基于细粒度特征权重专家网络的社交机器人检测方法

张怀博[1,2,3]，高金华[1,2]，廖逸之[1,2,3]，辛永辉[4]，程学旗[1,2]
1. 智能算法安全全国重点实验室，北京 100190；
2. 中国科学院计算技术研究所，北京 100190；
3. 中国科学院大学，北京 101408；
4. 国家计算机网络应急技术处理协调中心，北京 100029

摘要
近年来，社交机器人检测领域的研究已逐步从个体特征分析演进至群体特征挖掘，从传统特征工程升级为深度学习方法。其中，基于图网络的方法展现出显著优势，该方法能够融合账号行为特征、文本语义特征与网络拓扑特征，将社交机器人检测转化为图节点分类任务。然而，现有检测方法大多采用通用模型进行检测，未考虑不同类型社交机器人在细粒度特征上的差异，导致跨业务场景下的检测精度受限。基于此，提出一种基于细粒度特征权重专家网络的社交机器人检测方法。该方法通过构建业务专家网络，使每个专家专注于学习细粒度特征的差异化权重组合，然后借助多专家特征融合与综合研判，实现对潜在多业务类型社交机器人的融合检测。在公开推特数据集上的实验结果显示，该方法的性能优于现有主流检测方法，其中F1指标相对提升1.52%。

关键词
社交机器人检测；细粒度特征权重；混合专家网络

中图分类号：TP391.9 文献标志码：A doi：10.11959/j.issn.2096-0271.2026013

Social bot detection method based on fine-grained feature weighted expert network

Zhang Huaibo[1,2,3], Gao Jinhua[1,2], Liao Yizhi[1,2,3], Xin Yonghui[4], Cheng Xueqi[1,2]
1. State Key Laboratory of AI Safety, Beijing 100190, China
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
3. University of Chinese Academy of Sciences, Beijing 101408, China
4. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China

Abstract
In recent years, research in the field of social bot detection has gradually evolved from individual feature analysis to group feature mining, and from traditional feature engineering to deep learning methods. Among them, graph network-based methods have shown significant advantages: they can integrate account behavior features, text semantic features, and network topology features, converting social bot detection into a graph node classification task. However, most existing detection methods adopt general models for detection and fail to consider the differences in fine-grained

features among different types of social bots, which limits the detection accuracy across business scenarios. Based on this, a social bot detection method based on a fine-grained feature expert attention mechanism was proposed. The method constructed a business expert network, enabling each expert to focus on learning differentiated weight combinations of fine-grained features. Through multi-expert feature fusion and comprehensive analysis, it achieved integrated detection of potentially multi-type social bots across various business scenarios. Experimental results on a public Twitter dataset for social bot detection demonstrated that this method outperformed existing mainstream detection methods, with a relative improvement of 1.52% in the F1-score.

Key words
social bot detection, fine-grained feature weight, mixture of experts network

0 引言

社交机器人是通过程序自动化或半自动化控制的社交平台账号，社交机器人账号往往活跃在各大社交平台，在信息传播、评价刷单和热度控制方面发挥了较大的作用。恶意的社交机器人被用来发布虚假信息、不良舆情引导、广告营销和网络诈骗，给互联网生态带来负面影响。早在2010年就有学者展开了与社交机器人检测相关的研究，当时聚焦于推特平台自动化账号方面的检测[1]。早期的社交机器人呈现出更高的机械行为特征，其在互联网上发布的内容话术较为固定，发文频次存在明显规律，因此，早期研究侧重通过账号的异常行为特征进行检测。随着互联网数据规模的增大，社交机器人账号逐渐开始进行群体行为，批量账号在某个事件中发布语义相似的引导性帖文，且账号之间通过互关互粉的方式建立一定的影响力。随着机器人检测技术的提升，社交机器人通过隐蔽社交关系和随机关注等方法规避检测，增大了检测的难度。随着大语言模型与更多人工的介入，社交机器人在行为特征和发文内容上呈现出更高的仿真趋势，这对社交机器人检测方法提出了更高的要求。

在社交机器人检测研究领域，传统的基于机器学习的检测方法是应用最广泛的一类。该方法通过构建账号的多维度特征，使用监督学习的方式，形成社交机器人检测分类器。近年来，基于图网络模型的深度学习检测方法成为研究热点。相较于传统的机器学习方法，深度学习模型可以捕捉更丰富的行为特征和文本语义特征，尤其是结合图网络模型后，可以将社交关系特征纳入节点表征中，从而实现良好的检测效果。然而，现有的检测方法仍面临以下问题：一是社交机器人存在不同业务类别，但现有基于图网络模型的检测方法大多采用通用模型，忽略了账号的细粒度特征差异，影响了多业务类别场景下社交机器人的检测效果；二是现有基于专家网络的方法缺乏多特征组合的差异性，难以兼顾模型的可扩展性和可解释性。

为解决基于图网络模型的检测方法存在的问题，本文提出了一种基于细粒度特征权重专家网络（fine-grained feature weighted expert network，FFW-EN）的检测方法。该方法通过特征处理层挖掘账号的细粒度特征，并构建专家网络层来学习细粒度特征的差异化权重组合，最后通过门控层融合多专家输出，得到综合分类研判结果。

本文的主要贡献总结如下。

（1）提出了基于细粒度特征权重的专家注意力结构，并构建了不同专家多特征注意力权重正交损失，提升了专家在多特征组合下的学习能力。

（2）提出了融合多专家输出的混合专家网络结构，通过门控层融合通用网络专家和特征注意力专家的输出，提升了模型在跨业务类别社交机器人融合场景中的检测能力。

（3）提出了基于细粒度特征专家注意力网络的检测方法FFW-EN，并在公开推特数据集上验证了模型的提升效果，分析了模型的可解释性。

1 相关工作

1.1 基于机器学习的方法

基于机器学习的方法主要是通过统计学习方法构建分类模型，通过构建账号特征进行特征表征，然后利用监督学习的方法构建分类器，识别社交机器人账号。Benevenuto等[2]提出了一种基于用户行为和内容特征的监督学习方法，从用户行为和推文内容中提取了62个特征，使用支持向量机模型进行分类，将用户分为垃圾信息发送者和非垃圾信息发送者。Chu等[3]提出了一种自动化分类系统，将推特（Twitter）用户分为3类（人类、机器人和人机结合账户），基于用户行为、内容特征及账户属性设计了一个系统模型，用于捕捉自动化行为的周期性和规律性，并使用线性判别分析对多特征用户类型进行综合判断。Hall等[4]提出了一种基于机器学习的分类器模型，通过行为特征和非正式编辑特征检测未标记的机器人活动，采用随机森林和梯度提升分类器两种集成学习方法，并结合行为模式和修订评论特征进行训练。Yang等[5]对推特垃圾邮件账户的逃避策略进行了全面实证分析，并设计了图基特征、邻居基特征、自动化特征和时序特征4种新的检测特征，在面对垃圾邮件账户的演化逃避策略时，检测精度明显提高。Yang等[6]提出了推特社交机器人检测工具Botometer，基于随机森林与集成分类器提取用户行为、内容、网络等6类特征，已发布多个版本（V1～V4）模型，并持续更新以应对新型机器人。综上，基于机器学习的方法在模型架构层面已经相对成熟，但在账号特征的构建方面仍需要增加大量针对性设计，且随着社交机器人的更新换代，账号特征要不断迭代，才能有效检测具有对应行为特征的社交机器人，否则该方法会缺乏泛化能力。

1.2 基于自然语言处理的方法

基于自然语言处理的文本分析方法是较早发展的技术路径之一。早期研究中，胡舜良等[7]曾提出通过哈希值匹配识别网络机器人的方案，具体是将待检测文本转换为哈希值，再与灌水等不良文本的哈希值进行比对，以此完成机器人账号的初步筛查。沈一等[8]则聚焦于账号昵称特征，提出了一种水军检测方法，该方法通过计算昵称的最小哈希值（minHash）来衡量不同账号间的相似度，进而对相似度较高的昵称群体进行聚合分析，判断其是否符合机器人账号的特征。随着深度学习技术的发展，相关研究逐渐转向基于预训练模型与神经网络的方法。例如，Dukic等[9]构建了基于BERT的社交机器人检测模型，通过引入额外特征与上下文嵌入表示，有效提升了模型的检测性能；Harrag等[10]设计的模型专门针对阿拉伯语文本，利用迁移学习技术识别由GPT-2生成的机器文

本，其核心是捕捉机器生成文本的独特风格特征；Heidari[11]进一步提出了融合情绪分析与心理学知识的集成学习方法，通过深度双向变换器模型实现社交机器人检测，显著增强了模型对文本潜在特征的挖掘能力。以上基于自然语言处理的检测方法在社交机器人识别中展现出一定有效性，但这类方法多依赖单一文本特征，未能充分发挥多特征组合带来的性能增益，仍存在提升空间。

1.3 基于图网络模型的检测方法

基于图网络模型的检测方法主要是将账号特征作为节点表达，将社交关系作为图的边，然后利用图神经网络方法将邻居节点的信息更新至当前节点，从而实现图网络信息的汇聚。Zhao等[12]提出了一种基于多属性异构图卷积网络的社交机器人检测框架Bot-AHGCN，通过构建多属性异构信息网络和基于权重学习的相似性嵌入，将社交机器人检测转化为异构图上的半监督节点分类任务。Feng等[13]提出多重异构图卷积神经网络模型BotRGCN，通过构建账号的文本类信息、数值类信息和属性类信息，以及基于粉丝关系和关注关系的关系图卷积网络，实现了对推特社交机器人的有效检测。Tan等[14]提出了基于社区层级的机器人检测框架BotPercent，专注于估算特定社交网络社区中的机器人百分比，该方法结合了特征、文本和图网络模型，并通过信心校准解决了现有模型的泛化问题。Feng等[15]提出了一种基于异构图关系变换器的推特机器人检测框架RGT，构建了以用户为节点、多样化交互关系为边的异质信息网络，然后设计了关系图变换器来动态捕捉用户间不同类型关系的差异化影响力强度，该模型能有效识别机器人集群的协同行为模式，同时具备优秀的数据效率和表征学习能力。以上基于图网络的社交机器人检测方法多采用通用二分类模型，将节点的多种特征进行拼接后直接处理，未对细粒度特征实施有效的权重调控。这一问题使得其在多业务类别的社交机器人检测场景中的检测效果受限。

1.4 基于专家模型的检测方法

基于专家模型的检测方法是近年来具有创新性的研究方法，目标是利用多专家模型捕捉不同领域信息，从而获取更丰富的融合信息表达，并提升检测指标。Ng等[16]提出多平台社交机器人检测方法BotBuster，该方法通过混合专家（mixture of experts，MoE）架构构建了模块化检测模型，通过动态权重分配机制实现了对不完整数据的鲁棒处理，达到了跨平台泛化检测效果，但该方法存在的问题是专家类型相对固定，缺乏动态扩展的能力。Liu等[17]提出一种用于推特机器人检测的方法BotMoE，该方法构建了基于社区感知的模态特定专家混合架构的多模态联合检测模型，通过专家融合层实现跨模态一致性分析，从而达到了识别特征操纵型机器人的效果，但该方法存在的问题是缺少对特征的细粒度权重控制，且专家在社区感知方面的可解释性有待进一步研究。

2 基于细粒度特征专家注意力网络的检测方法

随着社交机器人在各类媒体平台的广泛应用，其行为模式会因业务目标的不同而呈现出显著差异。这使得通用的模型检测方法难以同时捕捉到特征差异较大的业务类别群体，进而削弱了整体检测效果。

因此，本文提出了一种FFW-EN检测方法：一方面，构建细粒度特征权重专家网络，使单个专家学习不同细粒度特征权重的组合；另一方面，构建混合专家网络层，通过门控机制融合通用网络专家与基于细粒度特征权重的专家，以此提升模型在跨业务类别社交机器人融合场景中的检测能力。

2.1 模型架构

本文提出的FFW-EN模型架构如**图1**所示，该模型分为特征构建层、专家网络层、专家特征融合层和输出层。

特征构建层将账号的元信息数据、发文信息数据以及社交关系数据分别进行处理，尤其针对不同数据类型的特征数据采用不同的处理方式。其中，数值类特征需要经统计后做全局标准化处理；属性类特征需要根据不同的类型选择枚举或布尔类型；文本类特征采用稳健优化的BERT预训练方法（robustly optimized BERT pretraining approach，RoBERTa）获取分类标记（classification token，CLS）编码信息。在不同类型特征处理完成后，将每个特征作为独立的细粒度特征输入专家网络层。

专家网络层包括1个固定的异质网络专家和多个细粒度特征注意力专家。其中，异质网络专家用于学习包含粉丝、关注等社交关系网络特征账号的表征信息；细粒度特征专家根据输入的细粒度特征，自主学习对每个细粒度特征的注意力权重，结合多专家之间注意力权重正交化损失，尽可能学习不同的细粒度特征权重组合。

专家特征融合层是将所有专家的输出特征进行汇总的模块，该模块采用门控权重设计，即每个专家的输出通过学习一个门控权重系数加权获得最终的专家输出结

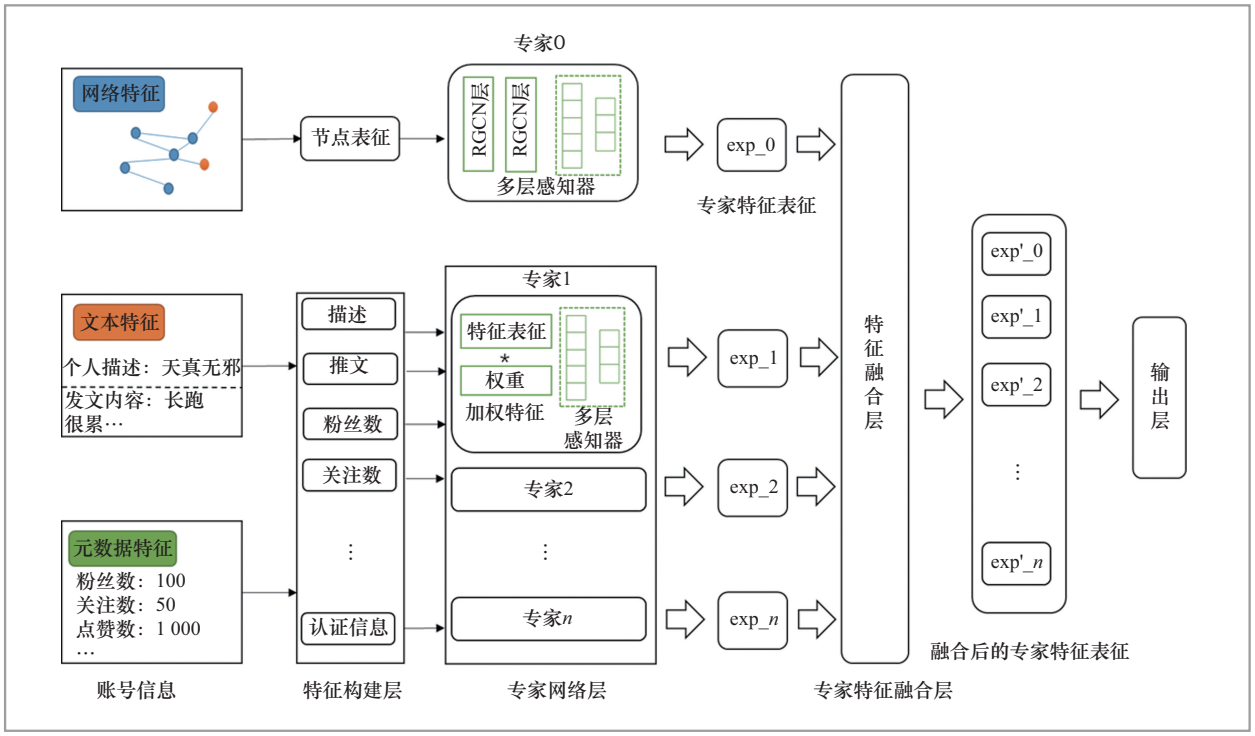


图1 FFW-EN模型架构

果。输入 exp_n 表示第 n 个专家的特征表征，exp′_n 表示经过专家特征融合层后的第 n 个专家的特征表征。

输出层定义了1个全连接层和1个二分类输出层，以输出二分类结果，用于账号标签的最终研判。

2.2 账号特征设计

FFW-EN方法在账号特征设计方面，整合了用户的描述信息、发文信息、元数据信息和社交网络信息等。根据数据类型，这些特征可划分为文本类特征、数值类特征、属性类特征和社交网络关系特征。不同数据类型的账号特征字段存在差异，因此，数据处理和编码位数也不相同。账号特征见**表1**。

在数据特征处理方面，文本类数据特征主要包括用户描述信息和用户推文信息。为了更好地表征文本中的语义信息，本文采用RoBERTa模型对文本信息进行向量表征。其中，用户描述可以直接表征为768维的向量；用户推文信息取账号近期发布的最多50条推文的编码向量，求平均后得到768维向量。为了减少该特征与其他特征的维度差异，采用主成分分析（principal component analysis，PCA）降维，将文本信息的表征维度降低到64维。标准化数据主要涵盖数值类特征，如转发数、评论数、情感值等。由于每个账号的发文数量不同，且每条帖文的数值类特征不同，因此，有必要对账号的数值特征进行标准化处理，以规范数值类数据的表示

表1 账号特征

特征名称	特征描述	编码位数
description	用户自我介绍	64
tweets	用户发布的推文	64
followers_count	粉丝数	1
friends_count	好友数	1
listed_count	此用户所属的公共列表的个数	1
favourites_count	点赞的推文数	1
statuses_count	用户发出的推文的数量	1
screen name length	昵称长度	1
protected	是否处于受保护状态	1
contributors enabled	投稿人模式是否启用	1
is translator	是否开启了“翻译者”功能	1
is translation enabled	用户是否启用了推文翻译功能	1
profile_user_background tile	用户个人资料背景是否设置为平铺模式	1
geo_enabled	用户是否启用地理定位	1
default_profile	是否默认主题或背景	1
default_profile_image	是否默认图像	1
default profile	用户是否使用默认资料的字段	1
verified	用户是否经过认证	1

范围。标准化过程主要包括两个步骤：一是针对账号的所有发文特征求均值，以此表征该账号的特征；二是针对所有账号的特征数值进行标准化计算，使标准化后的数据均值和方差都稳定在一定范围。对单个账号求数值特征的方式为：

$$\boldsymbol{F}(u)=\overline{\sum_{i} M_{ui}} \tag{1}$$

其中，M_{ui}为账号u发布的第i条消息的数值特征表示。当获取到所有账号的数值特征矩阵后，按照式（2）对整个矩阵进行标准化处理。假设μ表示数值特征的均值，σ表示数值特征的方差。标准化处理的方式为：

$$\boldsymbol{F}_z(U)=z\text{score}\left(\boldsymbol{F}(U)\right)=\left[\frac{\boldsymbol{F}(u)-\mu}{\sigma}\right] \tag{2}$$

其中，U代表账号信息，$\boldsymbol{F}(U)$表示账号特征矩阵，$z$score(·)表示对矩阵进行标准化操作。

属性类特征主要是布尔型数据，涵盖账号的属性设置信息，如账号是否处于受保护状态、投稿人模式是否启用、是否使用默认图像以及是否使用默认资料等。在编码过程中，本文用“1”表示True，用“0”表示False，以此实现对属性类特征的编码表达。

2.3 专家层设计

专家层设计主要包含两种类型：第一类是可以动态学习特征权重组合的注意力专家；第二类是固定学习关系图卷积网络信息的专家。第一类专家的输入是账号的全部细粒度特征，第二类专家的输入是账号的特征和社交网络关系。

2.3.1 细粒度特征注意力专家设计

细粒度特征注意力专家模块的架构如图2所示。该模块的输入为账号的全量细粒度特征，所有特征经过1个两层前馈神经网络处理后，得到专家对特征的注意力权重向量，其中前馈神经网络第一层的长度与输入的特征嵌入的长度P保持一致，

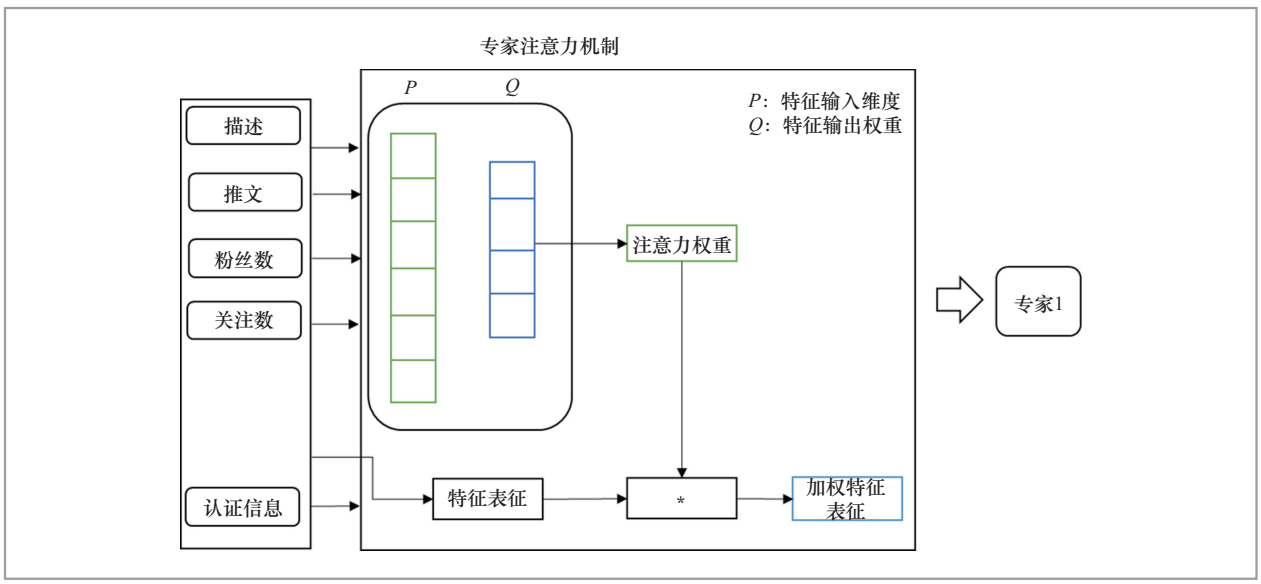


图2 细粒度特征注意力专家模块的架构

第二层输出层的长度与特征数量 Q 保持一致。随后，将特征按照注意力权重进行加权，得到最终的加权输出特征向量。细粒度特征注意力专家在针对不同业务类型场景的社交机器人检测中发挥着重要作用，这是因为不同业务类型的社交机器人往往存在较大的行为特征差异，多专家学习不同重要特征的组合可以更好地捕捉多业务类别社交机器人所处的业务场景特征。

2.3.2 关系图卷积网络专家设计

关系图卷积网络专家模块主要用于学习社交关系网络特征信息。该模块借助账号节点表征和关系图卷积网络将不同社交图网络结构信息更新至节点表达。

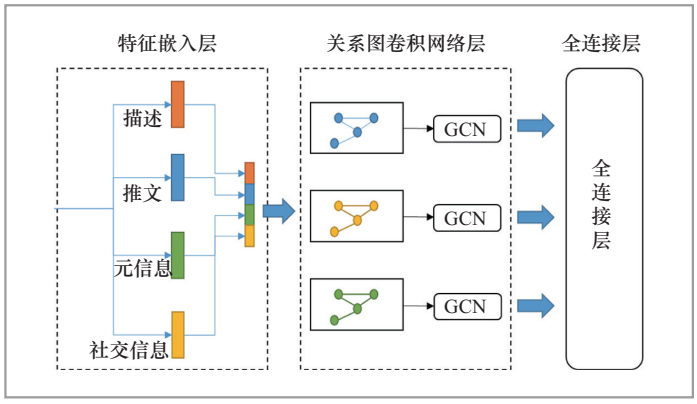


图3 关系图卷积网络专家模块的架构

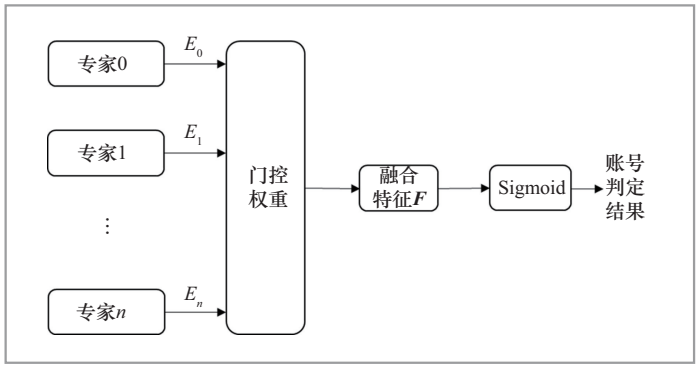


图4 门控输出层模块的架构

关系图卷积网络专家模块的架构如**图3**所示。该模块主要包括特征嵌入层、关系图卷积网络层和全连接层。其中，特征嵌入层将账号的不同特征信息进行拼接，从而实现账号节点特征的表达；关系图卷积网络层通过构建账号的粉丝网络、关注网络等图结构，利用图卷积网络实现不同网络信息的汇聚；最后通过全连接层汇聚全部输出信息，并将向量信息按照专家输出维度进行转换。

2.4 输出层设计

输出层设计采用专家特征门控融合的方式，将多个专家的输出分别学习一个向量权重，从而通过多权重专家向量加权融合的方式得到输出结果，最后通过线性层生成社交机器人检测的二分类输出。

门控输出层模块的架构如**图4**所示。假设专家的输出为 E，则通过专家层得到了专家输出特征向量 $<E_1,E_2,\cdots,E_N>$。门控网络层通过对输入特征进行处理，生成与专家数量对应的权重向量 $<w_1,w_2,\cdots,w_N>$，权重满足归一化条件，即权重之和 $\sum_{i=1}^{N} w_i = 1$，其中 w_i 表示第 i 个专家的权重值。将每个专家的输出向量与对应的权重向量相乘之后，得到融合后的特征向量 $\boldsymbol{F} = \sum_{i=1}^{N} w_i \cdot E_i$。最后，将融合后的特征向量 $\boldsymbol{F}$ 输入线性层，再结合 Sigmoid 激活函数得到正常账号和社交机器人账号二分类的概率结果。

3 实验评估

本节描述了整体实验环境的设置、实验数据集、评价指标和基线方法，并在此基础上，介绍了主流方法的对比实验、消融实验，对实验结果进行了分析。

3.1 实验环境设置

实验环境配置为：在软件框架方面，机器学习相关实验依托PyTorch、Torch-geometric及Scikit-learn实现；CUDA计算套件采用10.2.89版本；硬件采用16核32线程的CPU、128 GB内存，以及NVIDIA Tesla V100 GPU，用于支持高效的并行计算。

3.2 模型参数

3.2.1 超参数设计

实验针对模型的专家数量、RGCN层数以及训练轮次和学习率等指标进行了优化，并采用GridSearch方法选取最优参数。以Twibot-20数据集的专家个数为例，专家数量参数设计实验曲线如**图5**所示。

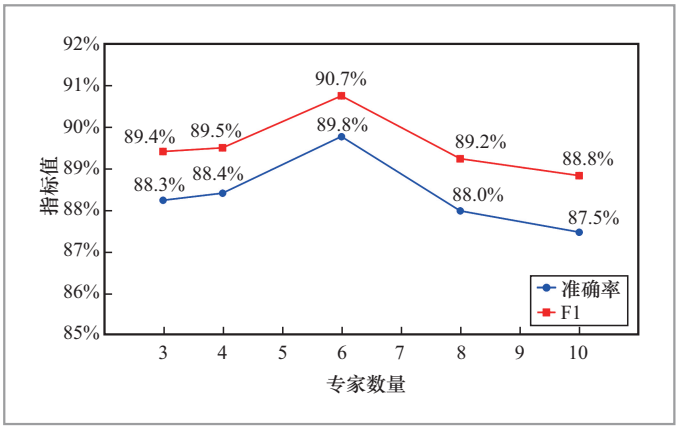


图5 专家数量参数设计实验曲线

FFW-EN模型架构设计中包含1个固定的社交网络专家，至少包含2个细粒度特征权重专家，因此专家数量最小是3，最大数量设定为10。由图5可知，当专家数量为6时，整体的准确率和F1指标最佳。这是因为专家数量较少时，无法充分学习更多的细粒度特征权重组合，专家数量过多时存在的正交损失的约束会使每个专家只关注固定的少量特征，无法发挥多特征权重组合的优势。因此，专家数量需要根据输入的特征数量进行合理设置。按照相似的方法，实验超参数最终确定了训练轮次Epochs为200、RGCN层数为2、学习率LR为1×10^{-3}、每个专家输出特征维度expert_dim为32。

3.2.2 损失函数设计

FFW-EN模型的损失函数包括交叉熵损失和专家正交损失，最终损失函数是交叉熵损失与正交损失的融合损失，本文通过定义超参数λ来平衡两者之间的权重，最终损失函数表示为：

$$\ell_{\text{total}}=\ell_{\text{main}}+\lambda\cdot\ell_{\text{ortho}} \tag{3}$$

损失函数中的第1项ℓ_{main}为交叉熵损失。交叉熵损失用于评估预测数据标签与真实数据标签之间的差距，计算式为：

$$\ell_{\text{main}}=\frac{1}{N}\sum_{i}-\left[y_i\cdot\ln\left(p_i\right)+\left(1-y_i\right)\ln\left(1-p_i\right)\right] \tag{4}$$

其中，N为样本个数，y_i表示样本i的标签，机器人标签为1，非机器人标签为0，p_i表示样本i预测为机器人的概率。

损失函数中的第2项ℓ_{ortho}为专家正交损失。该损失函数计算的核心在于，通过约束不同专家的关注权重向量尽可能正交，让不同专家聚焦于不同的细粒度特征，以提升专家间学习信息的多样性。本文采用Gram矩阵减去单位矩阵的方式设计正交损失，核心是通过约束专家权重向量的Gram矩阵逼近单位矩阵，从而实现专家之间的正交性约束。为了使Gram矩阵的非对角线元素尽可能接近0，采用如下定义。

首先，定义Gram矩阵G，其中元素$G_{ij}=\alpha_i\cdot\alpha_j$，$\alpha_i$代表第$i$个专家关注特征的权

重向量，$i=1,2,\cdots,N$（N为专家个数）。然后，对每个专家的权重向量进行L2归一化，得到单位向量$\widehat{\alpha_i}=\alpha_i/||\alpha_i||_2$，归一化操作是为了避免模长对Gram矩阵元素的干扰。假设专家对特征的关注权重矩阵为A，经过归一化后的权重矩阵为$\hat{A}$，由于$\widehat{\alpha_i}$是L2归一化后的单位向量，满足$\widehat{\alpha_i}\cdot\widehat{\alpha_i}=1$，因此$\hat{A}$归一化后的Gram矩阵$\hat{G}$的对角线元素为1。最后，正交损失可定义为Gram矩阵与单位矩阵I差值的范数，即：

$$\ell_{\text{ortho}}=\left\|\hat{G}-I\right\|_{\text{F}}^{2}=\sum_{j=1}^{N}\left(\widehat{G}_{ij}-I_{ij}\right)^{2} \quad (5)$$

其中，I为$N\times N$的单位矩阵。

3.3 对比实验

本节主要对FFW-EN方法与主流的社交机器人检测方法进行对比，并介绍相关数据集和评价指标。

3.3.1 实验准备

本文在基线方法上选取了近年来主流的社交机器人检测方法，涵盖了经典的社交机器人检测方法SGBot[18]、文献[19]方法、RoBERTa[20]、文献[21]方法、SATAR[22]、Botometer等，基于关系图卷积网络的检测方法BotRGCN、RGT，以及基于专家网络的检测方法BotBuster、BotMoe等。

本文采用的数据集为公开的社交机器人推特数据集Cresci-15和Twibot-20，两个数据集的数据规模适中且具备多种社交网络关系信息。数据集按照训练集60%、验证集20%、测试集20%的比例进行划分，在训练过程中每项实验结果采用5次实验取平均值的方式，尽可能减小参数波动带来的影响。为了更好地评价检测方法的效果，选取了准确率和F1值两个指标。F1值用来综合评价模型的精确率和召回率，更加贴近社交机器人检测的业务效果。

3.3.2 实验分析

本文将FFW-EN方法与主流的10种方法进行实验对比，实验结果见**表2**。

在Cresci-15数据集上的实验结果显示，FFW-EN方法在准确率和F1值上的表现均达到了最优，准确率达到99.44%，F1值达到99.56%；其次是BotMoe方法，F1值为98.82%；再之后是RGT方法和BotBuster方法。该实验结果说明基于专家网络的方法在该数据集上取得的效果最优，其次是基于关系图卷积网络的方法。

在Twibot-20数据集上的实验结果显示，FFW-EN方法在准确率和F1值上的表现仍然是最优的，准确率达到89.77%，F1值达到90.74%，相较于次优的BotMoe，其准确率提升2.01%，F1值提升1.52%。这说明基于细粒度专家注意机制的FFW-EN方法在两个推特数据集上的指标比较理想，优于实验中的其他对比方法。

3.4 消融实验

消融实验的设计主要考虑两部分内容：一是每个专家的细粒度特征权重带来的增益效果；二是多专家之间的正交损失带来的增益效果。因此，实验分别对比了去除对应部分与完整的FFW-EN模型。考虑数据集的规模和多样性，实验采用Twibot-20数据集。

3.4.1 基于细粒度特征权重的专家设计

实验对比了完整的FFW-EN模型和去除细粒度特征权重的情况。其中，去除细粒度特征权重是将原始的特征进行拼接后直接输入专家层进行处理，不具备特征权重的加权信息。FFW-EN专家注意力消融实验结果见**表3**。

由**表3**可知，去除细粒度特征权重后，FFW-EN模型的准确率和F1值均有一定程度的下滑，准确率降低1.18%，F1值降低0.98%，这表明细粒度特征权重专家网络对模型指标起到了增益作用，也侧面验证了数据集中存在特征差异化的账号信息。

3.4.2 专家正交损失

实验对比了完整的FFW-EN模型和去除专家正交损失的情况。其中，FFW-EN模型采用交叉熵损失+专家正交损失，后者仅采用交叉熵损失。

表4中，TotalLoss代表FFW-EN模型采用交叉熵损失和正交损失的组合，在去除交叉熵损失后，实验的准确率和F1值均有一定程度的下滑，准确率降低1.69%，F1值降低1.43%，这表明增加专家正交损失可以让不同专家的组合增益更高。

为了进一步验证专家正交损失带来的实际增益效果，本文对带正交损失的特征权重Top5和无正交损失的特征权重Top5进行了分析。带正交损失和无正交损失的特征权重Top5分别如**图6**、**图7**所示。由**图6**可知，增加专家正交损失可以让不同专家尽可能学习到多样化的关键特征组合。固定专家0仅关注RGCN表征信息，专家1更关注认证信息和默认头像信息，专家2更关注推文信息，专家3更关注地理位置和用户描述信息，专家4更关注扩展信息和是不是贡献者，专家5更关注用户活跃天数和背景图片信息。这使每个专家都能学习到账号与该部分关键信息相关的内容，有利于提升存在多种业务类别的社交机器人的融合检测能力。

由**图7**可知，在去除专家正交损失，仅采用交叉熵损失的情况下，专家1、专家2和专家5都对账号描述信息有较强的依赖，专家3和专家4对推文信息有较强的依赖。这是因为多个专家学习到的信息特征存在趋同，不利于学习更多的特征组合信息。

表2 FFW-EN方法与主流方法的实验结果

方法	Cresci-15数据集		Twibot-20数据集	
	准确率	F1	准确率	F1
SGBot	77.10%	77.91%	81.60%	84.90%
文献[19]	96.10%	82.65%	71.30%	57.33%
RoBERTa	96.90%	95.86%	75.50%	73.09%
文献[21]	93.20%	94.73%	78.70%	81.08%
SATAR	93.40%	95.05%	84.00%	86.07%
Botometer	57.90%	66.90%	53.10%	53.13%
BotRGCN	96.50%	97.30%	85.70%	87.25%
RGT	97.20%	97.78%	86.60%	88.01%
BotBuster	96.90%	97.53%	77.24%	81.18%
BotMoe	98.50%	98.82%	87.76%	89.22%
FFW-EN	**99.44%**	**99.56%**	**89.77%**	**90.74%**

表3 FFW-EN专家注意力消融实验结果

模型	准确率	F1
FFW-EN	89.77%	90.74%
-FeatureAttention	88.59%	89.76%

注：“-”表示在上一个方法的基础上减去该项。

表4 FFW-EN正交损失消融实验结果

模型	准确率	F1
TotalLoss	89.77%	90.74%
-OrthogLoss	88.08%	89.31%

注：“-”表示在上一个方法的基础上减去该项。

3.5 全局特征可解释性分析

FFW-EN模型的优势在于，可以获取

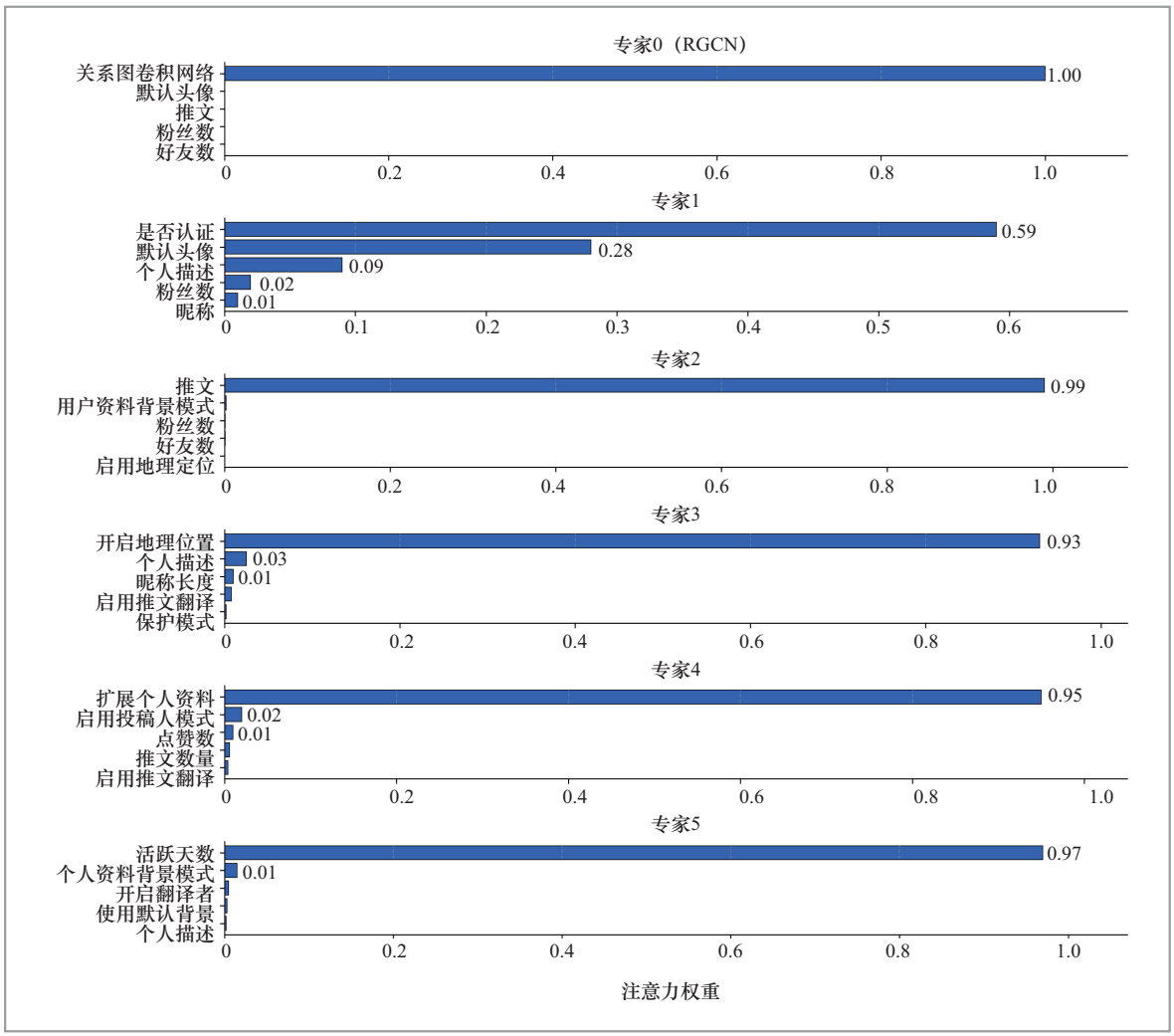


图6 带正交损失的特征权重Top5

每个专家对细粒度特征的权重，同时获取到专家融合层的每个专家的权重，从而在特征和专家两个方面具备可解释能力。3.4节验证了专家对特征采用的多样性组合学习策略。该部分通过计算全局特征分数来验证所有输入特征的重要性。

专家输出权重占比如图8所示。由图8可知，专家2在所有专家特征融合中的贡献最高，达到了51.2%，其次是专家1（25.9%）和专家0（11.6%）。由于每个专家专注的特征组合不同，不同专家的融合特征在输出结果研判中起到的作用存在差异，为了进一步分析所有细粒度特征的综合重要性，本文定义了特征重要性指标，计算式为：

$$\mathrm{WF}[j]=\sum_{i}^{E}\sum_{j}^{F}\mathrm{EA}[i][j]\times\mathrm{score}[i] \quad (6)$$

其中，i表示专家模块索引（从0到E，E为专家数量），j表示特征索引，WF[j]表示第j个专家的权重，EA[i][j]表示特征权重得分，即第i个专家对第j个特征关注权重，score[i]为第i个专家对应的门控分数。最终对所有特征的加权重要性得分做归一化处理：

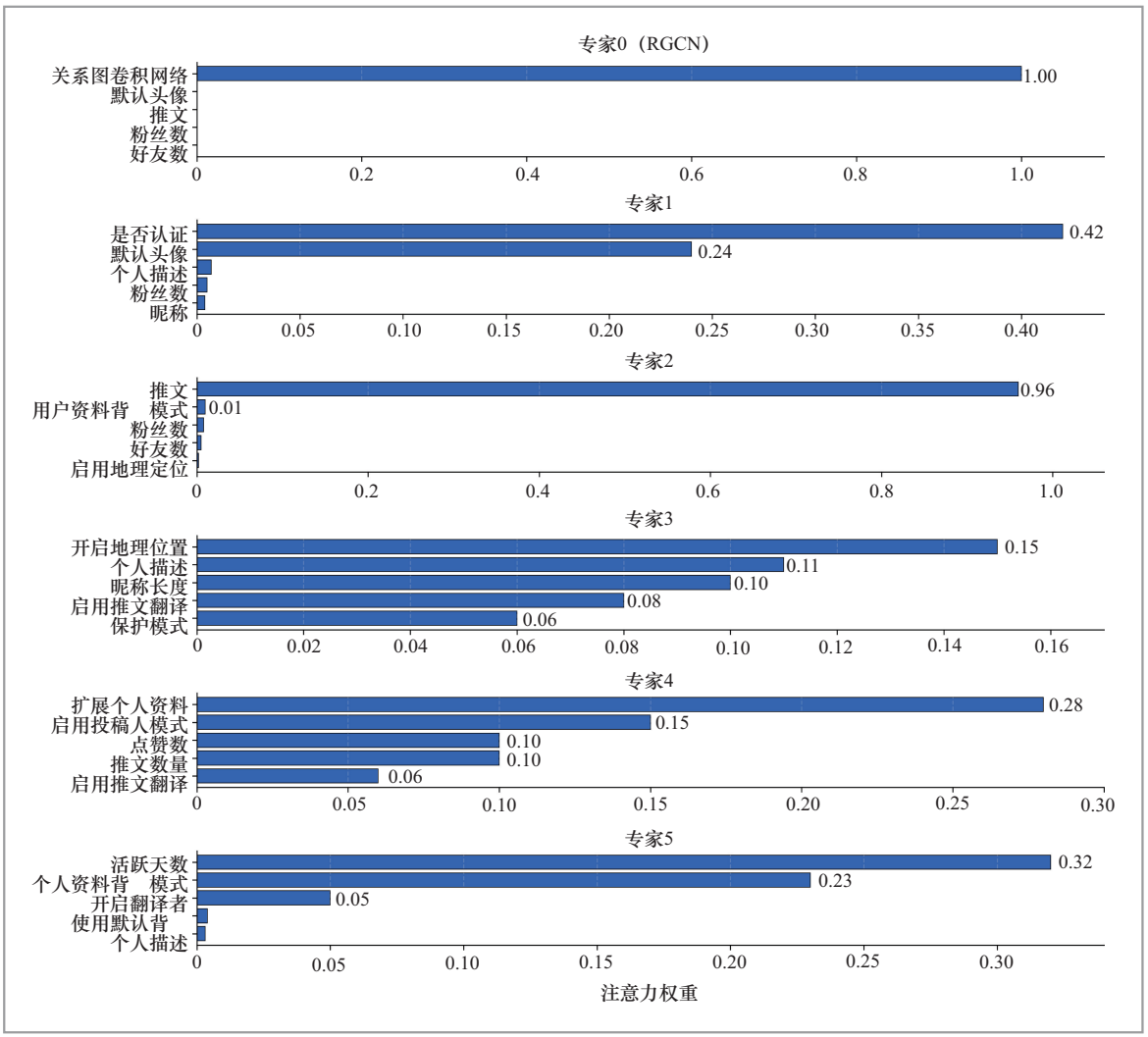


图7 无正交损失的特征权重Top5

$$F[j]=\mathrm{WF}[j]/\sum_{j}^{F}\mathrm{WF}[j] \qquad (7)$$

其中，$F[j]$为各个特征的全局重要性得分，反映第j个特征在模型决策过程中的相对重要程度。

模型在评估过程中的全局特征重要性占比Top10见**表5**。由**表5**可知，所有细粒度特征中推文信息特征的重要性占比最高，其次是账号验证信息、地理位置信息、账号默认头像信息、账号描述和活跃时长等，这些特征按照重要程度与实际人工研判过程中的重要性基本相似，这为重点特征的筛选提供了理论依据。

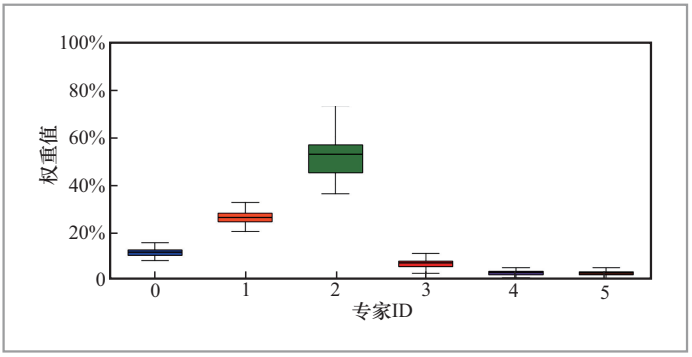


图8 专家输出权重占比

表5 全局特征重要性占比Top10

特征	重要性占比
tweet	57.49%
verified	17.28%
geo_enabled	9.37%
default_profile_image	8.38%
description	2.89%
active_days	1.76%
has_extended_profile	0.94%
followers_count	0.38%
screen_name_length	0.27%
profile_use_background_image	0.22%

4 结束语

为解决现有通用模型在检测多业务类别社交机器人中关注特征差异较大而导致检测效果不佳的问题，本文提出了基于细粒度特征权重专家网络的FFW-EN方法。该方法通过构建细粒度特征权重专家网络模型，结合混合专家网络搭建了多特征注意力权重的专家网络层，并通过门控层融合通用网络专家与多个基于特征权重的专家，有效提升了模型在跨业务类别社交机器人融合场景下的检测能力。实验结果表明，FFW-EN方法在社交机器人公开数据集上均取得了良好的效果，对实验结果的分析充分验证了该方法的有效性和可解释性。

未来工作将从以下方面展开深入探索：首先，在细粒度特征划分层面，可进一步结合特征的来源属性和不同特征类型的组合模式，进一步提升模型对复杂特征组合的捕捉能力；其次，在专家协作机制上，进一步研究专家间的动态协作模式，通过自适应专家数量、动态调整专家融合权重等方式，提升模型对多样化输入数据的处理效率与表达能力；最后，当前实验主要基于现有的公开推特数据集，后续可在更多真实社交平台数据场景中进行应用验证，以检验该方法在多平台的泛化能力。

参考文献：

[1] Panagiotis T M, Mus-Tafaraj E. From obscurity to prominence in minutes: political speech and real-time search[J]. WebSci10: Extending the Frontiers of Society Online, 2010: 1-7.

[2] Benevenuto F, Magno G, Rod-Rigues T, et al. Detecting spammers on Twitter [C]//Proceedings of the 7th Annual Collaboration, Electronic Messaging, Antiabuse and Spam Conference (CEAS). New York: ACM, 2010, 6: 12.

[3] 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. New York: ACM, 2010: 21-30.

[4] Hall A, Terveen L, Halfaker A. Bot detection in wikidata using behavioral and other informal cues[J]. Proceedings of the ACM on Human-Computer Interaction, 2018, 2: 1-18.

[5] 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.

[6] Yang K C, Ferrara E, Menczer F. Botometer 101: social bot practicum for computational social scientists[J]. Journal of Computational Social Science, 2022, 5(2): 1511-1528.

[7] 胡舜良，凌宝红，胡东辉. 基于文本认知Hash的网络水军发帖检测技术研究[J]. 安庆师范学院学报(自然科学版)，2012，18(2)：69-72.

Hu S L, Ling B H, Hu D H. Research on the technology of detection the posts of "net navy" based on text's Hash[J]. Journal of Anqing Teachers College (Natural Science Edition), 2012, 18(2): 69-72.

[8] 沈一, 鲍新平. 一种用户标识识别方法和装置: CN106095813A[P]. 2016-11-09.
Shen Y, Bao X P. A method and device for user identification: CN106095813A [P]. 2016-11-09.

[9] Dukić D, KečA D, Stipić D. Are you human? Detecting bots on Twitter using BERT[C]//Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Piscataway: IEEE Press, 2020: 631-636.

[10] Harrag F, Debbah M, Darwish K, et al. BERT transformer model for detecting Arabic GPT2 auto-generated tweets[J]. arXiv preprint, 2021: 2101.09345.

[11] Heidari M. Ensemble learning of deep bidirectional transformers for emotion and psychology analysis of textual data for social media bot detection[D]. Fairfax: George Mason University, 2022.

[12] Zhao J, Liu X D, Yan Q B, et al. Multi-attributed heterogeneous graph convolutional network for bot detection[J]. Information Sciences, 2020, 537: 380-393.

[13] Feng S B, Wan H R, Wang N N, et al. BotRGCN: Twitter bot detection with relational graph convolutional networks [J]. arXiv preprint, 2021: 2106.13092.

[14] Tan Z X, Feng S B, Sclar M, et al. BotPercent: estimating bot populations in twitter communities[J]. arXiv preprint, 2023: 2302.00381.

[15] Feng S B, Tan Z X, Li R, et al. Heterogeneity-aware Twitter bot detection with relational graph transformers[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(4): 3977-3985.

[16] Ng L H X, Carley K M. BotBuster: multi-platform bot detection using a mixture of experts[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2023, 17: 686-697.

[17] 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.

[18] Yang K-C, Varol O, Hui P M, et al. Scalable and generalizable social bot detection through data selection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 1096-1103.

[19] Wei F, Nguyen U T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings[C]//Proceeding of the 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). Piscataway: IEEE Press, 2019: 101-109.

[20] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv preprint, 2019, arXiv: 1907.11692.

[21] 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. 2017, 11(1): 280-289.

[22] Feng S B, Wan H R, Wang N N, et al. SATAR: a self-supervised approach to Twitter account representation learning and its application in bot detection[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 3808-3817.

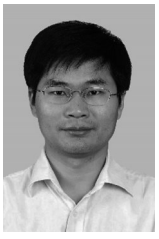
作者简介

张怀博（1990– ），男，中国科学院大学博士生，中国科学院计算技术研究所工程师，主要研究方向为社交机器人检测、图神经网络、数据挖掘。

高金华（1989– ），男，博士，中国科学院计算技术研究所高级工程师，主要研究方向为自然语言理解、社交媒体分析。

廖逸之（1997– ），男，中国科学院大学硕士生，主要研究方向为社会网络分析。

辛永辉（1990– ），男，博士，国家计算机网络应急技术处理协调中心工程师，主要研究方向为网络安全、信息安全。



程学旗（1971– ），男，博士，中国科学院计算技术研究所研究员，主要研究方向为数据科学与大数据分析系统、网络科学与社会计算、Web搜索与挖掘。

收稿日期: 2025-11-17

通信作者: 辛永辉,xinyh@cert.org.cn