引用格式：屈津, 杜姝颖, 宫继兵, 等. 基于多模态信息不变和特定表示的社交机器人检测方法 [J]. 电子科技大学学报, 2026, 55(2): 232-243.
QU J, DU S Y, GONG J B, et al. Social bot detection method based on multimodal information invariant and specific representation[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(2): 232-243.

基于多模态信息不变和特定表示的社交机器人检测方法

屈 津[1]，杜姝颖[1]，宫继兵[1,2]，彭吉全[1]，王开宇[1,2*]

（1. 燕山大学人工智能学院（软件学院），秦皇岛 066004；2. 燕山大学河北省计算机虚拟技术与系统集成重点实验室，秦皇岛 066004）

摘要：社交机器人在发展过程中不断进化，给现有检测模型带来挑战。为此，提出了一种新的社交机器人检测框架BotSAI。该框架首先采用定制化编码器从用户元数据、文本和社交网络异构图中提取多维度特征表示。其中，图编码器通过过采样和局部关系转换器实现邻居信息的高效均衡聚合。随后，利用多通道表示器将用户表示映射至不变子空间和特定子空间以增强其特征。最后，通过自注意力机制对增强后的用户表示进行整合与提炼。实验结果表明，BotSAI 在两大权威社交机器人检测基准上均超越了现有方法。此外，系统性实验揭示了不同社交关系对检测精度的影响，为社交机器人检测提供了新的研究视角。

关键词：社交机器人检测；社交网络异构图；不变与特定子空间；自注意力机制
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2024277

Social bot detection method based on multimodal information invariant and specific representation

QU Jin[1], DU Shuying[1], GONG Jibing[1,2], PENG Jiquan[1], and WANG Kaiyu[1,2*]

(1. School of Artificial Intelligence (School of Software), Yanshan University, Qinhuangdao 066004, China;
2. Key Laboratory of Computer Virtual Technology and System Integration of Hebei Province, Yanshan University, Qinhuangdao 066004, China)

Abstract: Social bots have continuously evolved during their development, posing significant challenges to existing detection models. To address this, we propose a novel social bot detection framework, named BotSAI. This framework first employs customized encoders to extract multi-dimensional feature representations from user metadata, text, and heterogeneous social network graphs. Specifically, the graph encoder achieves efficient and balanced aggregation of neighborhood information through oversampling and a local relation transformer. Subsequently, a multi-channel representor maps user representations into invariant subspaces and specific subspaces to enhance their features. Finally, the enhanced user representations are integrated and refined using a self-attention mechanism. Experimental results demonstrate that BotSAI outperforms state-of-the-art methods on two authoritative social bot detection benchmarks. Furthermore, systematic experiments reveal the impact of different social relationships on detection accuracy, providing new research perspectives for social bot detection.

Key words: social bot detection; heterogeneous graph of social networks; invariant and specific subspaces; self-attention mechanism

随着互联网技术的迅猛发展和移动互联网设备的普及，国内外出现了众多社交网络平台，社交网络用户的数量也在持续增加。在这个过程中，出现了一类由程序自动控制并能模仿真实用户行为的社交机器人账户，其中部分机器人账户具有虚假性质。这些机器人账户通过传播错误信息[1-2]、干扰信息传播秩序[3-4]、发起隐私侵犯行为[5]及扩散低质量内容[6]等方式影响舆论方向，对在线社区的稳定运行构成挑战。因此，迫切需要有效的社交机器人检测方法来减轻其负面影响。

收稿日期：2024−10−11
基金项目：河北省创新能力提升计划资助项目（22567626H）；中央引导地方科技发展资金项目（246Z0306G）
作者简介：屈津，主要从事社交网络与安全方面的研究。
*通信作者 E-mail: wangky@ysu.edu.cn

社交机器人在养殖与检测之间已经开始了激烈的博弈。早期使用随机信息创建的简单社交机器人，通过发布重复和虚假的内容来产生舆论影响。社交机器人检测模型通过从元数据中提取的用户属性[7]、用户活动时间线[8]等特征输入传统分类器可以轻松进行检测[9-10]。随后，社交机器人通过篡改用户的元数据逃避传统检测方法，研究人员则基于语义来识别恶意推文，完成检测任务。文献 [11] 提出基于 STT-RNN 模型，用于联合建模文本、时间和空间数据，通过门控循环单元 GRU 处理序列数据，捕捉文本的上下文结构生成特征向量进行识别。文献 [12] 使用预训练的 BERT-base 模型生成推文的上下文嵌入，提取特定标记的嵌入作为主要特征，结合表情符号、转发、URL 等特征，构建混合特征向量进行社交机器人识别。随着生成式语言模型（如 GPT 系列）的快速发展，社交机器人能够生成语法正确、逻辑连贯且情感表达自然的文本，其语义层面与人类用户差异逐渐模糊，仅依赖推文内容分析可能导致漏检。

近年来，图论领域涌现了大量基于深度学习的研究，研究者利用社交用户及用户间社交关系构建图形结构。文献 [13] 提出一种 BotRGCN 框架，利用多模态用户信息构建异构图，并应用关系图卷积网络 R-GCN 来捕捉用户间的社区行为。文献 [14] 通过构建包含多样化关系的异质信息网络 HIN，利用关系图 Transformer 建模用户间的影响强度，并结合语义注意网络聚合用户和关系的信息，实现异质关系感知的社交机器人检测。当前基于图结构的方法，虽相比基于文本的方法具有更好的检测性能和更强的泛化能力[15-17]，但仍面临多重局限。1）主流数据集（如 MGTAB[16]、TwiBot-20[18]）存在严重类别失衡问题，导致模型对多数类过度拟合，现有方法缺乏有效的偏差矫正机制；2）现有社交关系建模过度简化，仅依赖关注和被关注二元关系，忽视了提及、回复和引用等多维动态交互模式，这些模式对用户行为表征具有关键作用；3）目前图论中最优框架 RGT[14] 提出的语义注意力机制，采用全局统一权重为同一社交关系类型中的邻居节点分配注意力权重，未能充分考虑同类型社交关系中邻居节点对目标用户的异质性影响。

当今高级社交机器人通过窃取真实用户描述信息构建仿生账户，并实施关注、评论等交互伪装行为，深度渗透至真实用户社交网络[19]，从而规避传统检测并扩大影响力。在此背景下，多模态协同机制通过融合异构数据有效识别跨模态伪造冲突，已成为提升检测鲁棒性的必然路径。文献 [20] 提出了 BIC 检测框架，BIC 不仅分别对文本和社交关系网络信息进行建模，还通过交互模块和语义一致性模块让这 2 种模态的信息在学习过程中相互作用，以全面评估社交媒体上的机器人活动。文献 [21] 提出了一种基于社区感知混合专家的多模态社交机器人检测框架 BotMoE，该框架通过特定编码器提取用户元数据、文本内容和社交关系网络 3 种模态特征，采用混合专家层 MoE 及门控网络将用户动态分配到不同社区专家，以捕获社区间机器人行为差异。虽然上述方法在多模态交互与语义一致性建模上展现了创新性，但未对齐的模态表示与浅层交互机制，显著限制了其检测性能。因此，本文致力于研究多模态表征学习中的特征解耦方法，通过将异构模态信息映射到共享不变子空间和模态特定子空间的联合表征框架，构建鲁棒性更强的多模态融合机制以应对对抗性新场景的挑战。

在跨域视觉表征领域，DSN[22] 提出双流架构，将图像特征解耦为私有子空间（捕获领域特异性模式）和跨域共享子空间（提取领域不变性特征），有效缓解了跨域图像分类任务中训练集与测试集分布不一致的问题。针对多任务文本分类的噪声干扰问题，文献 [23] 设计分层特征提取网络，采用共享编码器与任务特定编码器的协同架构，缓解了传统多任务学习中共享特征易受任务间噪声干扰的问题。在跨模态情感计算方向，MISA[24] 通过模态不变和特定子空间的双重映射机制，实现视频、语音与文本模态的信息增强与深度交互，其解耦式表征学习方法在情感识别准确率上取得突破性提升。

基于多模态表征学习中模态不变子空间与模态特定子空间映射范式在多个领域的成功实践，本文面向社交机器人检测任务，提出了一种新型的多模态解耦检测框架 BotSAI，该框架采用了不变和特定子空间感知的混合架构。

1　社交机器人检测模型

图 1 展示了本文提出的社交机器人检测框架 BotSAI，该框架具有多模态用户信息编码及模态不变与特定子空间映射的可交互学习能力。具体来说，首先对用户的元数据、文本内容和社交网络异构图信息进行特征提取，分别使用重新设计的元数据编码器、文本编码器和图编码器，从不同角度获取全面的社交平台用户表示。其次，将提取的每部

分特征分别映射到模态不变子空间和模态特定子空间。随后，对增强的特征向量表示执行多头自注意力机制进行交互融合，使每个向量能够学习到其他跨模态和跨子空间的表示。通过这种融合方式，每个表征可以从其他表征中诱导潜在信息，这些信息对于整体的社交机器人检测任务具有协同作用。最后，将输出的融合特征向量串联起来，使用构造出的联合向量用于检测任务并学习模型参数。

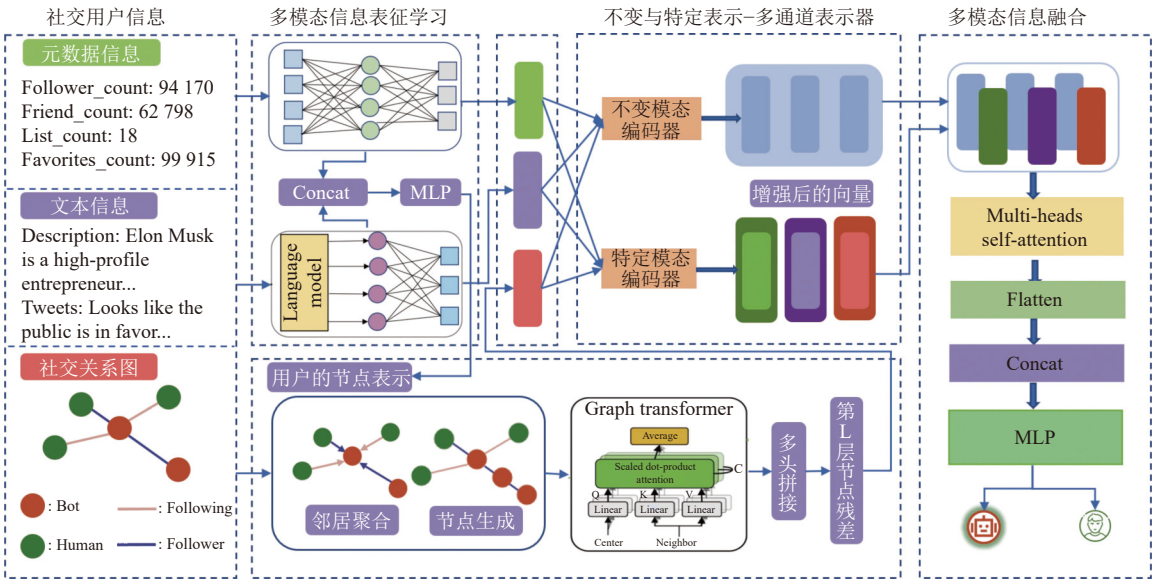


图 1 社交机器人检测框架 BotSAI 概述

1.1 问题定义

本文利用多模态信息来进行社交机器人检测，对于检测任务的输入包括 3 部分特征序列，分别是图形模态特征G、文本模态特征T和元数据特征M。由用户社交关系形成的异构图定义为：设$G=G(U,E,\varphi,R^e)$，其中$U=\{u_1,u_2,\cdots,u_n\}$是用户的集合，E为边集，$\varphi:E\to R^e$表示关系映射函数，R^e是关系类型集。社交平台用户u_v的邻居可以从G导出记为$N_v=\{u_{v,w}\}_{w=1}^{W_v}$，其中$W_v$为邻居的数量。文本模态包含用户$u_v$的第$a$条描述信息记为$B_v=\{t_v^a\}_{a=1}^{Q_a}$及第$b$条推文信息记为$T_v=\{t_v^b\}_{b=1}^{Q_b}$，其中$Q_a$和$Q_b$表示总数量。目标是寻找检测函数$f:f(u_v)\to\hat{y}\in\{0,1\}$使得$\hat{y}$近似于真实标签值$y$，以使得其预测精度最大化。BotSAI 框架分为 2 个阶段：多模态信息表征学习阶段和多模态信息交互阶段。

1.2 多模态信息表征学习

社交机器人操纵者常利用网络爬虫获取真实用户的元数据和文本信息，创建高度混淆账号，这使得单凭文本信息难以区分机器人与真人用户。虽然引入用户间的社交网络关系可显著提升机器人的可区分性，但单纯依赖网络结构可能忽视文本内容的重要性，而后者正是社交机器人传播不实信息的主要手段。因此，本文提出结合用户元数据、文本内容及社交网络的异构图信息，构建用户全方位表示模型，以提高社交机器人检测的准确性。

1.2.1 元数据模态编码器

对于来自用户配置文件的元数据，基于可用性和简单性考虑 5 个量（关注者、追随者、收藏、活跃天数、用户名长度）和 3 个分类元数据（用户名、时间戳、默认配置文件图像）。将这些特征嵌入记为$\boldsymbol{m}_v$，进行 Z-Score 归一化，随后应用 2 层 MLP来学习用户u_v的元数据的特征表示记为$\boldsymbol{x}_v^M$：

$$\boldsymbol{x}_v^M=\mathrm{MLP}(\boldsymbol{m}_v) \tag{1}$$

1.2.2 文本模态编码器

本文使用预训练的语言模型对社交平台用户推文及描述信息进行编码，为了公平起见，利用与其他基线模型相同的 RoBERTa[25] 作为文本模块的编码器，编码过程如下所示：

$$\boldsymbol{x}_v^T=\frac{1}{n}\sum_{a=1}^{n}\frac{1}{l_a}\sum_{b=1}^{l_a}\mathrm{LM}(\boldsymbol{t}_v)^{a,b} \tag{2}$$

式中，LM表示预训练语言模型编码器；$\boldsymbol{t}_v$表示用户u_v的文本信息；a表示推文数据中的第a条推文；b表示描述信息中的第b条属性；n表示文本内容总数；l_a表示第a条输入文本序列的长度。设用户u_v的推文信息提取出的特征记为$\boldsymbol{t}_v^a$，用户u_v的描述信息中提取的特征记为$\boldsymbol{t}_v^d$，并将其输入一个两层的MLP中，得到用户u_v低维的文本表示$\boldsymbol{x}_v^T$。

1.2.3　图模态编码器

本文通过用户间多种社交关系构建关系异构图，以捕获用户间的复杂社交联系。社交关系网络图中的节点初始信息使用用户元数据、个人描述和推文数据特征构造组成，从而获得全面的用户节点嵌入，节点u_v的原始特征属性表示为：

$$\boldsymbol{h}_v = \boldsymbol{W}_D(\boldsymbol{x}_v^M \| \boldsymbol{x}_v^T) + \boldsymbol{b}_D \tag{3}$$

1）图节点过采样。对于机器人账号和真实用户账号数量不平衡问题，通过图神经网络将社交图用户节点的领域节点特征映射到特征空间，在特征空间采用过采样的方式为少数类生成节点样本，来处理数据集的不平衡问题。首先使用 2-Hop 邻域聚合，通过邻域聚合的信息表示为$\boldsymbol{q}_v$。由于用户社交关系网络中存在机器人账号藏匿在与多个真实用户交互中，为防止通过 2-Hop 领域聚合得到表示过度平滑进而藏匿的机器人账号，所以将$\boldsymbol{h}_v$和$\boldsymbol{q}_v$进行连接作为用户节点的嵌入表示，即：

$$\boldsymbol{x}_v = [\boldsymbol{q}_v \| \boldsymbol{h}_v] \tag{4}$$

式中，$\boldsymbol{x}_v$为用户节点u_v的嵌入表示。通过这种连接原始节点信息和聚合的邻居信息的方式，既保留了原始节点信息，也使这些机器人节点的嵌入与人类节点相似。随后，利用合成节点生成器[26]来生成少数类节点，可以通过设置超参数ω（过采样尺度）控制少数类生成的节点数量。$\Gamma(\cdot)$表示特征空间中以欧氏距离测量的少数类u_v的k近邻的集合。从$\Gamma(u_v)$中选择一个与u_v具有相同标签的随机节点u_w，在u_v和u_w之间的直线上选择一个随机点为u_k，过程如下所示：

$$\boldsymbol{x}_k = (l-\delta)\boldsymbol{x}_v + \delta \boldsymbol{x}_w \tag{5}$$

式中，$\boldsymbol{x}_k$是生成虚拟节点在特征空间中的嵌入；δ是一个随机变量，服从均匀分布，取值范围从 0 到 1。这种方法生成了分布均匀的少数类样本$\boldsymbol{x}_k$，其中$\boldsymbol{x}_k$与$\boldsymbol{x}_v$、$\boldsymbol{x}_w$具有相同的标签。

2）局部关系图转化器。在社交网络中广泛存在影响的异质性[14]、全局关系异质性[14]和局部关系异质性。本文所指的局部关系异质性，是在同一类社交关系中，用户因不同邻居节点的影响而表现出的异质性，通过对这种异质性的建模与应用，能够有效捕捉用户行为特征中的个性化差异，从而提升对用户行为的理解和分析能力。受 Transformer[27]在自然语言处理方面的成功启发，本文提出了局部关系图转化器，使用一种包含 Transformer 在异构信息网络上运行的 GNN 架构，去建模影响异质性及局部关系异质性。本文使用多头注意力机制来聚合计算社交关系异构图中中心节点与邻居节点之间的影响，首先对中心节点特征$\boldsymbol{x}_v^{(l-1)}$和邻居节点特征$\boldsymbol{x}_u^{(l-1)}$进行线性变化，得到的 GNN 第l层基于关系r的第i个注意头的查询、键和值，即：

$$\begin{aligned}
\boldsymbol{Q}_v^{(l),r,i} &= \boldsymbol{W}_Q^{(l),r,i}\boldsymbol{x}_v^{(l-1)} + \boldsymbol{b}_Q^{(l),r,i} \\
\boldsymbol{K}_u^{(l),r,i} &= \boldsymbol{W}_K^{(l),r,i}\boldsymbol{x}_u^{(l-1)} + \boldsymbol{b}_K^{(l),r,i} \\
\boldsymbol{V}_u^{(l),r,i} &= \boldsymbol{W}_V^{(l),r,i}\boldsymbol{x}_u^{(l-1)} + \boldsymbol{b}_V^{(l),r,i}
\end{aligned} \tag{6}$$

式中，$\boldsymbol{W}_*^{(l),r,i}$和$\boldsymbol{b}_*^{(l),r,i}$为可学习参数；$\boldsymbol{Q}_v^{(l),r,i}$、$\boldsymbol{K}_u^{(l),r,i}$、$\boldsymbol{V}_u^{(l),r,i}$作为 Transformer 的查询、键以及值。随后，通过计算中心节点与其邻居节点之间的注意力分数，衡量邻居节点对中心节点的重要性，即:

$$\boldsymbol{e}_{vu}^{(l),r,i} = \frac{\boldsymbol{Q}_v^{(l),r,i}\boldsymbol{K}_u^{(l),r,i\mathrm{T}}}{\sqrt{d_k}} \tag{7}$$

式中，$\boldsymbol{e}_{vu}^{(l),r,i}$为中心节点$\boldsymbol{x}_v^{(l-1)}$在确定关系$r$下关于邻居$\boldsymbol{x}_u^{(l-1)}$的重要程度分数；$d_k$为键向量的维度，用于缩放查询与键进行点积运算的结果。对所有邻居节点注意力分数进行softmax归一化得到注意力权重，即：

$$\boldsymbol{\alpha}_{vu}^{(l),r,i} = \mathrm{softmax}\left(\boldsymbol{e}_{vu}^{(l),r,i}\right) = \frac{\exp\left(\boldsymbol{e}_{vu}^{(l),r,i}\right)}{\sum\limits_{k\in N_r(v)}\exp\left(\boldsymbol{e}_{vk}^{(l),r,i}\right)} \tag{8}$$

式中，$\boldsymbol{\alpha}_{vu}^{(l),r,i}$为中心节点$\boldsymbol{x}_v^{(l-1)}$在确定关系$r$下关于邻居$\boldsymbol{x}_u^{(l-1)}$的注意力权重。中心节点在给定关系下对邻居节点的注意力权重反映了其与邻居节点基于社交关系分类的重要性分布，同时考虑了局部关系的异构性。随后，使用归一化的注意力权重对邻居节点用户的值进行加权求和，得到中心节点$\boldsymbol{x}_v^{(l-1)}$第i个注意力头的输出，即：

$$\boldsymbol{z}_v^{(l),i} = \sum_{r\in R}\sum_{u\in \mathrm{N_r(v)}}\boldsymbol{\alpha}_{vu}^{(l),r,i}\boldsymbol{V}_u^{(l),r,i} \tag{9}$$

式中，$\boldsymbol{z}_v^{(l),i}$为 GNN 中第l层中心用户节点$\boldsymbol{x}_v^{(l-1)}$的第i个注意力头输出。随后，将每个注意力头的输出进行拼接，并通过线性变换映射回到原始维度，即：

$$\boldsymbol{H}_v^{(l)} = \boldsymbol{W}_o^{(l)}\left[\boldsymbol{z}_v^{(l),1}\|\boldsymbol{z}_v^{(l),2}\|\cdots\|\boldsymbol{z}_v^{(l),h}\right] \tag{10}$$

式中，$\boldsymbol{H}_v^{(l)}$为中心用户节点$\boldsymbol{x}_v^{(l-1)}$的最终表示；$\boldsymbol{W}_o^{(l)}$为第l层可学习的线性变换矩阵；$\|$为列拼接操作。

3）节点残差。在用户节点的跨层表示中，将图神经网络中的残差连接引入用户节点的隐藏表

示，以确保平滑的表示学习，即：

$$\boldsymbol{x}_v^{(l)} = \beta \boldsymbol{H}_v^{(l)} + (1-\beta)\boldsymbol{x}_v^{(l-1)} \tag{11}$$

式中，β为可学习参数，用于控制经过多头注意力变换后特征$\boldsymbol{H}_v^{(l)}$与原始特征$\boldsymbol{x}_v^{(l-1)}$加权比例。经过聚合更新后提取出最终的用户表示$\boldsymbol{x}_v^G$。

1.3 多模态信息交互

本研究采用多通道表示器将元数据、文本和图结构模态的特征向量分别投影至不变子空间和特定子空间进行表示增强。不变子空间通过最小化模态差异来提取跨模态共性特征，将多模态数据映射到统一的共享子空间，使不同模态的数据分布趋于一致，从而捕捉用户行为的共性模式，这对识别用户真实意图至关重要。特定子空间则专注于挖掘各模态的独有特征，如语言风格、兴趣偏好等个性化信息，这些特征虽看似细微，却能显著提升检测精度。通过融合共享子空间的通用特征和特定子空间的个性化特征，可构建更全面的用户画像，为机器人检测提供更精准的判别依据。

为了学习这 2 个子空间，本文将分布相似损失应用于不变子空间，将正交损失应用于特定子空间，并通过重建损失确保隐藏表示能够捕获各自模态的细节。随后，使用融合增强后的用户表示进行社交机器人检测。

1.3.1 多通道表示器

为了充分利用社交平台用户多模态信息，同时受不变子空间与特定子空间架构在图形分类[22]和情感分类[24]方面应用的成功启发。本文利用多通道表示器将每个账号在元数据、文本和图形模态编码器得到的向量表示分别投射到 2 个空间中。首先，将每种模态特征向量投影到不变子空间，使用以下编码函数，即：

$$\boldsymbol{h}_i^G = E_i\left(\boldsymbol{x}_v^G;\boldsymbol{\theta}_i\right), \boldsymbol{h}_i^T = E_i\left(\boldsymbol{x}_v^T;\boldsymbol{\theta}_i\right), \boldsymbol{h}_i^M = E_i\left(\boldsymbol{x}_v^M;\boldsymbol{\theta}_i\right) \tag{12}$$

式 (12) 中的约束有助于消除不同模态数据间的结构性差异，使多模态信息能够更好地在共同的特征空间中融合与协同工作，从而在不同模态之间建立一致的特征表示。随后，将每种模态特征向量投影到特定子空间，用于捕获各模态下的独特特征，使用以下编码函数，即:

$$\boldsymbol{h}_s^G = E_s\left(\boldsymbol{x}_v^G;\boldsymbol{\theta}_s^G\right), \boldsymbol{h}_s^T = E_s\left(\boldsymbol{x}_v^T;\boldsymbol{\theta}_s^T\right), \boldsymbol{h}_s^M = E_s\left(\boldsymbol{x}_v^M;\boldsymbol{\theta}_s^M\right) \tag{13}$$

式 (13) 中的方法可以有效地分离每种模态中蕴含的特有信息，使得每种模态的特征能够在其独特的特征空间中得到充分表达。上述公式中E表示前馈神经层，E_i中$\boldsymbol{\theta}_i$为 3 个模态间共享参数，E_s中的$\boldsymbol{\theta}_s^G$、$\boldsymbol{\theta}_s^T$和$\boldsymbol{\theta}_s^M$是为每个模态分配的单独参数。通过多通道表示器特征增强，提供了融合所需的整体视图。

1.3.2 多模态特征融合

获取每个模态投影到特定子空间与不变子空间的特征表示后，应用多头自注意力机制[24]捕捉跨模态表征间的潜在关联特性。最终通过拼接 6 个经空间映射与关系建模的转换向量，构建下游社交机器人检测任务的联合特征表示。

首先，将串联投影到不变子空间与特定子空间的进行增强后的特征向量表示，形成矩阵$\boldsymbol{M} = [\boldsymbol{h}_i^G, \boldsymbol{h}_i^T, \boldsymbol{h}_i^M, \boldsymbol{h}_s^G, \boldsymbol{h}_s^T, \boldsymbol{h}_s^M]$。随后，对该矩阵应用多头自注意力机制。具体而言，通过线性变化得到用于Attention计算的查询、键和值，即：

$$\boldsymbol{Q}_i = \boldsymbol{W}_i^Q \boldsymbol{M} + \boldsymbol{b}_i^Q,\ \boldsymbol{K}_i = \boldsymbol{W}_i^K \boldsymbol{M} + \boldsymbol{b}_i^K,\ \boldsymbol{V}_i = \boldsymbol{W}_i^V \boldsymbol{M} + \boldsymbol{b}_i^V \tag{14}$$

式中，$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$和$\boldsymbol{W}_i^V$以及$\boldsymbol{b}_i^Q$、$\boldsymbol{b}_i^K$和$\boldsymbol{b}_i^V$为可学习的权重和偏置项，将得到的查询、键和值代入到缩放点积函数的注意力模块，得到第i个注意力头的输出：

$$\mathbf{head}_i = \mathrm{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}_i \tag{15}$$

随后，将多个注意力头的输出进行串联，并通过线性变换层进行融合，最终得到多头自注意力的输出表示：

$$\overline{\boldsymbol{M}} = \mathrm{MultiHead}(\boldsymbol{M}) = (\mathbf{head_1} \oplus \cdots \oplus \mathbf{head}_n)\boldsymbol{W}_o \tag{16}$$

式中，$\boldsymbol{W}_o$为可学习参数，$\oplus$代表串联。最后堆叠矩阵$\overline{\boldsymbol{M}}$中的列向量构造一个联合向量记为$\boldsymbol{h}_{\mathrm{out}}$。使用任务检测函数得到最终的标签，即：

$$\hat{\boldsymbol{y}} = D(\boldsymbol{h}_{\mathrm{out}};\boldsymbol{\theta}_{\mathrm{out}}) \tag{17}$$

1.4 损失函数

对模型整体预测结构的优化通过最小化以下损失函数进行，即：

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \alpha\mathcal{L}_{\mathrm{sim}} + \beta\mathcal{L}_{\mathrm{diff}} + \gamma\mathcal{L}_{\mathrm{recon}} \tag{18}$$

式中，α、β和γ为超参数，用于控制每部分损失对于总损失的贡献权重。这些分量损失负责实现所需的子空间性质。

1.4.1 相似性损失$\mathcal{L}_{\mathrm{sim}}$

相似性损失用于表征模态的不变表示，最小化相似性损失可以将每个模态的共享表示差距缩小。在不变子空间中使用相似度损失，有助于将跨模态特征对齐。本文使用中心力矩差异[28] CMD度量来

实现这一目的。CMD是一种先进的距离度量，它通过计算两个表示的顺序矩差来度量它们分布之间的差异，CMD距离会随着两个分布变得更加相似而减小。设$\boldsymbol{X}$和$\boldsymbol{Y}$是在$[a,b]$依概率p和q分布的有界随机样本。中心矩差异正则化器CMD_K被定义为CMD度量的经验估计，即：

$$\mathrm{CMD}_K(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{|b-a|}\|E(\boldsymbol{X})-E(\boldsymbol{Y})\|_2+\sum_{k=2}^{K}\frac{1}{|b-a|^k}\|C_k(\boldsymbol{X})-C_k(\boldsymbol{Y})\|_2 \tag{19}$$

式中，$E(\boldsymbol{X})=\frac{1}{|\boldsymbol{X}|}\sum_{x\in\boldsymbol{X}}x$为样本向量$\boldsymbol{X}$的期望；$C_k(\boldsymbol{X})=E\left((x-E(\boldsymbol{X}))^k\right)$为样本向量$\boldsymbol{X}$的第$k$阶中心矩。对于社交机器人检测任务来说，本文计算两种模态间不变表示之间的CMD损失为：

$$\mathcal{L}_{\mathrm{sim}}=\frac{1}{3}\sum_{(m_1,m_2)\in\{(G,T),(G,M),(T,M)\}}\mathrm{CMD}_K(\boldsymbol{h}_{m_1}^c,\boldsymbol{h}_{m_2}^c) \tag{20}$$

1.4.2　差异损失$\mathcal{L}_{\mathrm{diff}}$

差异损失用于表征模态间特定表示，最小化差异损失是为了确保模态不变和特定表示捕获输入的不同方面，且可以有效分离每种模态中蕴含的特有信息。不变子空间与特定子空间特征表示的差异性，通过在每种模态表示的 2 个子空间施加软正交性约束来实现。在同一批次训练的社交平台用户中，先将矩阵$\boldsymbol{H}_i^G$和矩阵$\boldsymbol{H}_s^G$转换零均值的单位 L2 范数，其行表示在不变子空间与特定子空间中图形模态特征表示的隐藏向量$\boldsymbol{h}_i^G$和$\boldsymbol{h}_s^G$。文本和元数据模态向量做相同变换，图形模态向量对的正交性约束计算为：

$$\left\|\boldsymbol{H}_i^{G\mathrm{T}}\boldsymbol{H}_s^G\right\|_{\mathrm{F}}^2 \tag{21}$$

式中，$\|\cdot\|_{\mathrm{F}}^2$为 Frobenius 范数的平方。除了不变子空间中的向量和特定子空间中的向量之间的约束外，添加模态特定子空间向量之间的正交约束表示为：

$$\mathcal{L}_{\mathrm{diff}}=\sum_{m\in\{G,T,M\}}\left\|\boldsymbol{H}_i^{m\mathrm{T}}\boldsymbol{H}_s^m\right\|_{\mathrm{F}}^2+\sum_{(m_1,m_2)\in\{(G,T),(G,M),(T,M)\}}\left\|\boldsymbol{H}_i^{m_1\mathrm{T}}\boldsymbol{H}_s^{m_2}\right\|_{\mathrm{F}}^2 \tag{22}$$

1.4.3　重建损失$\mathcal{L}_{\mathrm{recon}}$

强制实施差异损失可能导致模态特定编码器学习到无意义表示，本文引入重建损失以增强隐藏表示对模态细节的捕获能力。首先使用解码器函数$\hat{\boldsymbol{x}}_m=D(\boldsymbol{h}_m^i+\boldsymbol{h}_m^s;\boldsymbol{\theta}^d)$，重建向量$\boldsymbol{x}_m$得到向量$\hat{\boldsymbol{x}}_m$，随后，计算向量$\boldsymbol{x}_m$与$\hat{\boldsymbol{x}}_m$均方误差损失。$\|\cdot\|_2^2$为欧几里得范数的平方，重建损失表示为：

$$\mathcal{L}_{\mathrm{recon}}=\frac{1}{3}\left(\sum_{m\in\{G,T,M\}}\frac{\|\boldsymbol{x}_m-\hat{\boldsymbol{x}}_m\|_2^2}{d_h}\right) \tag{23}$$

1.4.4　分类任务损失$\mathcal{L}_{\mathrm{task}}$

为了训练优化 BotSAI，本文采用交叉熵损失来对抗过拟合。任务损失函数可表示为：

$$\mathcal{L}_{\mathrm{task}}=-\sum_{i\in U}[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)]+\lambda\sum_{\omega\in\theta}\omega^2 \tag{24}$$

式中，U表示训练集中所有的用户；λ为超参数；θ表示模型中所有训练参数；y_i表示用户i的真实标签；$\hat{y}_i$表示 BotSAI 框架对于用户i的预测标签。

2　实验

2.1　实验设置

2.1.1　数据集

本文提出的 BotSAI 检测模型采用社交关系图神经网络与异构信息感知的双重架构，其技术实现依托图结构化数据范式。为探究多维社交关系组合策略对分类性能的影响机制，研究设计跨平台对比实验框架，选用国际公认的 TwiBot-20[21] 和 MGTAB[16] 基准检测数据集。

TwiBot-20 数据集包含来自体育、经济、娱乐和政治领域的 229 573 名社交平台用户，其中 11 826 个账户被标记为真实账户或机器人账户，还包括 33 488 192 条推文，数据集支持用户间的关注和被关注社交关系信息。

MGTAB 数据集包含超过 155 万社交平台用户和 1.3 亿条推文，其中构建了 7 种社交关系的用户网络。本文选用了 MGTAB 数据集 7 种用户社交关系中的 5 种真实关系，包含关注、被关注、提及、回复和引用关系，并将 10 199 个账户标记为真实账户或机器人账户。

本文对所有数据集进行 7∶2∶1 的随机分割作为训练集、测试集和验证集，以便与以前的工作进行比较。神经网络的训练在一定程度上是随机的，因此即使在固定的超参数和数据分割的情况下，每次迭代后学习到的模型参数和误差也会有轻微变化。为了避免训练过程中的随机性，对模型进行 5 次迭代的训练和测试，但使用相同的分区数据，进行重复实验和平滑随机波动，提供了更稳定的模型有效性评估。本文选取 Accuracy 和 F1-Score 作为

评价指标，在两个数据集上进行实验。

2.1.2 基线模型

本文对 BotSAI 进行了全面的比较研究，考虑了以下各种基线。

Botometer[29] 从用户元数据、发布的内容和用户交互中，抽取了超过 1 000 个特征进行检测，使用现有数据集进行训练。

RoBERTa[25] 通过使用预训练模型 RoBERTa 对用户的文本信息进行编码，将编码后的用户特征输入 MLP 中用于机器人检测。

SGBot[7] 通过随机森林分类器使用从用户元数据中提取特征，用于可扩展、可推广的机器人检测任务。

HGT[30] 通过节点−边类型依赖的注意力机制建模异构关系，结合类型特定的线性投影计算注意力权重与消息传递，并设计 HGSampling 算法均衡采样大规模图的节点类型。

S-HGN[31] 基于 GAT，通过引入可学习类型嵌入利用类型信息、残差连接增强模型能力及对输出嵌入进行 L2 归一化来改进 HGNN。

RGT[14] 在由社交平台用户构成的关系异构网络中，使用关系图转换器及语义关注网络以增强社交机器人检测。

BotRGCN[13] 通过从跟随关系中构建异构的推特用户社交网络，利用关系图卷积网络进行用户表示学习及社交机器人检测。

SATAR[32] 使用社交平台用户的语义、属性和邻域信息，通过自监督方法进行用户表示学习，并在机器人检测数据集上进一步微调。

BIC[20] 通过文本−图形交互模块进行跨模态的信息交换，并使用基于注意权重的推文语义一致性进行建模的框架，用于社交机器人检测任务。

BotMoE[21] 联合利用多种社交平台用户信息，包括用户元数据、文本内容以及网络结构，同时结合了一个社区感知的专家混合层 MoE，能够出色地检测社交机器人并可以泛化到不同的推特社区。

2.1.3 实验具体实现

使用 Pytorch[33]、Pytorch Geometric[34]、Scikit-learn[35] 和 Transformer[27] 来实现 BotSAI，表 1 给出了检测框架的超参数设置，以方便复现实验结果。本文在一个拥有 4060GPU、8 GB 显存、16 个 CPU 核心和 32 GB 内存的服务器上进行所有实验。在 TwiBot-20 和 MGTAB 上训练 BotSAI 大约需要 4 h 和 6 h。

表 1 BotSAI 超参数设置

超参数	TwiBot-20	MGTAB
maximum epochs	400	400
hidden size	256	256
optimizer	Adam	Adam
learning rate	10^{-4}	10^{-4}
batch size	128	128
dropout	0.4	0.4
Oversampling scale(ω)	0.25	1.50
layer count(L)	2	2
L2 regularization(λ)	5×10^{-6}	5×10^{-6}
relational edge set(E)	{follower, following}	{follower, following, reply}
α, β, γ	{0.7, 0.3, 1.0}	{0.7, 1.0, 1.0}

2.2 实验结果与分析

本研究首先系统评估了各基线方法对文本数据、图结构信息以及多模态特征交互的利用程度，随后在 TwiBot-20 和 MGTAB 这 2 个基准数据集上进行了实验。实验结果表明，BotSAI 模型显著优于 10 种基于特征工程、文本分析、图神经网络和多模态融合的基线方法。如表 2 所示，所有方法均经过 5 次独立实验，性能指标以“均值（标准差）”形式呈现，其中粗体标注最优结果，下划线标注次优结果，“−”表示该方法无法适用于 MGTAB 数据集。

实验分析表明。1）在 2 个数据集上，BotSAI 始终显著优于所有基线方法。具体来说，与以往先进的 BotMoE 相比，BotSAI 在 TwiBot-20 上的准确率提高了 0.81%，F1-score 提高了 1.21%；在 MGTAB 数据集上准确率提高了 2.21%，F1-score 提高了 0.05%。在社交机器人检测任务中，本文所提出的 BotSAI 检测框架实现了显著的性能提升。2）实验结果显示，BotSAI 在 MGTAB 数据集上的检测准确率提升幅度超过了在 TwiBot-20 数据集上的提升幅度。这表明，在社交网络中对多种社交关系组合的合理利用（MGTAB 数据集中应用关注、被关注和回复社交关系的组合，TwiBot-20 数据集中应用关注和被关注社交关系的组合）能够显著提升社交机器人检测的效果。3）实验表明，融合多模态信息的检测模型，如 BotSAI、BIC 和 BotMoE 在各种测试环境和数据集上均表现出卓越性能，显著优于仅依赖单一模态的传统方法。特别是在面对复杂和多变的社交媒体环境时，整合了多模态信息融合的模型展示出更强的适应能力和检测精度。此外，实验再次证实了基于图形的方法相较基于文本

的方法具有更好的检测效果。4）本文提出了将多模态信息映射到不变与特定子空间进行特征增强的社交机器人检测框架，该框架在 2 个数据集上的综合基准测试中获得了最佳性能。这些结果证明了利用社交平台用户多模态信息进行检测的必要性以及本文方法的有效性。

表 2　BotSAI 在两个数据集上的准确率和 F1 分数

分类模型	模态			TwiBot-20		MGTAB	
	文本	图形	多模态交互	Accuracy/%	F1-score/%	Accuracy/%	F1-score/%
Botometer				53.09	53.09	—	—
RoBERTa	√			75.50(±0.13)	73.09(±0.60)	79.03(±0.13)	75.80(±0.30)
SGBot	√			81.60(±0.50)	84.90(±0.40)	84.70(±0.40)	86.10(±0.48)
HGT		√		85.90(±0.60)	85.60(±0.60)	89.48(±0.30)	87.50(±0.40)
S-HGN		√		85.40(±0.30)	85.30(±0.20)	89.37(±1.50)	88.70(±0.60)
RGT		√		86.57(±0.11)	88.01(±0.40)	89.30(±0.40)	90.20(±0.50)
BotRGCN		√		86.30(±0.50)	86.60(±0.40)	89.26(±0.50)	87.20(±0.70)
SATAR	√	√		84.00(±0.80)	86.07(±0.70)	88.12(±0.10)	87.50(±0.10)
BIC	√	√	√	87.61(±0.21)	89.13(±0.15)	89.34(±0.50)	89.77(±0.30)
BotMoE	√	√	√	87.76(±0.20)	89.22(±0.30)	89.50(±1.30)	90.30(±0.30)
BotSAI	√	√	√	**88.57(±0.15)**	**90.43 (±0.30)**	**91.71(±0.20)**	**90.35(±0.40)**

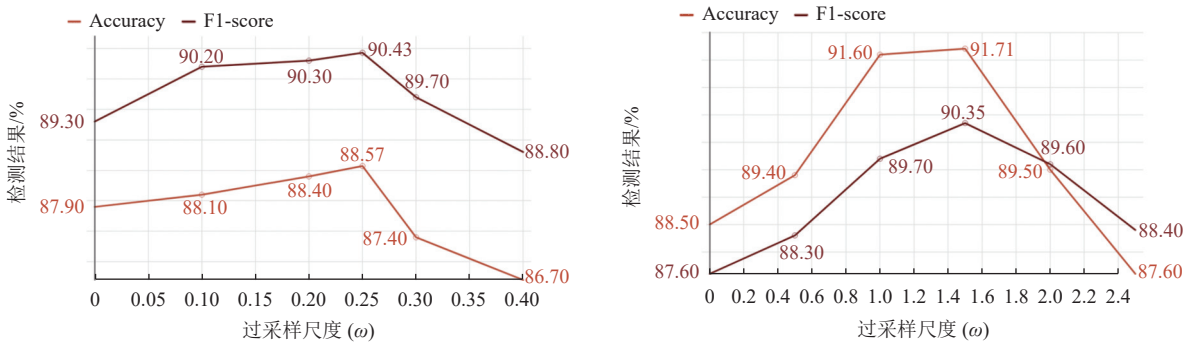
3　BotSAI 模型组件有效性研究

从以下几个方面继续展开研究。1）探讨图形数据增强对检测性能的影响；2）验证所提出的局部关系转换器的有效性；3）分析不同社交关系组合在社交网络异构图中的应用及其对检测效果的贡献；4）探究不变子空间与特定子空间在机器人检测方法中的作用与意义。

3.1　图形数据增强有效性研究

在 TwiBot-20 和 MGTAB 数据集中，用户被分为机器人和人类两类，然而这 2 个数据集均存在显著的类别不平衡问题。具体而言，TwiBot-20 数据集中包含 5 237 个真实用户样本，而机器人样本数量达到 6 589 个，机器人与人类的比例约为 1.26∶1。相比之下，MGTAB 数据集的不平衡现象更为突出，其真实用户样本量为 7 554 个，而机器人样本仅有 2 830 个，机器人与人类的比例约为 0.37∶1。这种显著的类别分布不均衡现象可能会对模型的训练和评估产生重要影响。

当训练数据不平衡时，模型可能难以从少数类中学习到足够的特征，导致倾向于将大多数样本分类为多数类。对于这 2 个数据集，尽管以往方法的总体准确率很高，但对多数类的预测存在明显偏差，少数类的分类准确率要低得多[36]。在 BotSAI 框架汇聚异构图中有有效信息时，本文引入过采样技术的数据增强策略。这种策略能够根据异构图中不同子图的节点特征分布差异，自适应调整采样尺度ω，以平衡少数类样本在不同子图空间的表征密度。依赖本文提出的 BotSAI 模型，仅改变采样尺度ω，保持其他参数不变进行实验。实验结果（图 2）表明，检测性能随过采样系数ω的增大呈先升后降趋势。在 TwiBot-20（ω=0.25）和 MGTAB（ω=1.5）数据集中分别达到最优准确率和 F1 分数，过采样增加了少数类样本的多样性，使模型能够学习到更全面的特征表示；当ω超过最优阈值，过度的过采样导致多数类样本的特征分布被稀释，同时模型逐渐过拟合合成样本的噪声特征，致使性能显著下降。



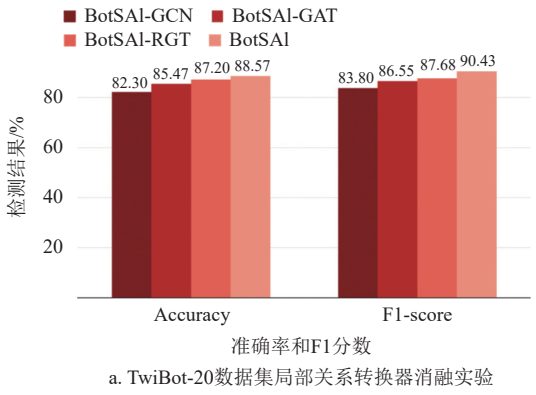
a. TwiBot-20数据集过采样尺度调优实验　　b. MGTAB数据集过采样尺度调优实验

图 2　过采样尺度ω对检测结果影响的对比分析图

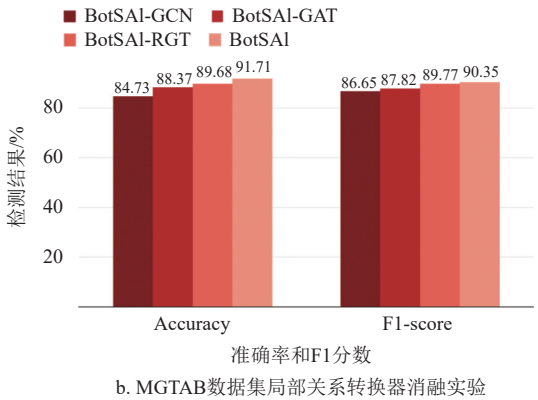
3.2 局部关系转换器有效性研究

互联网社交中，机器人账号表现出显著的目的导向性：高级机器人通常基于预设身份（如科技博主）策略性地关注和点赞特定领域用户以提升影响力。相比之下，真实用户行为更具随机性，其关注和点赞行为不仅限于特定对象，还会受到系统推荐的影响而呈现多样化特征。

本文提出的局部关系转换器相较于当前先进的 RGT 图神经网络分类模型中的语义注意网络，通过引入更多权重参数，强化了用户在同质社交关系中对不同邻居节点的差异化建模能力，从而实现了更精细的用户行为特征提取。为了深入评估本文提出的局部关系转换器对 2 个数据集上模型性能的影响，本文将局部关系转换器分别替换为 GCN、GAT 和 RGT，同时确保其余模型框架与实验参数保持不变，以此来进行消融实验，结果如图 3 所示。在 TwiBot-20 数据集中相比 RGT 检测准确率提升了 1.37%，F1 分数提升 2.75%；在 MGTAB 数据集中相比 RGT 检测准确率提升 2.03%，F1 分数提升 0.58%。实验结果表明本文提出的局部关系转换器能够进行更为细致的信息聚合，从而显著提高检测性能。



a. TwiBot-20数据集局部关系转换器消融实验



b. MGTAB数据集局部关系转换器消融实验

图 3 不同组件替换对局部关系转换器性能影响的对比实验结果

3.3 多种社交关系组合有效性研究

网络中存在着多样化的社交方式，同时先进的机器人则通过模仿与真实用户的多种社交互动来掩饰其非人类身份。目前，多数检测方法仅依赖 Follower 和 Following 社交关系。然而，操纵这两种社交关系的机器人账号成本低廉，通过大规模的简单操作就能轻易地与真实用户混淆。因此仅使这两种社交关系聚合邻居信息，会导致机器人特征的过度平滑，从而严重影响检测效果。

为探究不同社交关系对检测性能的影响，本研究基于 MGTAB 数据集构建了多种社交关系网络异构图进行对比实验。实验在 BotSAI 框架下进行，仅调整社交网络异构图结构，其他参数保持不变。每个基线模型采用不同随机种子运行 5 次，实验结果以均值和标准差形式呈现，如表 3 所示。最优结果以粗体标注，主被动关系对比结果以下划线斜体标识。

实验结果总结如下。1）对于社交机器人检测任务，通过综合运用多种社交关系组合构建异构网络图，能够更有效捕捉机器人与真实用户之间的差异。2）在单一社交关系的消融实验中，Follower 组相比 Following 组检验准确率高出 3.12%，在使用复合关系对比实验中 B+A 组对比 B+C 组检验准确率提高 1.46%。结果证明被动社交关系 Follower 相比主动社交关系 Following、Mention 蕴含更多机器人与真实用户的差异信息。3）本文中，最优社交关系组合 A+B+D 相较于采用全部社交关系的组合 A+B+C+D+E，检测准确率提升了 2.16%。值得关注的是，提升检测效果的关键并非在于盲目增加社交关系的种类，而在于选取关键社交关系以构建图网络。通过这种方式，模型能够聚焦于最相关的信息，同时有效避免因无关关系引入的噪声干扰，从而显著提升模型的整体性能与预测准确性。

实证研究表明，社交机器人检测效能的核心在于对被动关联特征的精准识别与分析。这种检测机制的有效性源于被动关系形成的天然壁垒：社交机器人操纵者难以通过程序化手段模拟真实用户自然形成的社交网络特性，尤其体现在高可信度用户往往自发形成具有声望同质性的关注网络，其追随者群体通常具有相近的社会影响力层级。为模仿此类真实账号，社交机器人需要实现与高影响力节点的双向互动（如获得高声望用户的主动关注或内容回

应）。这种基于主被动关系构建机制相较于单向关注行为，在技术实现层面存在更高的复杂性和操作门槛，这也正是被动关联特征能够成为有效检测依据的关键所在。

表 3　多种社交关系的应用对检测效果的影响

类型	关系	关系描述	Accuracy/%	F1-score/%
单一社关系	Follower: A	User A is followed by user B	87.45(±0.40)	85.48(±0.40)
	Following: B	User A follows user B	84.33(±0.30)	80.77(±0.45)
	Mention: C	User A mentions user B in tweets	86.73(±0.20)	82.57(±0.30)
	Reply: D	User A replies to tweet of user B	86.53(±0.25)	84.49(±0.15)
	Quoted: E	User A quotes tweet of user B	85.93(±0.15)	83.46(±0.10)
社交关系组合	B+A	User A combines active and passive behaviors	87.89(±0.10)	87.60(±0.15)
	B+C	User A passive behavior combination	86.43(±0.25)	84.68(±0.30)
	B+C+D+E	User A combines all passive relationships	88.73(±0.10)	86.45(±0.20)
	A+B+C+D+E	User A full relationship set	89.55(±0.30)	86.57(±0.35)
	BotSAI: A+B+D	Optimal relationship combination	**91.71(±0.20)**	**90.35(±0.40)**

3.4　模态不变与特定子空间研究

3.4.1　子空间特性研究

为了探索子空间对社交机器人检测结果的影响，研究 BotSAI 模型的 3 个变体。1）BotSAI-BASE 是 1 个基线版本，通过为每个模态使用 3 个独立的编码器，不学习任何子空间，直接对编码后的特征进行融合，使用融合后的向量进行检测。2）BotSAI-SF 变体在表示学习阶段与 BotSAI 相同，只使用特定于模态的特征 $(\boldsymbol{h}_s^G \boldsymbol{h}_s^T \boldsymbol{h}_s^M)$ 进行融合和检测。3）BotSAI-IF 变体在表示学习阶段与 BotSAI 相同，只使用模态不变的特征 $(\boldsymbol{h}_i^G \boldsymbol{h}_i^T \boldsymbol{h}_i^M)$ 进行融合和检测。

本研究通过对比实验揭示了多模态表征学习策略对检测性能的影响。如图 4 所示，在 TwiBot-20 数据集上，BotSAI-SF 相较于 BotSAI-BASE 在检测精度（−0.69%）和 F1 值（−0.78%）上均呈现性能下降；而在 MGTAB 数据集上，其检测精度降幅达 1.13%，F1 值降低 0.64%。这种性能衰减可归因于 BotSAI-SF 变体过度强调跨模态差异性，导致模态间噪声干扰增大。相比之下，BotSAI-IF 通过有效捕捉多模态共性特征，在 2 个基准数据集上分别取得 0.92% 和 1.35% 的 F1 值提升，验证了共性特征学习的重要性。基于此发现，本文提出的 BotSAI 模型构建了模态不变空间（Invariant Space）和模态特定空间（Specific Space）的协同优化机制。实验结果表明，这种架构在 TwiBot-20 和 MGTAB 数据集上分别实现了 88.57% 和 91.71% 的检测准确率，较最优基线提升 0.81% 和 2.21%，证实了这种联合优化策略的有效性。

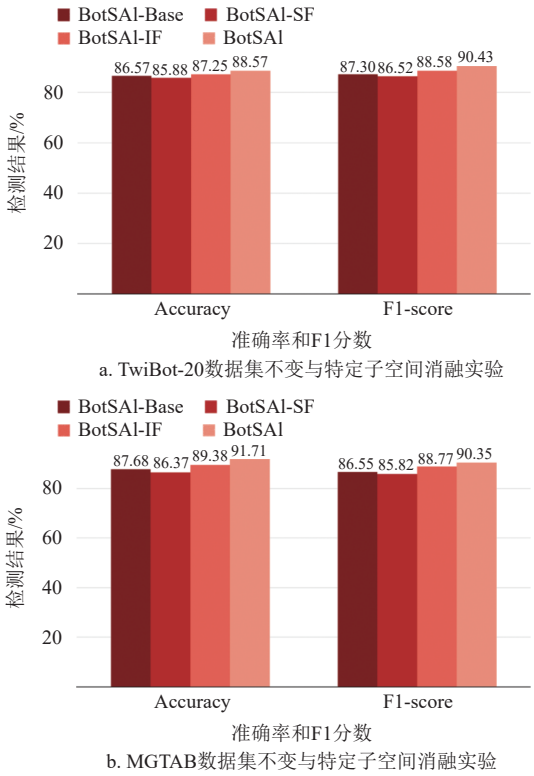
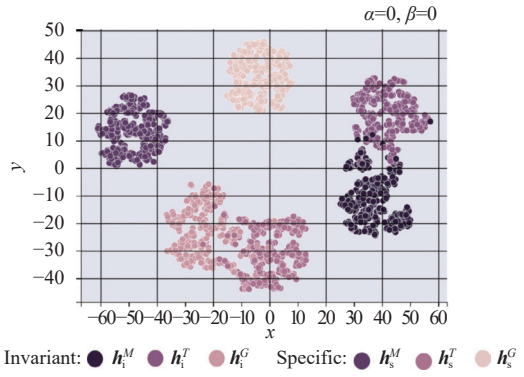


图 4　子空间消融实验

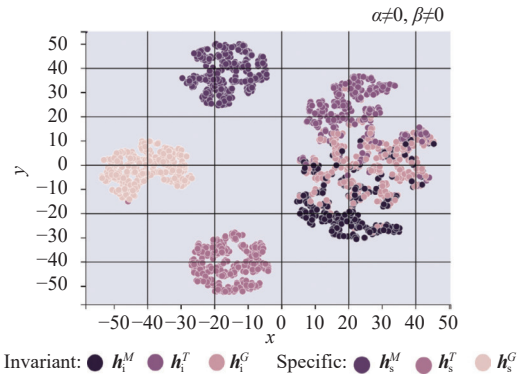
3.4.2　子空间可视化表示

为深入分析模态不变子空间与模态特定子空间对特征表示泛化能力的影响机制，本研究采用 t-SNE

降维技术对 TwiBot-20 和 MGTAB 测试集的用户表征进行可视化分析（如图 5 和图 6 所示）。可视化结果表示，当$\alpha=0$、$\beta=0$时模型的特征分布呈现明显的模态间可分性不足的现象，难以有效捕捉跨模态的共性与特定特征。然而，在相似性与差异性损失函数的驱动下，$\alpha\neq0$、$\beta\neq0$，模态不变子空间呈现跨模态表征的显著交叠，而特定子空间则实现跨模态的有效分离性。这表明 BotSAI 成功实现了预期的子空间学习目标，验证了方法的有效性。

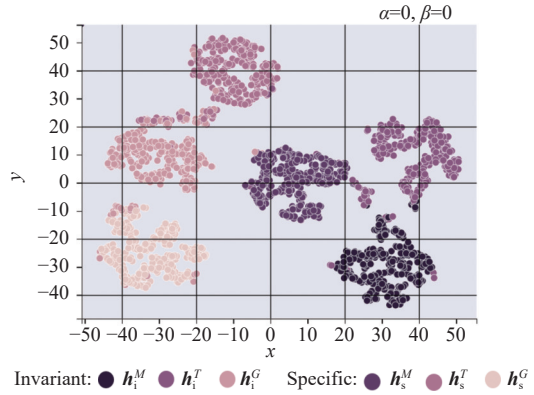


a. TwiBot-20数据集中关闭相似性和差异损失

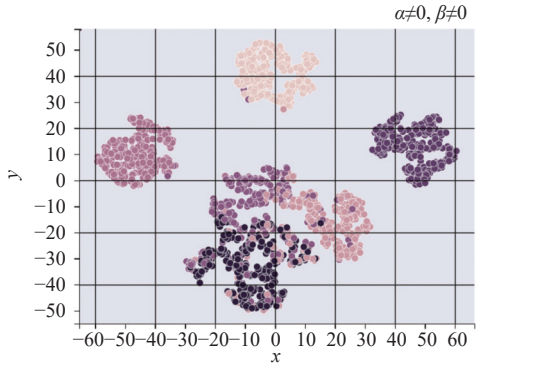


b. TwiBot-20数据集中启用相似性和差异损失

图 5 TwiBot-20 数据集不变与特定子空间中隐藏表示的可视化



a. MGTAB数据集中关闭相似性和差异损失



b. MGTAB数据集中启用相似性和差异损失

图 6 MGTAB 数据集不变与特定子空间中隐藏表示的可视化

4 结束语

社交机器人检测是一项重要且具有挑战性的任务。本文提出了一个新的检测框架 BotSAI，检测过程分为多模态信息表征学习与多模态信息交互 2 个阶段，利用多模态特征向量投影到不变与特定子空间的方法，增强模态间共性凸显特性，旨在解决高度复杂机器人伪装逃避检测的难题。大量实验表明，BotSAI 在 2 个社交机器人检测基准上始终优于现有方法。在未来，将侧重研究从被动社交关系中发掘真实用户与机器人的差异，并将构建具有多种可利用被动社交关系的数据集。

参 考 文 献

[1] CRESCI S, YANG K C, SPOGNARDI A, et al. Demystifying misconceptions in social bots research[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2303.17251.

[2] YANG K C, PIERRI F, HUI P M, et al. The covid-19 infodemic: Twitter versus facebook[J]. Big Data & Society, 2021, 8(1): 20539517211013861.

[3] FENG S, WAN H, WANG N, et al. What does the bot say? Opportunities and risks of large language models in social media bot detection[EB/OL]. [2024-07-04]. https://arxiv.org/pdf/2402.00371.

[4] PACHECO D. Bots, elections, and controversies: Twitter insights from Brazil′s polarised elections[C]//Proceedings of the ACM on Web Conference 2024. New York: ACM, 2024: 2651-2659.

[5] 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. Montreal: AAAI, 2017: 280-289.

[6] HE B, YANG Y, WU Q, et al. Dynamicity-aware social bot detection with dynamic graph transformers[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2404.15070.

[7] 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. New York: AAAI, 2020: 1096-1103.
[8] MAZZA M, CRESCI S, AVVENUTI M, et al. Rtbust: Exploiting temporal patterns for botnet detection on twitter[C]//Proceedings of the 10th ACM Conference on Web Science. Boston: ACM, 2019: 183-192.
[9] D'ANDREA E, DUCANGE P, LAZZERINI B, et al. Real-time detection of traffic from twitter stream analysis[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2269-2283.
[10] WEI F, NGUYEN U T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings[C]//2019 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). Los Angeles: IEEE, 2019: 101-109.
[11] DIAZ J, POBLETE B, BRAVO-MARQUEZ F. An integrated model for textual social media data with spatio-temporal dimensions[J]. Information Processing & Management, 2020, 57(5): 102219.
[12] DUKIC D, KECA D, STIPIC D. Are you human? Detecting bots on Twitter using BERT[C]//2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Sydney: IEEE, 2020: 631-636.
[13] FENG S, WAN H, WANG N, et al. BotRGCN: Twitter bot detection with relational graph convolutional networks[C]//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Netherlands: IEEE/ACM, 2021: 236-239.
[14] FENG S, TAN Z, LI R, et al. Heterogeneity-aware twitter bot detection with relational graph transformers[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2022: 3977-3985.
[15] FENG S, TAN Z, WAN H, et al. Twibot-22: Towards graph-based twitter bot detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 35254-35269.
[16] SHI S H, QIAO K, CHEN J, et al. MGTAB: A multi-relational graph-based twitter account detection benchmark[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2301.01123.
[17] 余尚戎, 肖景博, 殷琪林, 等. 关注社交异配性的社交机器人检测框架[J]. 信息网络安全, 2024, 24(2): 319-327.
YU S R, XIAO J B, YIN Q L, et al. A social heterophily focused framework for social bot detection[J]. Netinfo Security, 2024, 24(2): 319-327.
[18] FENG S, WAN H, WANG N, et al. Twibot-20: A comprehensive twitter bot detection benchmark[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2021: 4485-4494.
[19] 徐开元, 周安民, 陈艾琳, 等. 结合主动学习与关系图卷积神经网络的社交机器人检测[J]. 四川大学学报（自然科学版）, 2023, 60(5): 121-129.
XU K Y, ZHOU A M, CHEN A L, et al. Social bot detection based on active learning and relational graph convolutional neural networks[J]. Journal of Sichuan University (Natural Science Edition), 2023, 60(5): 121-129.
[20] LEI Z, WAN H, ZHANG W, et al. BIC: Twitter bot detection with text-graph interaction and semantic consistency[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2208.08320.
[21] 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. [S. l.]: ACM, 2023: 485-495.
[22] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: ACM, 2016: 343-351.
[23] LIU P F, QIU X P, HUANG X J. Adversarial multi-task learning for text classification[EB/OL]. [2024-04-19]. https://arxiv.org/pdf/1704.05742.
[24] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA: Modality invariant and specific representations for multimodal sentiment analysis[EB/OL]. [2024-05-08]. https://arxiv.org/pdf/2005.03545v2.
[25] LIU Y H, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[EB/OL]. [2024-07-26]. https://arxiv.org/pdf/1907.11692.
[26] SHI S, QIAO K, YANG J, et al. Over-sampling strategy in feature space for graphs based class-imbalanced bot detection[C]//Companion Proceedings of the ACM on Web Conference 2024. Singapore: ACM, 2024: 738-741.
[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS′17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: [s.n.], 2017: 6000-6010.
[28] ZELLINGER W, GRUBINGER T, LUGHOFER E, et al. Central moment discrepancy (cmd) for domain-invariant representation learning[EB/OL]. [2024-05-02]. https://arxiv.org/pdf/1702.08811.
[29] YANG K C, FERRARA E, MENCZER I, et al. Botometer 101: Social bot practicum for computational social scientists[J]. Journal of Computational Social Science, 2022, 5(2): 1511-1528.
[30] HU Z, DONG Y, WANG K, et al. Heterogeneous graph transformer[C]//Proceedings of The Web Conference 2020. [S. l.]: ACM, 2020: 2704-2710.
[31] LYU Q, DING M, LIU Q, et al. Are we really making much progress? Revisiting, benchmarking, and refining heterogeneous graph neural networks[EB/OL]. [2024-11-30]. https://arxiv.org/pdf/2112.14936.
[32] FENG S, WAN H, WANG 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. [S.l.]: ACM, 2021: 3808-3817.
[33] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in PyTorch [EB/OL]. [2024-11-30]. https://openreview.net/pdf?id=BJJsrmfCZ.
[34] FEY M, LENSSEN J E. Fast graph representation learning with PyTorch geometric[EB/OL]. [2024-03-25]. https://arxiv.org/pdf/1903.02428.
[35] SWAMI A, JAIN R. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2013, 12(10): 2825-2830.
[36] HU F, WANG L, WU S, et al. Graph classification by mixture of diverse experts[EB/OL]. [2024-04-29]. https://arxiv.org/pdf/2103.15622.

责任编辑　刘飞阳