基于意图嵌入的社交机器人检测方法

牛红峰[1,2]，李嘉伟[1]，宋云鹏[1]，蔡忠闽[1,2]
（1.西安交通大学智能网络与网络安全教育部重点实验室，陕西 西安 710049；
2.西安交通大学自动化科学与工程学院，陕西 西安 710049）

摘　要：人工智能生成内容技术显著提升了社交机器人的伪装能力，给现有的机器人检测方法带来新的挑战。通过对社交平台用户意图进行建模，提出一种基于意图嵌入的社交机器人检测方法，从而避免直接在行为层面检测伪装能力大幅提升的机器人这一难题。实验结果表明，采用意图嵌入的检测模型相比未使用意图嵌入的模型，社交机器人检测准确率提高了 5.58 个百分点，并增强了对不同类型社交机器人的识别能力，验证了意图嵌入在提升人机检测任务性能中的有效性。

关键词：社交机器人检测；意图表征；意图嵌入；人工智能生成内容
中图分类号：TP301.6
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024205

Intention embedding method based social bot detection

NIU Hongfeng[1,2], LI Jiawei[1], SONG Yunpeng[1], CAI Zhongmin[1,2]
1. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China
2. School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Abstract: Artificial intelligence generated content technology has significantly enhanced the disguise capabilities of social bots, presenting new challenges to existing bot detection methods. By modeling the intentions of social media users through intention representation, a intention embedding method based social bot detection was proposed, thereby avoiding the difficulty of directly detecting bots with enhanced behavioral camouflage at the action level on social platforms. Experimental results show that the detection model using intention embedding improves the accuracy of social bot detection by 5.58 percentage points compared to models not utilizing intention embedding, and it enhances the recognition capability of specific types of social bots, verifying the effectiveness of intention embedding in improving the performance of human-bot detection tasks.

Keywords: social bot detection, intention representation, intention embedding, artificial intelligence generated content

0　引言

推特上的自动用户（又称“推特社交机器人”）是一种能够在平台上执行特定任务的自动化软件，包括自动发布推文、回复消息、关注其他账户以及收藏和转发推文[1]。尽管这些社交机器人有时用于增强信息传播和用户互动，但它们也常被用于执行恶意活动，如传播虚假信息[2]、进行网络骚扰[3]和性别歧视[4]，因此这类社交机器人被称为恶意社交机器人[5]。社交平台上的恶意社交机器人检测已成为网络安全研究中的关键议题。

社交机器人的各种恶意活动给在线社交网络带来了严重的信任危机。为应对这一挑战，研究者提出多种检测方法来识别这些社交机器人。目前，推

收稿日期：2024–07–10；修回日期：2024–11–11
通信作者：蔡忠闽，zmcai@sei.xjtu.edu.cn
基金项目：国家自然科学基金资助项目（No.62102308）
Foundation Item：The National Natural Science Foundation of China (No.62102308)

特上社交机器人的检测主要有 3 种方法[2,6-8]，即基于特征的方法、基于文本的方法和基于图的方法。基于特征的方法[9]通过分析推特上的行为指标（如转发量和点赞数）为机器学习模型提供输入数据；基于文本的方法[10]通过应用自然语言处理技术分析推文内容，辨别人类用户与社交机器人账号之间的差异；基于图的方法[11]则通过研究推特用户网络的结构属性，构建反映用户互动（如转发、提及和回复）的网络图谱，以识别不寻常的连接模式，帮助发现被操纵的账户群体或虚假信息的传播途径。综上所述，现有的检测方法主要侧重于对用户表层行为的建模和检测。然而，随着人工智能生成内容（AIGC, artificial intelligence generated content）技术的持续发展，社交机器人与人类用户在上述 3 个层面的差异逐步缩小。一方面，人工智能可以模拟人类的行为特征；另一方面，它可以生成高度接近人类文本内容的输出。此外，人工智能技术还能在社交图网络中伪造出真实的社交关系和社区结构，使得机器人具备人类用户的动态交互模式。这种高度拟人化的图网络伪装显著增加了检测难度，导致传统基于图网络的检测方法难以有效识别社交机器人。因此，AIGC 的发展对现有检测方法构成了多层面的重大挑战。

在社交平台上，可以根据用户的潜在使用动机对其行为进行分类和分析，包括发推、转发、回复和关注等。这种基于行为模式和上下文信息的分类分析揭示了用户背后的意图，如营销推广、信息传播或社交互动等[10,12]。例如，转发型社交机器人的主要目的是增加特定信息的可见性；营销型社交机器人致力于最大化内容的推广效果，同时隐藏其商业属性；操纵型社交机器人则专注于推广特定议题或观点，常在评论或推文中加入相关话题标签 #hashtag[1]。由此可见，社交机器人的行为目的通常较为单一和集中，而人类用户在社交平台上的动机则更多样化。

意图通常被定义为个体执行特定行为的目的或动机[13]。近年来，基于用户意图的研究在学术界得到了广泛关注，特别是在半自动驾驶、商品推荐和活动识别等领域的应用[14-16]。在网络安全领域，基于意图的研究也取得了重要进展。例如，Wu 等[17]通过分析用户提交敏感信息的意图，及时提供安全路径以防止信息被误送至不匹配的目的地，有效抑制了钓鱼网站的攻击，并显著降低典型钓鱼攻击的欺骗率。随后，Ji 等[18]设计了一种难以通过意图分析破解的新型恶意软件，该软件能够隐藏其恶意意图并精确锁定目标受害者。Pang 等[19]提出估计模型 AdvMind，通过主动合成查询反馈策略，不仅精准揭示黑盒攻击中对手的意图，还显著提高攻击成本，展现出与其他防御策略的协同潜力。总之，意图在网络安全领域的恶意软件识别、网络钓鱼检测和网络流量异常分析等方面显示出显著成效，为基于意图的社交机器人检测研究提供了坚实的基础和支持。

在社交机器人检测领域，意图的应用同样具有重要意义。人类用户和社交机器人在使用意图上表现出明显差异，人类用户的操作通常灵活且多样，意图广泛且多变；而社交机器人则基于特定目标和功能设计，展现出相对一致和集中的使用意图。这一差异成为区分人类用户与社交机器人的重要特征。因此，通过分析用户的意图特征，可以有效提高社交机器人检测的准确性。

本文旨在探讨如何通过用户意图有效检测推特上的恶意社交机器人。首先，利用大语言模型 ChatGPT 结合人类知识，对推特推文的意图进行精确分类和定义。然后，将这些分类后的意图标签嵌入神经网络模型中，采用迁移学习和微调策略，应用于区分社交机器人与人类用户的任务中，以提高社交机器人检测的准确性。本文的主要贡献总结如下。

1) 探讨用户意图在社交机器人检测中的应用潜力，以应对 AIGC 增强伪装能力对现有检测方法的挑战，并且提出从意图层面开展检测的创新视角，拓展了检测策略的维度。

2) 提出一种结合大语言模型和人类知识的人机协作方法，以提高社交平台用户意图的定义和分类精度。该方法通过自然语言处理和人类知识的协同标注，提升了意图标注的准确性和一致性，从而有效应对 AIGC 背景下意图识别的挑战。

3) 设计一种基于 Transformer 的网络模型，利用意图表征技术对社交平台用户的意图进行建模，并开发和训练一个融合意图嵌入的机器人检测模型。实验结果表明，与未使用意图嵌入的模型相比，本文方法在二元分类人机检测任务中的准确率最高提升了 5.58 个百分点。此外，该方法在社交机

器人类型识别的多分类任务中性能提升显著，验证了意图嵌入在提高人机检测任务性能中的有效性。

1 方法概况

1.1 基于意图嵌入的社交机器人检测方法设计

本文提出一种基于意图嵌入的社交机器人检测方法，该方法主要由基于大语言模型的意图定义与标注以及基于意图嵌入的社交机器人检测两部分组成，如图 1 所示。

1) 意图定义与标注

该部分通过结合大语言模型 ChatGPT 和人类知识，定义与标注用户的意图，并对推文数据进行处理。

2) 社交机器人检测

该部分包括意图嵌入和人机检测 2 个环节。在意图嵌入环节，将带有意图标签的推文向量输入基于 Transformer 的网络模型，进行标准深度学习训练。在训练后，选择性能最优的模型，并冻结除最后一层外的所有权重，为人机检测阶段做准备。在人机检测阶段，通过微调策略调整冻结的权重，以适应人类与社交机器人之间的区分任务。网络模型接收来自人类和社交机器人的输入数据，并在未知数据集上进行部署，以评估其检测性能。

2 基于大语言模型的意图定义与标注

本节主要介绍本文实验使用的数据集，提出基于大语言模型的意图标注方法，并说明推文数据的预处理方法。

2.1 数据集描述

Cresci2017 数据集是一个在推特平台上广泛应用的社交机器人数据集[20]，专门用于研究和识别社交网络中的社交机器人。该数据集记录了推文内容等多种元数据，例如，账户创建日期、关注者数量和地理位置信息等，旨在支持对社交平台自动化账户活动的检测与分析。

1) 人类用户数据集：Cresci 等[20]通过混合众包策略随机选取推特账户，并审核其回复以确认是否为人类用户，未回应的账户被排除。

2) 1 号社交机器人：设计目标是支持特定政治候选人，通过模仿人类用户行为来快速增强在推特上的影响力。

3) 2 号社交机器人：设计目标是推广 Talnts 移动应用，通过发布常规内容和提及用户来推广产品。

4) 3 号社交机器人：设计目标是宣传亚马逊公司的商品，通过发布商品链接进行推广，同时保持一定程度的真实性。

一条推文数据包括内容和元数据，内容反映用户的观点或情绪；元数据包括发布时间、点赞数、回复数等，用于分析传播模式和影响力。本文利用推文内容来分析用户意图，并构建内容网络。

2.2 基于大语言模型的意图定义与标注方法

本文通过分析用户推文内容并对其中的意图进行标注，以提高模型区分人类用户与社交机器人的能力。由此可知，精确定义和分类推特上的意图至

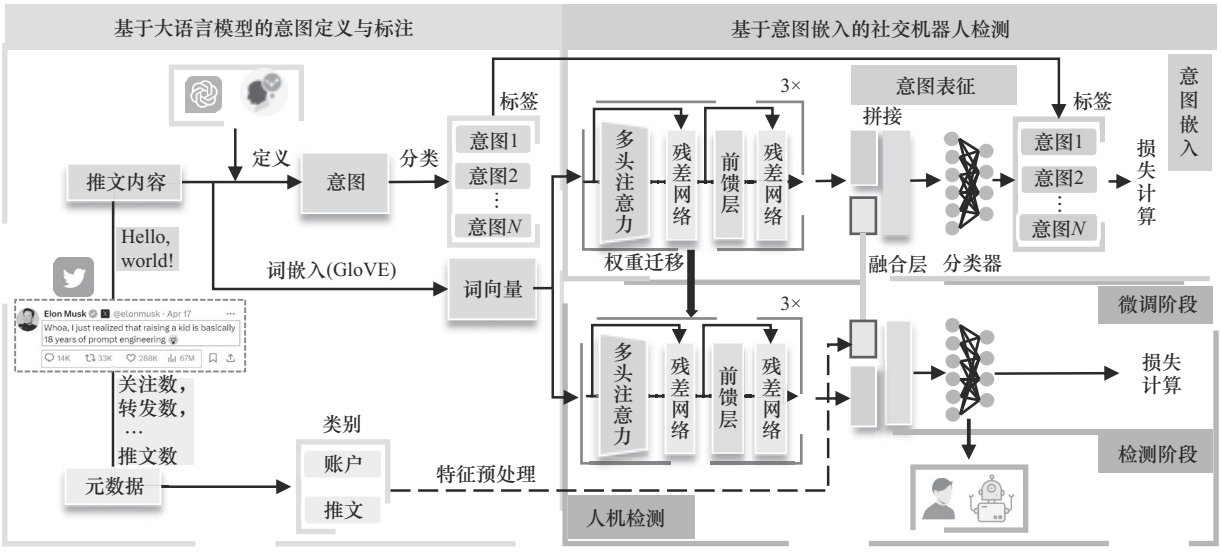


图 1 基于意图嵌入的社交机器人检测方法

关重要。为此，本文提出一种结合大语言模型ChatGPT与人类知识的意图定义、分类及标注方法。首先，通过开放编码对数据集进行初步定义与分类；然后，利用ChatGPT的知识库进一步细化这些定义；最后，借助人类专家的知识对意图分类进行调整，使模型的准确率提升至98%。该人机协作策略构建了精确的意图定义模型，为社交平台上的行为分析提供了有力支持。基于大语言模型的意图标注流程如图2所示。

2.2.1　推特用户意图的定义及分类

本文采用开放编码对数据集进行初步定义和分类，开放编码是一种定性研究中的常用方法，旨在识别和定义数据中的概念和类别[21]。随后，利用ChatGPT对用户意图进行详细定义与分类，最终完善所选数据集中的意图定义。

2.2.2　基于ChatGPT的意图标注模型构建

将定义和分类作为提示输入ChatGPT，以促使其学习并构建初步的意图标注模型，本文采用基于OpenAI GPT-4架构的ChatGPT进行实验分析。

2.2.3　人机一致性判断

将模型标注的意图结果与人工确定的意图标注进行比较，以评估模型的分类性能。本文目标是使模型的意图分类准确率至少达到98%，表明其已高度训练并具备精确分类的能力。经过上述3个阶段的训练和优化，最终获得一个更成熟且精确的意图定义模型，该模型能够有效理解和分类社交平台用户的意图，为数据分析和决策提供有力支持。具体的意图标注过程包括以下7个步骤。

1) 基于开放编码的初步意图定义：本文采用开放编码方法对选定数据集进行意图定义与分类，并根据不同社交机器人的设计目标明确相关意图类别。招募6名社交平台用户对数据集的意图进行分类。

2) 基于ChatGPT的意图定义：ChatGPT通过其知识库提供7个主要意图类别及详细描述，以帮助定义推特平台用户的行为意图。

3) 意图定义的确定：综合分析选定数据集和社交机器人的设计目标后，确定6个主要意图类别，并对各类别意图的具体定义进行调整，将其作为提示输入ChatGPT进行学习。

4) 真实意图标签生成：从数据集中随机抽取300个样本，基于新定义的6个意图类别进行人工标注，生成真实意图标签。

5) 意图标签一致性评估：将样本输入基于ChatGPT的意图标注模型，发现278个样本的标注结果与真实标签一致，结果一致的概率达92.7%。

6) 意图标注模型微调：对标注错误的样本进行分析后，针对性地调整输入提示，反复优化模型，旨在将意图标注准确率提高至98%以上。

7) 模型标注性能验证：经过多轮微调后，随机抽取200个样本进行验证，ChatGPT准确标注198个样本，意图一致的概率达99%，表明模型已达到预期的98%以上标准。

通过上述过程，本文将推特用户的意图细分为六大类，具体定义及实例如表1所示，并成功实现对30 000条人类用户推文和30 000条社交机器人推文的自动标注。

2.3　推文数据处理

本节旨在对推文内容和元数据进行预处理。推文内容用于构建文本神经网络模型，而元数据作为辅助信息，以增强对推文内容的理解。

2.3.1　推文内容向量嵌入

在将推文数据输入基于Transformer的模型处

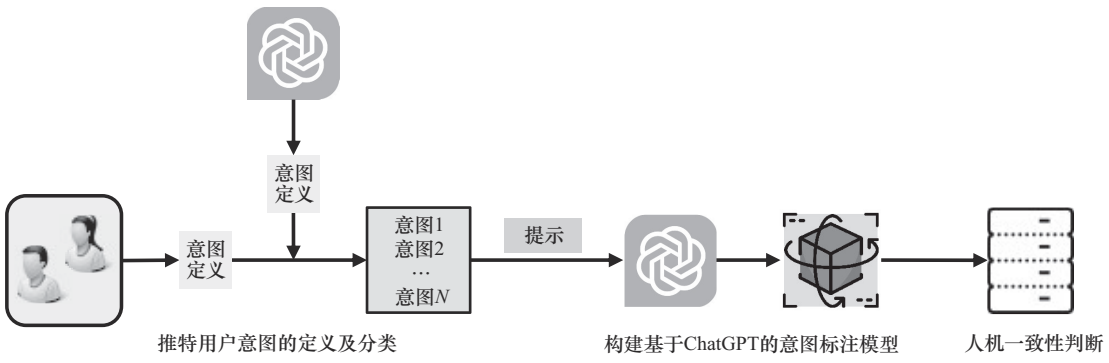


图2　基于大语言模型的意图标注流程

表1 推特用户意图分类

意图分类	定义	实例
个人表达和社交互动	分享个人生活	分享日常生活
	表达意见和情感释放	发表话题的意见
	建立和维护人际关系	转发、评论
信息获取和知识分享	追踪、获取新闻和时事更新	浏览新闻信息，收藏、评论新闻信息
	探索并获取教育和学习资源	收藏、转发知识
	分享专业知识和见解	发表专业的评论
娱乐和休闲	追踪娱乐、名人和粉丝动态	关注并跟踪偶像动态
	参与娱乐活动和挑战	歌曲分享
	游戏和趣味性内容的分享	游戏分享
商业、营销和品牌推广	推广产品和服务	产品广告宣传
	进行市场调研和客户互动	投票活动
	建立和提升品牌形象，建立品牌维护	公司企业文化宣传
职业发展和专业网络扩展	职业机会的探索和分享	招聘信息的分享
	行业洞察和趋势分析	行业信息报告
	建立专业网络和合作关系	与领域相关的博主进行互动
政治与社会运动	高拟人程度和宣传政治议题	纸质话题发表意见
	参与社会运动和公共议题	高拟人程度社会活动
	组织和动员支持者	政治呼吁

理之前，本文通过自然语言处理方法进行预处理，以提升数据质量和模型效率。核心过程包括以下4个步骤。

1) 文本清洗：移除无关字符和标点，规范化文本，消除噪音。

2) 词汇表构建与文本分割：使用字节对编码（BPE, byte pair encoder）算法[22]对文本进行分割，构建包含基础标记单元的词汇表，以增强模型的泛化能力。

3) 数据增强：通过同义词替换和随机插入，增强数据多样性并提升模型的鲁棒性[1]。

4) 词嵌入：采用GloVETwitter-100词向量模型进行词嵌入，以捕捉推特的语义和句法关系[1,23]。

上述步骤确保了文本数据的清洁、规范化和丰富的语义表示，为后续深度学习模型提供了可靠基础。

2.3.2 特征选择

推特元数据包括账户元数据和推文元数据，为用户行为和内容传播分析提供关键信息。账户元数据（如推文数量、关注者数）揭示账户属性与社交连接性，推文元数据（如转发数、回复数）反映其影响力和传播模式。本文基于Cresci2017数据集，选取6个推文特征和8个账户特征，具体定义如表2所示。

3 基于意图嵌入的社交机器人检测

基于意图嵌入的社交机器人检测包含意图嵌入和人机检测2个环节。意图嵌入通过深度学习的训练模式将用户意图嵌入基于Transformer的网络模型中；人机检测通过微调的方式将上述意图嵌入后的深度网络迁移应用于人机检测任务中。

3.1 基于Transformer的网络模型

本文设计了2种基于Transformer的网络模型输入结构，一种专注于推文内容，另一种将推文内容与元数据相结合。图3展示了这2种结构的具体处理流程，模型采用Transformer架构，由3个整合多头注意力机制和前馈网络的编码器层构成，嵌入残差连接以提升稳定性和效率，从而优化自然语言处理任务的性能。

表 2　　推特元数据及其描述

类别	特征	定义
元数据（账户）	推文数	指一个推特账户发布的推文总数,反映用户的活跃度和内容发布的频率
	关注者数	指账户的关注者数量,通常被视为衡量账户影响力和受欢迎程度的指标
	关注数	表示该账户关注的其他推特账户的数量,可以提供用户社交活跃性的信息
	收藏数	指一条推文被用户点赞(曾称为“喜欢”)的次数,是衡量用户对推文内容赞赏或认同的一个简单指标
	被列入表数	指该账户被其他用户添加到推特列表的次数,是衡量账户在特定领域或社群中的影响力和重要性的一个指标
	默认配置文件	如果用户没有对其推特配置文件进行大量个性化设置,则此项为真,可能表明账户较新或用户对个性化设置不关心
	地理定位启用	指用户是否启用了推文地理位置标记功能,如果启用,则推文可以包含用户的地理位置信息
	使用背景图片的个人资料	指用户是否在其推特个人资料中使用了背景图片,是个人化设置的一部分
元数据（推文）	转发数	指一条推文被其他用户转发的次数,转发数越高,通常表示该内容在推特社区中受欢迎或引起共鸣的程度越高
	回复数	指一条推文收到的回复数量,它可以衡量一条推文的互动性或争议性,反映用户的直接互动
	点赞数	显示了一条推文被用户点赞(曾称为“收藏”)的次数,是衡量用户对推文内容赞赏或认同程度的一个简单指标
	话题标签数	统计一条推文中使用的话题标签数量,话题标签是对内容进行分类并使其易于发现的一种方式,也可用于强调或作为在线活动的一部分
	网址数	指一条推文中包含的网址或网络链接的数量。链接通常用于引导关注者到推特外的更详细内容,如新闻文章、博客、视频
	提及数	该特征统计了一条推文中提及其他推特用户的次数,通过‘@’符号后跟用户名来提及,用于与其他用户直接互动或归属内容

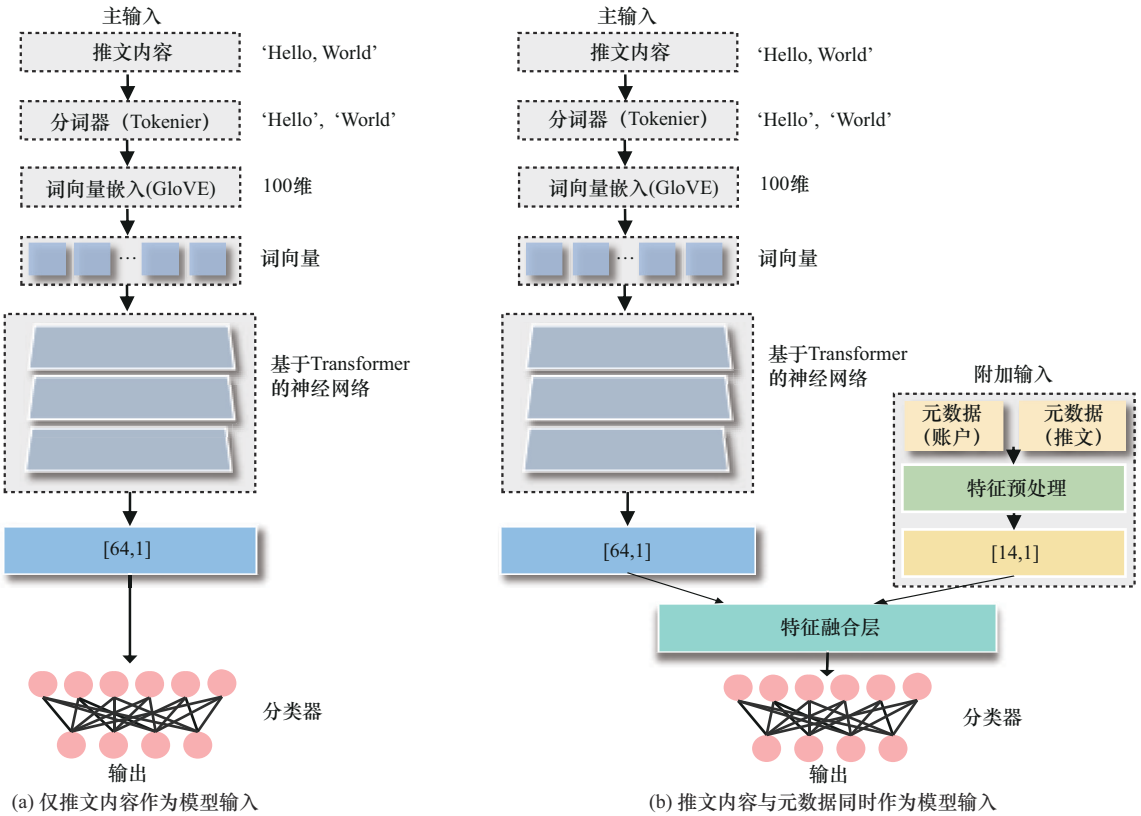


图 3　2 种不同的模型输入结构

分类器由2个全连接层构成。特征向量先经过第一全连接层并经ReLU激活后，再作为第二层的输入。第二层的输出经Softmax函数处理，生成最终的类别预测概率。该设计通过非线性变换和概率归一化提升了模型的分类效能。

关于输入数据的处理，推文内容首先经过分词，并利用GloVETwitter-100生成100维词向量。然后，这些词向量输入至基于Transformer的网络模型中，该模型输出64维深度特征向量，用于进一步分析。

1) 推文内容：推文内容作为主要输入送入设计精良的基于Transformer的深度神经网络，生成一个64维特征向量。该向量随后被输入分类器中，分类器专门从文本内容中提取用于分类的特征。

2) 推文内容−元数据融合：在以推文内容为主要输入的基础上，引入归一化的元数据作为模型的辅助输入，并将其与推文内容的特征向量结合形成联合特征向量。该联合特征向量被输入分类器中，以增强模型对推文的综合理解能力，从而提高分类的准确性和鲁棒性。

3.2 意图嵌入环节

本文采用深度学习预训练方法，将推文内容及元数据输入基于Transformer的模型，并嵌入意图标签。训练过程遵循标准流程，旨在精准融入用户意图。训练完成后，除最后一层外，其余权重被冻结，以用于后续人机检测任务。

3.3 人机检测环节

人机检测环节包括微调阶段和检测阶段。在微调阶段，使用特定的人类用户与社交机器人数据集对预训练深度学习模型进行再训练和参数微调，以增强模型区分人类用户和社交机器人的能力。在检测阶段，微调后的模型被部署在实际社交机器人检测任务中，以验证方法的有效性。

3.3.1 微调阶段

将意图嵌入的网络模型迁移至人机检测任务中，包括人类用户与社交机器人的二元分类任务以及具体社交机器人类型识别任务。通过微调模型，以适应不同任务的需求。模型输出经过Softmax层处理后，计算损失以优化模型。本文探讨了2种常见的微调模式来实现对意图嵌入预训练模型的微调。

1) 微调模式1：在此模式下，预训练模型中除了最后一层全连接网络外，其他部分均保持冻结状态，仅对最后一层全连接网络进行微调。

2) 微调模式2：对整个网络进行重新训练。在该模式下，本文采用对整个网络进行重新训练的微调方法，即对预训练模型的所有层执行微调操作。在此策略下，整个网络被纳入学习过程，以便更好地适应新的任务需求。

3.3.2 检测阶段

在检测阶段，将微调后的模型应用于新的、未知的社交机器人检测任务中，以评估不同微调策略在实际应用中对模型检测性能的影响。

4 实验

4.1 数据集

本文将每条推文视为一个独立的研究样本，针对人类用户推文，本文实验选取60 000条推文组成人类用户的数据集。针对3种社交机器人（1号社交机器人、2号社交机器人、3号社交机器人），每种社交机器人分别选取20 000条推文，共计60 000条，构成社交机器人样本数据集。本文将上述数据集划分为3个子集，按照50%、25%和25%的比例进行分配，分别命名为数据集1、数据集2和数据集3。3个子集的具体划分与用途描述如下。

1) 数据集1：包含30 000条人类用户推文和1号、2号、3号社交机器人各10 000条推文，共计60 000个样本。该数据集用于人类与社交机器人意图嵌入训练。

2) 数据集2：包含15 000条人类用户推文和1号、2号、3号社交机器人各5 000条推文，共计30 000个样本。该数据集用于对人机检测模型进行微调训练。

3) 数据集3：包含15 000条人类用户推文和1号、2号、3号社交机器人各5 000条推文，共计30 000个样本。该数据集用于评估微调后的模型在人机检测任务中的性能，包括人机二元分类和社交机器人类型识别的多分类实验。

4.2 评估指标

本文采用真阳性率、真阴性率和准确率3个核心指标进行评估，以衡量模型在区分人类用户与社交机器人时的检测效果及分类准确性。

1) 真阳性率：表示测试中的机器人样本中，

被正确分类为机器人的比例。

2) 真阴性率：表示测试中的人类样本中，被正确分类为人类的比例。

3) 准确率：表示所有测试样本中，被正确分类的比例。

4.3　基准模型

4.3.1　LSTM

本文参考文献[1]中提出的基于LSTM模型的方法作为深度基准模型，构建了一个双层LSTM模型用于社交机器人检测，具体的模型参数如表3所示。

表3　　模型参数

参数	数值
优化器	SGD
学习率	0.001
损失函数	交叉熵损失函数
Batch Size	16
Dropout	0.5

交叉熵损失函数定义为

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log q(x_i) \tag{1}$$

其中，$p(x_i)$是在真实分布p下第i个事件x_i发生的概率；$q(x_i)$是在预测分布q下第i个事件x_i发生的概率。

本文设计的基于Transformer的网络模型采用相同的参数设置。此外，考虑BiLSTM相较于LSTM的优势在于其双向结构可以同时捕捉序列中的前后文信息，从而提升对上下文的理解能力和预测精度，因此选择BiLSTM作为第二种基准模型。

4.3.2　机器学习模型

本文选择支持向量机（SVM, support vector machine）、逻辑回归（LR, logistic regression）和多层感知机（MLP, multilayer perceptron）3种经典的机器学习[23]模型作为基准进行比较，所有模型均基于Sklearn 0.24.2库实现，使用默认参数进行部署。

本文选择14种元数据作为机器学习模型的输入。为提高模型的性能并确保特征变量之间的可比性，在将所有特征变量输入模型之前，进行z-score标准化处理，以实现数据归一化[24]。

4.3.3　消融实验

本文进行了消融实验，探讨不同特征组合作为输入对模型性能的影响，具体分析了4类特征组合，包括仅推文内容、元数据（推文）+推文内容、元数据（账户）+推文内容、元数据+推文内容。此外，还探究意图嵌入对人机检测模型性能的影响，比较2种意图嵌入方法，即基于推文内容的嵌入与结合方法、推文内容和元数据的信息融合方法。具体实验包括人类与社交机器人的二元分类检测和社交机器人类型识别的多分类任务，以评估这些策略在不同任务中的效果影响。

此外，本文还引入一组对照实验，以评估未应用意图嵌入的深度神经网络模型的性能。实验设计中，仅使用数据集2进行模型训练，并以数据集3为性能评估的数据来源。

4.4　基于意图嵌入的检测实验

本文实验旨在评估本文提出的基于意图嵌入的社交机器人检测方法的有效性，并探讨4种不同特征组合在2种微调模式下对人类与社交机器人样本检测准确率的影响。

表4的结果表明，引入意图嵌入后，4种特征组合的检测模型性能均显著提升。具体而言，当输入模型的特征组合为“元数据+推文内容”时，准确率提升了5.58个百分点；而输入为“元数据（推文）和推文内容”时，准确率提升幅度最小，仅为1.79个百分点，两者相差3.79个百分点。这说明用户意图嵌入显著提高了检测效率，并且特征选择对模型性能优化具有至关重要的作用。

2种微调模式在不同特征组合下的表现差异明显，其中，微调模式2的性能显著优于模式1，可能是因为对网络的全面调整增强了模型的适应能力。

本文方法的检测准确率达99.39%，相较于基准模型LSTM和BiLSTM，准确率分别提升了12.94和11.79个百分点，显示了本文方法的显著优势。此外，未采用意图嵌入的Transformer模型准确率为93.81%，通过引入意图嵌入后，模型的性能得到了显著提升，进一步证明了意图嵌入策略在增强社交机器人检测中的有效性。与SVM模型相比，本文方法的准确率高出17.47个百分点，这表明在社交机器人检测任务中，深度学习方法表现出更强的优势。

综上所述，用户意图嵌入的检测模型被证明是一种有效的社交机器人检测策略。特征组合的选择

表4 基于意图嵌入的人机检测

输入	模型	准确率	真阳性率	真阴性率
元数据	SVM	81.92%	79.11%	84.74%
	Logistic Regression	80.06%	79.01%	81.04%
	MLP	80.09%	79.14%	81.03%
仅推文内容	Transformer（无）	73.93%	74.67%	73.19%
	Transformer（模式1）	74.96%	75.35%	74.57%
	Transformer（模式2）	79.91%	79.09%	80.73%
	提升	5.98%	4.42%	1.84%
元数据（推文）+推文内容	Transformer（无）	78.43%	78.54%	78.32%
	Transformer（模式1）	79.41%	78.68%	80.14%
	Transformer（模式2）	80.22%	80.20%	80.24%
	提升	1.79%	1.66%	1.92%
元数据（账户）+推文内容	Transformer（无）	78.50%	78.09%	78.91%
	Transformer（模式1）	81.46%	79.09%	83.83%
	Transformer（模式2）	83.86%	80.76%	88.95%
	提升	5.36%	2.67%	10.04%
元数据+推文内容	LSTM（文献[1]）	86.45%	85.19%	87.71%
	BiLSTM	87.60%	87.89%	87.31%
	Transformer（无）	93.81%	93.09%	94.54%
	Transformer（模式1）	94.36%	94.48%	94.25%
	Transformer（模式2）	99.39%	98.79%	99.98%
	提升	5.58%	5.70%	5.45%

对模型性能有显著影响，因此，为构建更精确的检测模型，应综合考虑推文内容与元数据的多源输入，以优化检测效果。

4.5 基于意图嵌入的社交机器人类别实验

在社交机器人类型识别任务的实验中，本文引入2种意图嵌入微调模式，以评估其对社交机器人类型识别性能的影响，实验结果如表5所示。由表5可知，当输入包含元数据与推文内容时，3号社交机器人的识别准确率从90.98%提升至99.24%，增幅达8.26个百分点。同时，对人类用户、1号社交机器人和2号社交机器人的识别准确率分别提高了1.01个百分点、1.84个百分点和13.82个百分点。实验结果表明，基于意图嵌入的神经网络模型能够有效提升社交机器人类型的识别性能。

此外，表5进一步证明，最优结果是由特征组合“元数据+推文内容”所获得的。在微调模式2下，模型在识别具体社交机器人类别和人类用户时的准确率均有所提升。相比之下，当输入为“仅推文内容”时，模型的识别性能表现不一致，虽然对人类用户和2号社交机器人的识别准确率分别提升了7.03和4.46个百分点，但对1号和3号社交机器人的检测准确率却有所下降。

当输入“仅推文内容”时，模型的识别准确率甚至低于机器学习模型。实验结果表明，结合元数据与推文内容作为输入，不仅提高了模型在人机分类任务中的识别准确率，还在与传统机器学习方法的对比中展现了优势，这证实了融合多种数据源对模型性能提升的重要性。

实验结果显示，在识别社交机器人类型时，“元数据+推文内容”的最佳微调策略为模式2，即

表5　基于意图嵌入的社交机器人类型识别

输入	模型	人类	1号社交机器人	2号社交机器人	3号社交机器人
元数据	SVM	78.21%	52.90%	36.50%	41.34%
	Logistic Regression	79.52%	55.10%	36.06%	51.62%
	MLP	79.09%	53.66%	36.42%	45.42%
仅推文内容	Transformer（无）	65.92%	95.02%	60.80%	75.64%
	Transformer（模式1）	78.85%	66.68%	16.70%	55.44%
	Transformer（模式2）	72.95%	94.28%	65.26%	67.88%
元数据 + 推文内容	Transformer（无）	89.61%	97.02%	34.76%	90.98%
	Transformer（模式1）	89.11%	95.02%	46.76%	98.72%
	Transformer（模式2）	90.62%	98.86%	48.58%	99.24%

对整个网络进行重新训练；而当输入仅为“推文内容”时，模型在识别具体社交机器人类型方面的表现不明显。上述结果进一步表明，元数据对于提升模型性能具有显著影响。

4.6　实验小结

本文通过分析人类用户与社交机器人在隐含意图层面的差异，明确了两者在意图特征上的差异性，即人类用户的文本通常内容丰富，而社交机器人的输出则较单调。在此基础上，本文提出一种基于意图嵌入的社交机器人检测方法，为模型提供深层次的上下文信息，从而增强对细微差异的识别能力。最后，通过针对人机检测任务的微调，本文模型在实际检测任务中的性能得到了显著优化，验证了意图嵌入技术在实际人机检测应用中的潜在价值。

实验结果表明，在人机二元分类和社交机器人类型识别的多分类任务中，结合推文内容与元数据信息作为模型输入的方法显著提高了检测性能。相较于仅使用推文内容，元数据提供额外的上下文信息，有助于捕捉用户特定的行为模式，从而弥补仅依赖文本内容的局限性。这种特征组合为区分不同样本提供了更有价值的信息，增强了模型的区分能力。通过综合多种数据源的信息，有效提升模型的检测能力和泛化性能。

5　讨论与局限性

5.1　意图的定义与标注

本文提出一种基于用户意图嵌入的社交机器人检测方法，通过挖掘人类用户与社交机器人的行为意图来识别恶意社交机器人，准确定义和标注用户意图是该方法的基础。为此，本文结合ChatGPT的知识与人类专家对数据集的理解设计了一种意图标注模型，精确定义了人类与社交机器人的互动意图。这种人机协同方法增强了意图分类的互补性。

尽管上述研究取得了显著进展，复杂的社交平台环境仍然带来了新挑战。用户意图并不总是显而易见，某些推文可能隐含深层次的意图。未来研究应开发更先进的算法，以更准确地识别潜在意图。本文基于人类用户与社交机器人的差异，将用户意图分为6类。实验结果证明了意图嵌入技术在提升检测性能方面的有效性，展示了其在广泛应用中的潜力。

5.2　检测模型的适应性挑战问题

本文检测方法依赖于意图嵌入网络模型，但意图分布可能会随着时间和地域的变化而改变。例如，在消费旺季和选举期间，社交机器人数量的波动对检测模型的准确性构成了挑战。本文方法与传统特征提取方法不同，其优势在于无需重新提取特征或开发新节点。该方法通过对新数据集进行再训练，简化了模型的适应过程，并增强了对新型场景的适应能力。

5.3　跨平台的社交机器人应用问题

本文聚焦于推特平台的社交机器人检测，验证了意图嵌入方法的有效性。推特提供了丰富的数据集，支持相关研究。本文方法基于消息内容和账户元数据，通过动态调整的方式适应其他平台，如新浪微博。

社交机器人检测的有效性可能受语言和文化背

景的影响。例如，新浪微博的中文用户群体与推特的多样性存在显著差异。因此，跨文化和跨平台的适应性成为开发检测工具的关键因素。

6 结束语

本文针对推特平台的社交机器人检测问题，提出一种基于意图嵌入的社交机器人检测方法。首先，利用ChatGPT和人类用户知识对用户意图进行定义和分类，并将这些意图嵌入基于Transformer的模型中作为推文标签。然后，探讨2种模型微调方法，以优化意图嵌入模型在人机二元分类和社交机器人类型识别的多分类任务中的表现。实验还分析了4种特征输入组合对模型性能的影响，实验结果表明，该方法明显提升了检测性能。然而，模型的泛化能力和跨社交环境的适应性仍需进一步验证。未来研究可以探索更多特征组合和更先进的模型结构，以增强模型的适用性与鲁棒性。

参考文献：

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

[2] FENG S B, TAN Z, WAN H N, et al. Twibot-22: towards graph-based twitter bot detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 35254-35269.

[3] ARIN E, KUTLU M. Deep learning based social bot detection on twitter[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1763-1772.

[4] GUNARATHNE P, RUI H X, SEIDMANN A. Racial bias in customer service: evidence from twitter[J]. Information Systems Research, 2022, 33(1): 43-54.

[5] 师文, 陈昌凤. 社交机器人在新闻扩散中的角色和行为模式研究: 基于《纽约时报》“修例”风波报道在Twitter上扩散的分析[J]. 新闻与传播研究, 2020, 27(5): 5-20, 126.
SHI W, CHEN C F. The role and behavior pattern of social bots in the diffusion of news: an analysis of the spread of coverages of anti-amendment bill campaign by the New York times on twitter[J]. Journalism & Communication, 2020, 27(5): 5-20, 126.

[6] OUNI S, FKIH F, OMRI M N. BERT- and CNN-based TOBEAT approach for unwelcome tweets detection[J]. Social Network Analysis and Mining, 2022, 12(1): 144-162.

[7] YE S, TAN Z X, LEI Z Y, et al. HOFA: twitter bot detection with homophily-oriented augmentation and frequency adaptive attention[J]. arxiv Priprent, arxiv: 2306.12870, 2023.

[8] ANTONAKAKI D, FRAGOPOULOU P, IOANNIDIS S. A survey of twitter research: data model, graph structure, sentiment analysis and attacks[J]. Expert Systems with Applications, 2021, 164: 114006.

[9] 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 Press, 2023: 485-495.

[10] NAJARI S, SALEHI M, FARAHBAKHSH R. GANBOT: a GAN-based framework for social bot detection[J]. Social Network Analysis and Mining, 2022, 12(1): 4-15.

[11] LINGAM G, ROUT R R, SOMAYAJULU D, et al. Social botnet community detection: a novel approach based on behavioral similarity in twitter network using deep learning[C]//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. New York: ACM Press, 2020: 708-718.

[12] HEIDARI M, JONES J H, UZUNER O. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter[C]//Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2020: 480-487.

[13] ZIMMERMAN M, BRATMAN M E. Review essay: intention, plans, and practical reason[J]. Philosophy and Phenomenological Research, 1989, 50(1): 189.

[14] PING Y K, GAO C, LIU T C, et al. User consumption intention prediction in meituan[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2021: 3472-3482.

[15] WU M, LOU W T, LAHIJANIAN M, et al. Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles[C]//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2019: 6210-6216.

[16] RAFFERTY J, NUGENT C D, LIU J, et al. From activity recognition to intention recognition for assisted living within smart homes[J]. IEEE Transactions on Human-Machine Systems, 2017, 47(3): 368-379.

[17] WU M, MILLER R C, LITTLE G. Web wallet: preventing phishing attacks by revealing user intentions[C]//Proceedings of the Second Symposium on Usable Privacy and Security-SOUPS’06. New York: ACM Press, 2006: 102-113.

[18] JI T T, FANG B X, CUI X, et al. Framework for understanding intention-unbreakable malware[J]. Science China Information Sciences, 2023, 66(4): 142104.

[19] PANG R, ZHANG X Y, JI S L, et al. AdvMind: inferring adversary intent of black-box attacks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1899-1907.

[20] CRESCI S, PIETRO R D, 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-WWW'17 Companion. New York: ACM Press, 2017: 963-972.

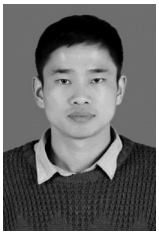
[21] STRAUSS A L, CORBIN J. Open coding[J]. Social Research Methods: A Reader, 2004: 303-306.

[22] GUTIERREZ-VASQUES X, BENTZ C, SAMARDŽIĆ T. Languages through the looking glass of BPE compression[J]. Computational Linguistics, 2023, 49(4): 943-1001.

[23] BISHOP C M. Pattern recognition and machine learning[J]. Springer Google Schola, 2006, 2: 1122-1128.

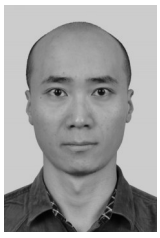
[24] SHEN C, CHEN Y F, GUAN X H. Performance evaluation of implicit smartphones authentication via sensor-behavior analysis[J]. Information Sciences, 2018, 430: 538-553.

[作者简介]



牛红峰（1989– ），男，河南新乡人，西安交通大学博士生，主要研究方向为网络安全与人机交互。

李嘉伟（2000– ），男，河北沧州人，西安交通大学硕士生，主要研究方向为网络安全与人机交互。



宋云鹏（1990– ），男，陕西宝鸡人，博士，西安交通大学副教授、硕士生导师，主要研究方向为人机交互、安全与隐私、混合增强智能。

蔡忠闽（1975– ），男，福建晋江人，博士，西安交通大学教授、博士生导师，主要研究方向为智能人机交互、混合增强智能、电力系统智能化、虚拟现实和增强现实。