directions[J]. Expert Systems with Applications, 2021, 186: 115742.

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

[7] 刘蓉, 陈波, 于泠, 等. 恶意社交机器人检测技术研究[J]. 通信学报, 2017, 38(增刊 2): 197-210.
LIU R, CHEN B, YU L, et al. Research on detection technology of malicious social robots[J]. Journal on Communications, 2017, 38(Sup 2): 197-210.

[8] ALI A S, BIN T R, NAJAFI P, et al. Detect me if you can: Spam bot detection using inductive representation learning[C]//Proceedings of the 2019 World Wide Web Conference. New York: ACM, 2019: 148-153.

[9] 毛文清, 徐雅斌. 基于深度图卷积网络的社交机器人识别方法[J]. 电子科技大学学报, 2022, 51(4): 615-622.
MAO W Q, XU Y B. Social bot identify method based on deep graph convolutional network[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(4): 615-622.

[10] 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.

[11] WEI F, NGUYEN U T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings[C]//Proceedings of the 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications. New York: IEEE, 2019: 101-109.

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

[13] LONG G H, LIN D Y, LEI J, et al. A method of machine learning for social bot detection combined with sentiment analysis[C]//Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. New York: ACM, 2022: 239-244.

[14] LEE K, EOFF B, CAVERLEE J. Seven months with the Devils: A long-term study of content polluters on twitter[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2021, 5(1): 185-192.

[15] HAYATI P, CHAI K, POTDAR V, et al. HoneySpam 2.0: Profiling web spambot behaviour[M]//Principles of Practice in Multi-Agent Systems. Heidelberg: Springer, 2009: 335-344.

[16] STRINGHINI G, WANG G, EGELE M, et al. Follow the green: Growth and dynamics in twitter follower markets[C]//Proceedings of the 2013 Conference on Internet Measurement Conference. New York: ACM, 2013: 163-176.

[17] 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.

[18] ZHANG Y H, ZHANG H, YUAN X, et al. Pseudo-honeypot: Toward efficient and scalable Spam sniffer [C]//Proceedings of the 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. New York: IEEE, 2019: 435-446.

[19] 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, 2018, 15(4): 561-576.

[20] CHAWLA V, KAPOOR Y. A hybrid framework for bot detection on twitter: Fusing digital DNA with BERT[J]. Multimedia Tools and Applications, 2023, 82(20): 30831-30854.

[21] 张开放, 苏华友, 窦勇. 一种基于混淆矩阵的多分类任务准确率评估新方法[J]. 计算机工程与科学, 2021, 43(11): 1910-1919.
ZHANG K F, SU H Y, DOU Y. A new multi-classification task accuracy evaluation method based on confusion matrix[J]. Computer Engineering & Science, 2021, 43(11): 1910-1919.

[22] IRANI D, WEB B S, PU C. Study of static classification of social spam profiles in mySpace[J]. Cancer Cytopathology, 2013, 121(10): 591-597.

[23] 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. New York: ACM, 2016: 273-274.

[24] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//Proceedings of the 26th Annual Computer Security Applications Conference. New York: ACM, 2010: 1-9.

[25] WANG A H. Detecting spam bots in online social networking sites: A machine learning approach[C]//IFIP Annual Conference on Data and Applications Security and Privacy. Berlin, Heidelberg: Springer, 2010: 335-342.

[26] TAVARES G, FAISAL A. Scaling-laws of human broadcast communication enable distinction between human, corporate and robot Twitter users[J]. PLoS One, 2013, 8(7): e65774.

[27] LUTS J, OJEDA F, VANDEPLAS R, et al. A tutorial on support vector machine-based methods for classification problems in chemometrics[J]. Analytica Chimica Acta, 2010, 665(2): 129-145.

[28] KHALED S, EL-TAZI N, MOKHTAR H M O. Detecting fake accounts on social media[C]//Proceedings of the IEEE International Conference on Big Data. New York: IEEE, 2018: 3672-3681.

[29] DUTTA H S, CHETAN A, JOSHI B, et al. Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services[C]//Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: IEEE, 2018: 242-249.

[30] YANG C, HARKREADER R, GU G F. Empirical evaluation and new design for fighting evolving twitter spammers[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(8): 1280-1293.

[31] BEĞENILMIŞ E, USKUDARLI S. Organized behavior classification of tweet sets using supervised learning methods[C]//Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. New York: ACM, 2018: 1-9.

[32] SOHRABI M K, KARIMI F. A feature selection approach to detect spam in the facebook social network[J]. Arabian Journal for Science and Engineering, 2018, 43(2): 949-958.

[33] AL-QURISHI M, HOSSAIN M S, ALRUBAIAN M, et al. Leveraging analysis of user behavior to identify malicious activities in large-scale social networks[J]. IEEE Transactions on Industrial Informatics, 2018, 14(2): 799-813.

[34] OENTARYO R J, MURDOPO A, PRASETYO P K, et al. On profiling bots in social media[C]//International Conference on Social Informatics. Cham: Springer, 2016: 92-109.

[35] YU H, KAMINSKY M, GIBBONS P B, et al. Sybilguard: Defending against Sybil attacks via social networks [C]//Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols For Computer Communications. New York: ACM, 2006: 267-278.

[36] MA W, HU S Z, DAI Q, et al. Sybil-resist: A new protocol for sybil attack defense in social network [C]//International Conference on Applications and Techniques in Information Security. Heidelberg: Springer, 2014: 219-230.

[37] YU H F, GIBBONS P B, KAMINSKY M, et al. SybilLimit: A near-optimal social network defense against sybil attacks[C]//Proceedings of the IEEE Symposium on Security and Privacy. New York: IEEE, 2008: 3-17.

[38] DANEZIS G, MITTAL P. SybilInfer: Detecting Sybil nodes using social networks[C]//Proceedings of the Network and Distributed System Security Symposium. California: DBLP, 2009: 1-15.

[39] CAO Q, SIRIVIANOS M, YANG X W, 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. New York: ACM, 2012: 15.

[40] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the International Conference on Learning Representations. Toulon: ICLR, 2017: 1-14.

[41] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. California: NIPS, 2014: 2672-2680.

[42] MA J, GAO W, WONG K F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]//Proceedings of the WWW'19: The World Wide Web Conference. New York: ACM, 2019: 3049-3055.

[43] NAJARI S, SALEHI M, FARAHBAKHSH R. GANBOT: A GAN-based framework for social bot detection[J]. Social Network Analysis and Mining, 2021, 12(1): 4.

[44] 李阳阳, 杨英光. 基于生成对抗网络的社交机器人检测[J]. 计算机与现代化, 2022(3): 1-6.
LI Y Y, YANG Y G. Social bots detection based on generative adversarial networks[J]. Computer and Modernization, 2022(3): 1-6.

[45] LI Y, PAN Q, WANG S H, et al. A generative model for category text generation[J]. Information Sciences: An International Journal, 2018, 450(C): 301-315.

[46] STANTON G, IRISSAPPANE A A. GANs for semi-supervised opinion spam detection[EB/OL]. [2023-07-28]. https://arxiv.org/pdf/1903.08289.

[47] YANG C, HARKREADER R, ZHANG J L, et al. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter [C]//Proceedings of the 21st International Conference on World Wide Web. New York: ACM, 2012: 71-80.

[48] WAGNER C, MITTER S, KÖRNER C, et al. When social bots attack: Modeling susceptibility of users in online social networks[J]. CEUR Workshop Proceedings, 2012, 838: 41-48.

[49] BRISCOE E J, APPLING D S, HAYES H. Cues to deception in social media communications[C]// Proceedings of the 2014 47th Hawaii International Conference on System Sciences. New York: ACM, 2014: 1435-1443.

[50] WALD R, KHOSHGOFTAAR T M, NAPOLITANO A, et al. Predicting susceptibility to social bots on Twitter[C]//Proceedings of the IEEE 14th International Conference on Information Reuse & Integration. New York: IEEE, 2013: 6-13.

[51] 罗云松, 黄慕宇, 贾韬. 重采样在微博机器人识别中的应用研究[J]. 中文信息学报, 2021, 35(12): 133-148.
LUO Y S, HUANG M Y, JIA T. The application of resampling in recognition of microblog robots[J]. Journal of Chinese Information Processing, 2021, 35(12): 133-148.

[52] 刘勘, 杜妤宸. 基于深度迁移网络的 Twitter 谣言检测研究[J]. 数据分析与知识发现, 2019, 3(10): 47-55.
LIU K, DU H C. Detecting twitter rumors with deep transfer network[J]. Data Analysis and Knowledge Discovery, 2019, 3(10): 47-55.

[53] 郑赟, 马玉良, 陈林楠, 等. 基于生成对抗网络的情绪识别数据增强方法[J]. 传感技术学报, 2022, 35(12): 1650-1654.
ZHENG Y, MA Y L, CHEN L N, et al. Data enhancement method of emotion recognition based on GAN[J]. Chinese Journal of Sensors and Actuators, 2022, 35(12): 1650-1654.

编 辑 叶 芳