2025 年 第 25 卷 第 10 期
2025, 25(10): 04183-09
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2403036
引用格式:王宇哲, 吴安昊, 闫钦与, 等. 基于变分推理与图神经网络的机器水军检测[J]. 科学技术与工程, 2025, 25(10): 4183-4191.
Wang Yuzhe, Wu Anhao, Yan Qinyu, et al. Bot detection by variational inference and graph neural network[J]. Science Technology and Engineering, 2025, 25(10): 4183-4191.

自动化技术、计算机技术

基于变分推理与图神经网络的机器水军检测

王宇哲, 吴安昊, 闫钦与, 颜靖华*
(中国人民公安大学信息网络安全学院, 北京 100038)

摘 要 随着互联网和社交平台的飞速发展,机器水军检测问题已成为构建和谐互联网环境的一大技术挑战。然而,从社交平台收集的用户数据存在信息缺失、数据噪声等问题。因此,针对图学习检测机器水军模型中,使用点估计作为权重的方法在数据单一或缺失数据的区域无法表达不确定性的问题。提出了一种融合变分推理的图神经网络机器水军检测模型VRGAT,它引入了权值的概率分布,导出了真实后验的变分近似,通过为均值和方差分别使用不同的卷积运算,更准确地捕捉数据的变异性。基于 Twibot-20 数据集开展了仿真验证,相较于已有的最佳机器水军检测基准(F_1 =88.12),VRGAT 模型实现了性能提升,达到 F_1 = 89.64。在鲁棒性实验中加入不同比例的随机噪声,VRGAT 模型的准确率下降相比其他基线模型明显减缓,表明其抗噪声能力优于已有基线方法。实验结果表明,引入变分推理能够提高机器水军检测效果及模型抗噪声能力。

关键词 机器水军检测; 变分推理; 图神经网络; 社交网络
中图法分类号 TP183; **文献标志码** A

Bot Detection by Variational Inference and Graph Neural Network

WANG Yu-zhe, WU An-hao, YAN Qin-yu, YAN Jing-hua*
(School of Information Network Security, People's Public Security University of China, Beijing 100038, China)

[**Abstract**] With the rapid development of the Internet and social platforms, the problem of spammer detection has become a major technical challenge in building a harmonious Internet environment. However, user data collected from social platforms are often subject to issues such as missing information and data noise. Therefore, in graph-based learning models for bot army detection, methods that use point estimation as weights fail to express uncertainty in regions with sparse or missing data. A graph neural network model for bot army detection, VRGAT, integrating variational inference, was proposed. It introduces a probability distribution for the weights and derives a variational approximation of the true posterior. By applying different convolution operations to the mean and variance, the model more accurately captures the variability in the data. Simulations based on the Twibot-20 dataset show that, compared to the best existing benchmark for bot army detection (F_1 = 88.12), VRGAT achieved an improved performance with an F_1 score of 89.64. In robustness experiments, when random noise was added at varying levels, the accuracy drop for VRGAT is significantly slower than for other baseline models, demonstrating its superior noise resistance. The experimental results demonstrate that the introduction of variational inference can enhance the effectiveness of spammer detection and improve the model's robustness against noise.
[**Keywords**] spammer detection; variational inference; graph neural network; social network

自动化程序创建的机器水军,发布虚假信息、扭曲事实,挑起敏感话题引导社会舆论或造成网络暴力,严重威胁了网络环境的和谐[1]。机器水军具有大规模协同操作和“自动养号”的特点,其行为模式模仿正常用户具有动态多变性,传统简单聚合用户特征的方法以无法满足具有高动态性的机器水军检测任务。例如, Kresovich 等[2]发现社交媒体平台充斥大量与电子烟话题相关的社交机器人,利用自动化机制广泛发布“电子烟可帮助戒烟”“电子烟比香烟安全”等言论,误导青少年群体的认知;Wang 等[3]发现机器人账号在“否定气候变化”话题中大规模扩散虚假信息,影响公众的环保态度与政策支持意愿;Zhang 等[4]以消费者舆论为背景,研究了机器人账号大量介入社交媒体商业领域的意见集群(opinion flocks),发现机器人会主动推送某些话题或情绪性信息从而造成普通用户误判舆论氛围,影响其购买行为与品牌信

收稿日期: 2024-04-24; **修订日期**: 2025-01-01
基金项目: 中国人民公安大学安全防范工程双一流专项(2023SYL08)
第一作者: 王宇哲(2000—),男,汉族,河北石家庄人,硕士研究生。研究方向:图神经网络。E-mail:infinite_zhe@sina.com。
* **通信作者**: 颜靖华(1980—),女,汉族,河北石家庄人,博士,副教授。研究方向:数据警务技术。E-mail:yanjinghua@ppsuc.edu.cn。

任度。因此,亟需开发高级机器水军检测技术,以保障信息传播的真实性和网络空间的安全。

为了设计适应高动态性的机器水军检测机制,主流研究大多基于机器学习技术。机器学习技术具有强大的数据处理能力和模式识别能力,能够从大量复杂的数据中学习到潜在的行为模式,因此,其能够很好地应对机器水军不断变化的策略和伪装技巧,中外相关学者也提出了众多机器学习框架。相关研究主要基于特征工程、自然语言处理(natural language processing, NLP)和图神经网络(graph neural network, GNN)3 种方法,这些方法分别通过分析社交网络中的用户特征、用户发文和网络结构特征来识别机器水军[5],而基于 GNN 的机器水军检测方法能够通过学习图结构中的节点和边的特征来深入分析图结构中的高维信息,能够高效地识别出机器水军的复杂行为模式。因此,在处理具有复杂数据结构的社交网络水军检测方面,基于 GNN 的水军检测方法相较于其他方法性能更优。

早期特征工程的方法是基于用户特征手工制作的,这些特征来自元数据[6-8]和用户文本信息[9],Yang 等[10]提出利用用户的社交圈、互动模式特征;Alceu 等[11]和 Hurtado 等[12]提出利用时间和网络信息特征,然后将这些特征与传统的分类算法[13-14]结合以识别机器水军。

基于 NLP 的检测方法通过分析用户发文的情感倾向性[15]、词汇丰富度和句法复杂性[16]来检测异常行为。杜锐等[17]对中文微博中主客观分类特征的选取进行了研究,通过词典与统计相结合的方法提取了基础情感词、语气词、程度词等 8 个候选特征。Feng 等[18]利用词向量和双向 LSTMs 来处理用户文本信息进行机器人检测。David 等[19]提出了一种基于 BERT 来分析用户推文的机器人检测模型。另外,EdaDyFe-FL 对网络暴力检测中的文本数据不平衡问题进行了有效处理[20]。然而 Cresci[1]发现机器水军可以通过创建具有操纵元数据和被盗推文的欺骗性帐户来逃避检测,这使得单纯基于特征工程和基于文本情感分析的方法已经不能满足机器水军检测的需要。

近年来 GNN 发展迅速,通过从网络结构和用户行为模式中自动学习特征,以区分机器人和人类用户的复杂特征。例如,lhosseini 等[21]提出了一种基于自编码器的无监督学习方法,以识别推特上的机器人账号。Eiman 等[22]提出了一种基于多视图注意力机制的框架,使用了标记和未标记的数据以预测社交机器人。然而,这种方法通过浅层的特征融合加图神经网络层难以精确地检测出高级机器人不同数据模态间的潜在不一致性。于是一种 Twitter 用户的自我监督表示学习框架(a self-supervised approach to twitter account representation learning and its application in bot detection, SATAR)[23]利用 Twittersphere 的图结构进行机器人检测。关系图变换神经网络(relational graph transformers, RGT)[24]关注于机器人在异构关系分布的异质性,通过 RGT 动态地结合和利用用户之间的各种关系和影响模式,捕捉用户间的复杂互动。由于社交网络用户交互爆发大量信息,仅依靠分析关系异质性并不能充分表征机器水军的复杂性,因此本文在考虑关系异质性的同时结合节点的热度融合语义的综合影响力。

现有的 GNN 水军检测技术主要依靠大量的标注数据来训练模型,深入挖掘社交网络中的交互模式和用户行为特征。然而,依然存在一些关键局限性。首先,以往 GNN 方法在数据处理时并没有充分考虑节点间相互作用产生的热度、语义和异构关系特征对网络的综合影响,单纯的特征分析不能充分表征机器水军节点。其次,机器水军检测实质上是二分类任务,而用于训练的社交网络数据集通常包含信息缺失和大量噪声,这些因素可能干扰图神经网络的学习过程,从而降低检测性能。此外,图神经网络在学习过程中的不透明性常使其被视为“黑箱”,导致模型决策过程难以理解,尤其是在需要质疑关键决策的场景中。

针对上述节点表征不全、信息缺失、数据噪声等问题,现设计一种融合变分推理的关系图神经网络模型,并对注意力网络进行扩展,即 VRGAT(variational relational graph attention neural networks)。变分推理的概率表示相比传统方法更能捕捉到数据的内在不确定性,通过构建机器水军网络的高维特征潜在空间 Z,重参数化后补全原本的图结构提升了原本数据集数据缺失的问题。通过残差连接和 L2 归一化,在传播前有效防止孤立节点的嵌入向量趋近于零,进一步提高模型对机器水军检测的准确率和抗噪声能力。

1 基于变分推理和图神经网络的机器水军检测模型

首先,本文模型学习每个节点所表示的语义、关系、热度的表示。接着,将语义、关系和热度信息融合的数据构建为表示异构的信息网络与多样化关系的多元异构网络 mHIN(multiple heterogeneous information network)。之后,采用全局视图的图形和动态聚合表示与语义、关系和热度的关系。最后,通过变分编码器捕捉节点潜在特征,将社交网络用户分类为机器水军或正常用户,并学习数据模式与其他参数。模型结构图如图 1 所示,算法流程如算法 1 所示。

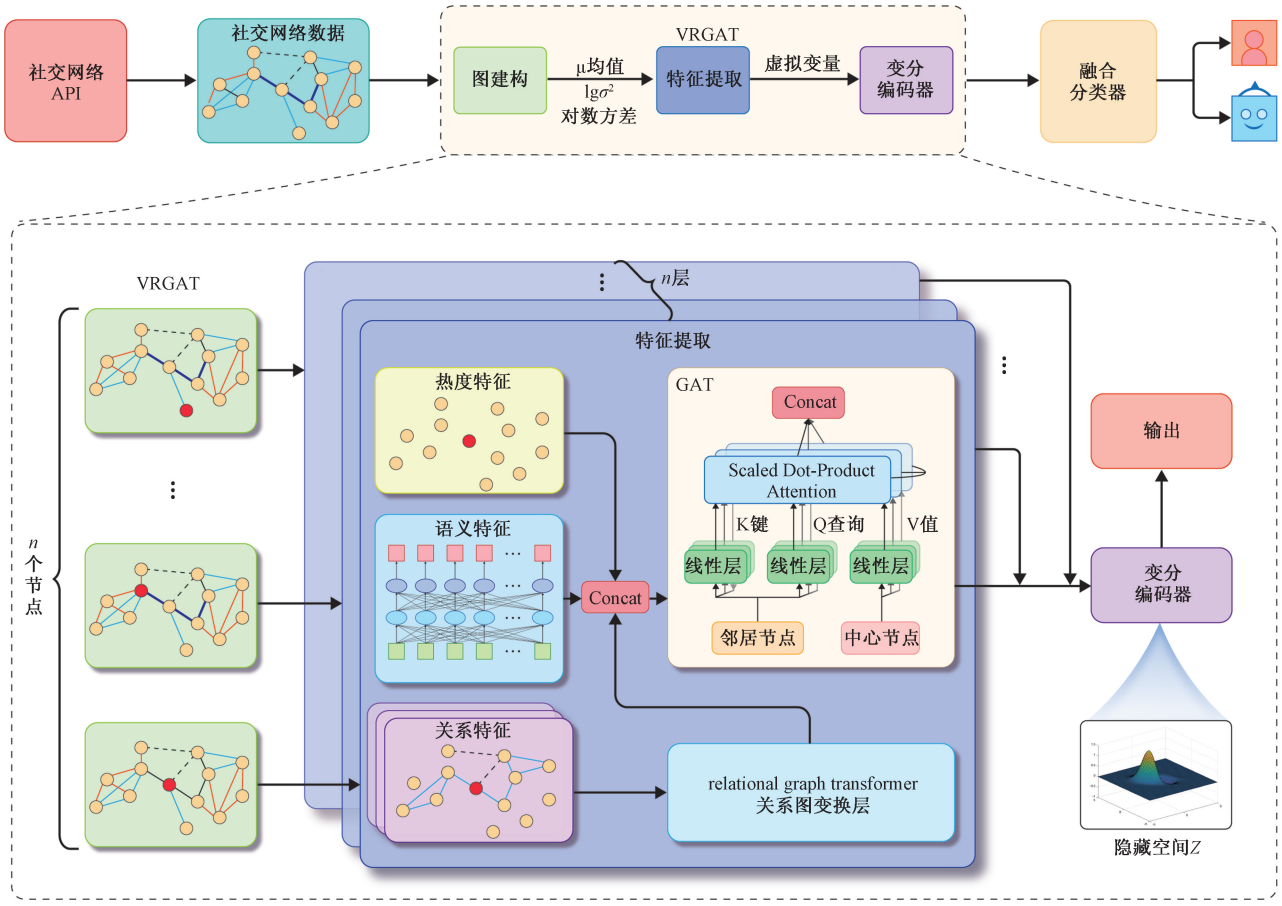

图 1　融合变分推理的关系图神经网络模型(VRGAT)结构示意图

Fig. 1　Structure of VRGAT

1.1　特征提取

受到 RGT[24] 的启发,本文研究构建了一个融合语义、关系和热度的多元异构的信息网络(mHIN),它同时考虑到语义、关系和热度的异质性。在特征提取阶段需要针对用户抽取热度、语义和关系信息,这些工作由对应不同的编码器来完成。

(1)热度编码器:社交网络用户的热度信息通常指的是用户在社交网络上的受欢迎程度、活跃度或影响力。首先,其中用户数字属性可以直接从社交平台 API 获取,编码器基于普遍性、直观性选择便于量化的 6 个数字(关注、被关注、点赞数、状态、活跃天数、屏幕名称长度),进行 z-score 归一化之后,应用两层 MLP 来学习用户数据的特征向量表示 。另选择 6 个用户状态数据特征(受保护、定位许可、验证、翻译许可、用户资料背景照片、默认配置文件图像)进行 One-hot 编码后经过全连接层,接着通过引入介数中心度来关注全局网络结构中的节点热度 $C_{\mathrm{B}}(v)$,具有高介数中心度的节点在网络中起到了关键的中介或桥接作用,能够控制或影响信息的流向。

$$C_{\mathrm{B}}(v)=\sum_{s\neq v\neq t}\frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

式(1)中:σ_{st} 为节点 s 到节点 t 之间所有可能的最短有向路径的总数;$\sigma_{st}(v)$ 是这些路径中经过节点 v 的路径数量。求和是对所有节点 s 和 t 的组合进行的,且 $s\neq v\neq t$,即节点 v 既不是路径的起点也不是终点。

$$f_i^{\mathrm{hot}}=f_i^{\mathrm{num}}+C_{\mathrm{B}}(v) \tag{2}$$

最后,将用户元数据、用户状态数据 f_i^{num} 与介数中心度 $C_{\mathrm{B}}(v)$ 合并形成用户热度编码 f_i^{hot}。

(2)语义编码器:使用预训练模型 RoBERTa[25] 将社交用户在平台上的发文作为语义编码的主要内容,同时将用户元数据中的(用户描述信息、个人签名)作为次要内容进行输入。

$$\overline{f_i^{\mathrm{text}}}=\varphi(W_i\mathrm{RoBERTa}\{t_i\}_{i=1}^{L}+b_i) \tag{3}$$

式(3)中:$\overline{f_i^{\mathrm{text}}}$ 为 RoBERTa 进行编码并将输出进行平均后得到用户的语义特征;t_i 为用户第 i 条发文,其长度为 $1\sim L$;φ 为激活函数 leaky-relu;W_i 和 b_i 为可学习参数。

(3)关系编码器:采用关系图转换器 RGT[24] 用来模拟用户之间不同关系的异构影响,并学习节点关系表示,如图 2 所示。具体来说,在实际的机器人检测任务中,把用户作为目标抽取联合特征表示是实用的、高效的,RGT 通过抽取节点与其他节点之间的关

系异质性(用户多维的活动产生)和影响异质性(由节点影响力决定),在注意力机制下有

$$\alpha_{c_{ij}}^{r}=\frac{\exp\left(\frac{\boldsymbol{q}_{c_i}^{r}\boldsymbol{k}_{c_j}^{r}}{\sqrt{D}}\right)}{\sum_{u\in N_{r(i)}}\exp\left(\frac{\boldsymbol{q}_{c_i}^{r}\boldsymbol{k}_{c_u}^{r}}{\sqrt{D}}\right)} \tag{4}$$

式(4)中:$\alpha_{c_{ij}}^{r}$ 为归一化分数,表示在给定关系和注意力头下,节点 j 对节点 i 的相对重要性;$\boldsymbol{q}_{c_i}^{r}$ 为节点 i 针对关系 r 和注意力头 c 的查询向量;$\boldsymbol{k}_{c_j}^{r}$ 为节点 j 针对同样的关系和注意力头的键向量;$\boldsymbol{k}_{c_u}^{r}$ 为每层中第 i 个节点隐藏表示 u 的键向量;D 为每个注意力头的隐藏层维度大小,用于缩放点积结果;$N_{r(i)}$ 为节点 i 在关系 r 下的邻居集合,点积表示注意力权重。

接着进行节点聚合表示,即

$$h_i^{r}=\tanh\left(\frac{1}{C}\sum_{c=1}^{c}\sum_{j\in N_{r(i)}}\alpha_{c_{ij}}^{r}v_{c_j}^{r}\right) \tag{5}$$

式(5)中:h_i^{r} 为节点 i 在关系 r 下聚合后的表示;C 为注意力头的数量。

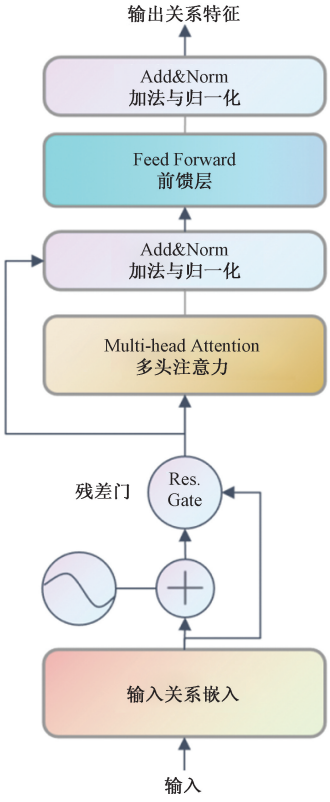

图 2 RGT 流程

Fig. 2 RGT flow

1.2 语义特征融合层

获得节点嵌入表示 α_i^{rht},对获得的节点局部特征和结合异构关系的全局特征进行抽取:对热度 f_i^{hot}、语义 f_i^{text} 特征取平均值得到用户节点特征 u_i^{ht},再将异构关系特征 h_i^{r} 与用户节点特征进行拼接得到节点的综合特征 $x_i^{\mathrm{rht}}\in\mathbf{R}^{N\times d}$。

通过引入多头注意力机制并行地关注不同的表示子空间来增强融合特征的表达能力。对于每个特征,通过线性变换得到对应的查询向量 $\boldsymbol{q}$、键向量 $\boldsymbol{k}$ 和值向量 $\boldsymbol{v}$。

$$\begin{cases}\boldsymbol{q}_i^{\mathrm{rht}}=x_i^{n-1}\boldsymbol{W}_i^{Q_n}+\boldsymbol{b}_i^{n}\\ \boldsymbol{k}_i^{\mathrm{rht}}=x_i^{n-1}\boldsymbol{W}_i^{K_n}+\boldsymbol{b}_i^{n}\\ \boldsymbol{v}_i^{\mathrm{rht}}=x_i^{n-1}\boldsymbol{W}_i^{V_n}+\boldsymbol{b}_i^{n}\end{cases} \tag{6}$$

式(6)中:$\boldsymbol{W}_i^{Q_n}$、$\boldsymbol{W}_i^{K_n}$、$\boldsymbol{W}_i^{V_n}$ 分别为查询、键和值的权重矩阵;$\boldsymbol{b}_i^{n}$ 为可学习偏置向量。

多头注意力在每个头 head_i 中分别计算注意力,融合不同特征,即

$$\begin{cases}\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})\\ \mathrm{Attention}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{v}\end{cases} \tag{7}$$

缩放点积注意力 Attention 使用 softmax 函数,其中 $\sqrt{d_k}$ 作为缩放因子用于稳定梯度,$\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}$ 是查询向量 $\boldsymbol{q}$ 与键向量 $\boldsymbol{k}^{\mathrm{T}}$ 的内积以得到每一对查询与键之间的相似度。最后,每个头的输出被聚合并通过另一线性变换进行整合,即

$$\alpha_i^{\mathrm{rht}(i)}=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_h)\boldsymbol{W}^{\mathrm{O}} \tag{8}$$

式(8)中:Concat 表示连接操作;$\boldsymbol{W}^{\mathrm{O}}$ 为另一个可学习的权重矩阵,用于将所有头的信息整合成最终的融合特征 $\alpha_i^{\mathrm{rht}(i)}$。

通过引入多头注意力机制,模型能够在不同的子空间中同时捕获多种特征间的复杂相互作用,从而生成一个全面且高度表达力的综合特征表示,以提高后续分类任务的效果。

1.3 变分推理

在 VRGAT 模型中,为节点引入潜在变量共同构成了一个潜在空间,潜在空间通常比原始输入空间更具表现力,能够更好地捕捉数据的内在结构和复杂性通过变分自编码器(variational auto-encoder, VAE)框架为每个节点引入潜在变量 z_i,这些变量捕捉了节点的潜在特性。

$$\mu_i,\lg\sigma_i^2=\mathrm{Encoder}(x_i^{I}) \tag{9}$$

式(9)中:μ_i 和 σ_i^2 分别为潜在变量 z_i 的均值和方差,x_i^{I} 为节点 i 在 VRGAT 模型最后一层 I 的输出,Encoder将 x_i^{I} 通过编码器映射到潜在变量 z_i 的分布 $q(z\mid x)$。由于在编码器中存在两个卷积操作,方差 $\alpha\mu^2$ 不能变为零,因此必须引入 Softplus 作为激活函数,如式(10),Softplus 是 ReLU 的平滑近似,它永远不会在 $x\to-\square$ 时为零,而 ReLU 在 $x\to-\square$ 时为零。

$$\mathrm{Softplus}(x)=\frac{1}{\beta}\lg[1+\exp(\beta x)] \tag{10}$$

式(10)中:β 用于调节函数平滑程度,由于重参化技巧需要 Softplus 函数能够提供足够的平滑性,同时不会导致过于平缓的反应,因此,此处默认 $\beta=1$。

最后,使用重参数化技巧使其能够高效反向传播,得到节点 i 的潜在表示 z_i。

$$z_i = \mu_i + \epsilon \odot \sigma_i \tag{11}$$

式(11)中:ϵ 表示从标准正态分布 $N(0,1)$ 中采样得到的随机噪声,它是一个随机变量,其目的是引入随机性,以模拟 z_i 的分布;$\odot$ 表示元素乘法(Hadamard乘法),用于将每个维度的随机噪声 ϵ 与对应的标准差 σ_i 相乘,以生成与 σ_i 相关的噪声项。

变分推理要求找到一个参数为 φ 的变分分布 $q_\varphi(\theta)$,其与后验分布 $p(\theta \mid X,Y)$ 尽可能近似。本文研究使用 Bayes by Backprop[26-27],用于学习权重 $\omega \sim q_\theta(\omega \mid D)$ 上后验分布的神经网络,其中 ω 可在反向传播中被采样。它通过最小化压缩成本来调整权重,压缩成本被称为变分自由能(边际似然的预期下界)。

采用本地重参数化技巧,通过对层激活而不是权重本身进行采样,实现计算加速,即

$$z_i = h_i^{\mathrm{ht}}\mu_i + \epsilon \sqrt{h_i^{\mathrm{ht}^2}(k_i\mu_i^2)}, \quad \epsilon \sim N(0,I) \tag{12}$$

式(12)中:k_i 为权重的方差与均值的乘积的系数;ϵ 为随机变量,使得在维持模型可微分性的同时考虑参数的不确定性,从而可以在模型训练过程中通过反向传播算法来优化这些参数。

1.4　输出层

在 VRGAT 中,以节点数作为总层数,关注与被关注关系作为异构图边集合 $\mathcal{R}$,对每个节点展开一张图,其中每一层都包含节点热度、语义和关系特征的抽取,经过 I 层变分推理编码器和重参数化之后得到最终的节点表示 z_i^I。最后结果 $\hat{y}_i$ 通过输出层和 softmax 层进行分类。

$$\hat{y}_i = \mathrm{softmax}[\boldsymbol{W}_0\sigma(\boldsymbol{W}_1 x_i^I + \boldsymbol{b}_1) + \boldsymbol{b}_0] \tag{13}$$

式(13)中:W_1 为输入层权重矩阵,用于将节点的输入特征 x_i^I 映射到隐藏空间;$\boldsymbol{W}_0$ 为输出层的权重矩阵,对输入的特征进行线性变换,将模型的隐藏层输出映射到输出空间;$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_0$ 分别为输入层的偏置和输出层的偏置;σ 为激活函数 Sigmoid。

1.5　损失函数

基于图神经网络的机器水军检测模型通常采用二元交叉熵损失(binary cross-entropy loss,BCE),但机器水军检测数据集中真实用户的数量往往远多于机器人,这导致类别不平衡的问题使得模型效果下降。因此本文模型采用改进的 Focal Loss 函数处理机器水军检测中的类别不平稳问题。

Focal Loss 通过增加难以分类的样本的相对损失,可以使模型更专注于学习区分这些难分的样本。损失函数包含三部分;第一部分是 Focal Loss 损失;第二部分是与数据相关的似然项,反映模型拟合数据的能力;第三部分是变分后验分布与先验分布之间的 KL 散度,引入先验知识。使用 KL 损失和正则化项来设计损失函数训练 VRGAT 以提高模型分类效果,即

$$\mathcal{L} = \mathcal{L}_{\mathrm{Focal}} + \lambda \sum_{w \in \theta} w^2 + \varepsilon \mathcal{L}_{\mathrm{KL}} \tag{14}$$

$$\begin{cases} \mathcal{L}_{\mathrm{KL}} = -\dfrac{1}{2}\displaystyle\sum_{j=1}^{d}(1 + \ln\sigma_{ij}^2 - \mu_{ij}^2 - \sigma_{ij}^2) \\ \mathcal{L}_{\mathrm{Focal}} = -\alpha_t(1-p_t)^\gamma \ln p_t \end{cases} \tag{15}$$

式中:λ、ε 为可调节参数,用来平衡似然项与 KL 损失对 $\mathcal{L}$ 的影响;i 为节点索引;j 为潜在空间维度;α_t 为模型将节点,检测为机器人样本的权重,用来平衡正负类别在损失函数中的贡献;p_t 为模型对机器人的预测概率;γ 为调节参数,用于减少易分类样本对总损失的贡献;w^2 为权重衰减(L2 正则化)项模型权重的平方和,用于防止过拟合。

算法 1:VRGAT

输入:社交网络机器水军检测图数据集 G
输出:优化后的模型参数 θ
1. 定义超参数 θ
2. 用户热度特征 f_i^{hot}、发文语义特征 f_i^{text}
3. 使用指定的超参数初始化 VRGAT
4. 使用 PyTorch Lightning 设置模型训练,并为模型检查点添加回调函数
5. while θ 尚未收敛 do
6. for $i \in I$ do
7. for 每个节点 i 邻接关系 do
8. RGT 得到节点异质关系
9. $u_i^r \leftarrow$节点关系特征
10. end
11. Average u_i^r
12. $f_i^{\mathrm{hot}} + f_i^{\mathrm{text}} = u_i^{\mathrm{ht}}$
13. for $c \leftarrow 1$ to C
14. for $j \in N_i^r$ do
15. $k_c^{r(i)}, q_c^{r(i)}, v_c^{r(i)}$
16. $\alpha_{c,ij}^{r(i)}$
17. end
18. end
19. $\alpha_i^{\mathrm{ht}(i)} \leftarrow$关系、语义、热度融合特征
20. VAE 变分编码器
21. μ_i, σ_i
22. 重参化技巧
23. $h_i^{\mathrm{ht}(i)} \leftarrow$节点隐藏特征
24. end
25. $z_i^I \leftarrow$节点潜在表示
26. Loss
27. $\theta \rightarrow$BackPropagate
28. end
29. return θ

2 实验结果及分析

本节评估 VRGAT 机器水军检测模型的性能。首先,对实验中的基本设置进行介绍,包括实验所用数据集和实验环境设置;接着将本文模型与 10 个基线模型方法在 Twibot-20 数据集上进行比较,以验证本文模型在检测机器水军任务中的准确率优越性;随后进行了消融实验,验证变分推理模块对模型结果的影响以及必要性;最后,探究模型加入随机噪声后的噪声敏感性以及鲁棒性。

2.1 数据集

小规模数据集不足以训练和稳定地衡量新的机器人检测措施,又由于高级机器水军的模糊性,为机器水军检测获得准确可靠的标签是困难的、成本高昂的、充满噪声的,这对机器水军检测数据集的建立带来了挑战[6]。TwiBot-20 是一个公开的大规模 Twitter 机器人检测数据集[5],实验遵循原基准测试中的数据集、验证集和测试集的划分以确保可对比性,数据集包含 229 573 个用户,其中标签有 5 237个人类用户和6 589 个机器人用户,33 488 192 条推文,8 723 736 个用户属性项目和 455 958 个关注关系。

2.2 实验设置

实验环境为 16 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz,120 G 内存,Ubantu22.04 操作系统,GPU 为 RTX 4090(24 GB),python版本为 3.10,深度学习框架为 pytorch_lightning,版本 0.10.1。VRGAT 超参数设置如表 1 所示。

采用准确率(A_{ccuracy})、精确率(P_{recision})、召回率(R_{ecall})和 $F_{\text{1-score}}$来评价模型的整体性能,它们被广泛应用于监督学习分类任务中的评价指标。

表 1 VRGAT 超参数设置

Table 1 Hyperparameter settings of VRGAT

超参数	值
优化器	AdamW
学习率	10^{-3}
L2 正则化 λ	3×10^{-5}
batch size	256
层数 L	2
dropout	0.5
KL 散度权重 β	1.2
隐藏层大小	128
最大 epochs	40
特征融合注意力头 C 数量	4
RGT 注意力头 D 数量	8
关系边集合 R	{follower,following}

$$A_{\text{ccuracy}} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

$$P_{\text{recision}} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$R_{\text{ecall}} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$F_{\text{1-score}} = 2\frac{R_{\text{ecall}}P_{\text{recision}}}{R_{\text{ecall}}+\text{P}_{\text{recision}}} \tag{15}$$

式(16)中:TP 为真阳性,表示模型正确地识别出机器人的数量;FP 为假阳性,表示模型错误地将正常用户识别为机器人的数量;TN 为真阴性,表示模型正确地将一个正常用户识别为正常用户的数量;FN 为假阴性,表示模型错误地将机器人挥别为正常用户的数量。

2.3 机器水军检测结果

在 Twibot-20 上对机器水军检测模型进行基准测试,如表 2 所示,对比算法的结果均使用开放的源代码和推荐参数进行实验所得,结果表明:①本文模型 VRGAT 优于所有的基线。将 VRGAT 与几种最先进的机器人检测模型进行对比,并增加了其他方法如支持向量机(support vector machine, SVM)、随机森林(random forest, RF)、NLP 进行比较;②通过以节点为中心抽取 n 个节点层的VRGAT,将社交网络的语义、热度与异构关系特征用注意力机制连接为节点特征联合表示,能够有效概括节点特征;③融合变分推理使得模型能够在高维空间中抽象出机器水军网络的潜在特征。

表 2 基于 Twibot20 数据集的基准测试

Table 2 Benchmarking on the Twibot20 dataset

方法	A_{ccuracy}	$F_{\text{1-score}}$	R_{ecall}
SVM[28]	0.728 9	0.764 6	0.804 0
RF[29]	0.819 1	0.854 6	0.801 7
Adaboost[30]	0.698 4	0.716 6	0.735 8
RoBERTa[25]	0.712 6	0.753 3	0.808 6
GCN[31]	0.749 2	0.750 4	0.751 6
GAT[32]	0.842 2	0.868 7	0.875 9
Botometer[33]	0.558 4	0.489 2	0.555 8
SATAR[23]	0.841 2	0.864 2	0.886 3
BotRGCN[34]	0.846 2	0.870 7	0.872 1
RGT[24]	0.866 4	0.881 2	0.889 1
VRGAT	0.880 2	0.896 4	0.892 5

通过构建隐藏空间,变分推理可以捕捉数据中的潜在结构和相关性,生成新的数据点以增强数据,以弥补数据集有限且不平衡缺陷下的训练效果。

2.4 消融实验

本文提出的 VRGAT 机器水军检测模型融合了语义、热度和异构关系特征作为节点联合特征进行

学习来捕捉机器水军与正常用户之间微妙的差别,以更好地识别机器人。为了验证 VRGAT 的有效性进行了消融实验,实验分别去除热度特征、语义特征和异构关系特征,将变分关系图注意力 Transformer应用在构建的 mHIN 上以学习节点表示,在实验中对特征融合前的注意力机制选用和、平均池化、最大池化和最小池化做注意力,分别将 VRGAT 替换为 GCN 和 SAGE,如表 3 所示。实验结果表明本文构建的语义热度关系图结构能够完整地表示节点特征,在分类结果上体现出优势。因此变分编码器、注意力机制是不可缺少的模型组件。

表 3　VRGAT 模型在 Twibot-20 数据集上的消融实验

Table 3　Ablation experiments of the model on Twibot-20 dataset

类别	hot	text	同质关系	同质关系	A_{ccuracy}	F_1
w/o hot	û	ü	ü	û	0.861 4	0.874 0
w/o text	ü	û	ü	û	0.836 8	0.851 1
w/o r	ü	ü	û	û	0.652 9	0.662 5
同质 r	û	û	û	ü	0.848 1	0.861 3
w/max	ü	ü	ü	û	0.861 7	0.871 9
w/min	ü	ü	ü	û	0.870 6	0.886 3
w/sum	ü	ü	ü	û	0.875 2	0.878 7
w/mean	ü	ü	ü	û	0.866 3	0.875 2
w/RGCN	ü	ü	ü	û	0.848 4	0.852 6
w/SAGE	ü	ü	ü	û	0.837 1	0.839 5
VRGAT(default)	ü	ü	ü	û	0.880 2	0.896 4

注:“w/”表示“包括”;“w/o”表示“不包括”;“û”表示不使用;“ü”表示使用;ot、text 分别为热度、语义特征。

如图 3 所示,使用多头注意力机制来影响节点的联合异质性,结果表明本文提出的 VRGAT 模型构建的融合语义、关系和热度的多元异构的信息网络(mHIN),节点特征往往由用户发文、不同的关系和热度来共同组成,可以通过学习节点的联合特征来学习节点之间细微区别的模式,有利于下一阶段神经网络的学习进而完成分类;为了验证联合特征的异质性的有效性,对模型使用的多头注意力机制和 RGT 中的语义注意力网络进行消融实验。结果表明,RGT 语义注意力头选择在 8,特征注意力网络头选择在 4 的效果最优。

2.5　鲁棒性分析

通过实验评估了模型在对噪声数据(包括对抗攻击噪声)的敏感性。为检验模型对噪声数据的敏感性,本文研究通过向数据集中添加随机噪声和错误标签,测试模型在面对数据质量问题时的鲁棒性。机器水军检测数据集的收集面临多重挑战,手工数据标注成本高且易出错[6],社交平台对数据访问有严格限制以保护用户隐私、水军行为多样性等

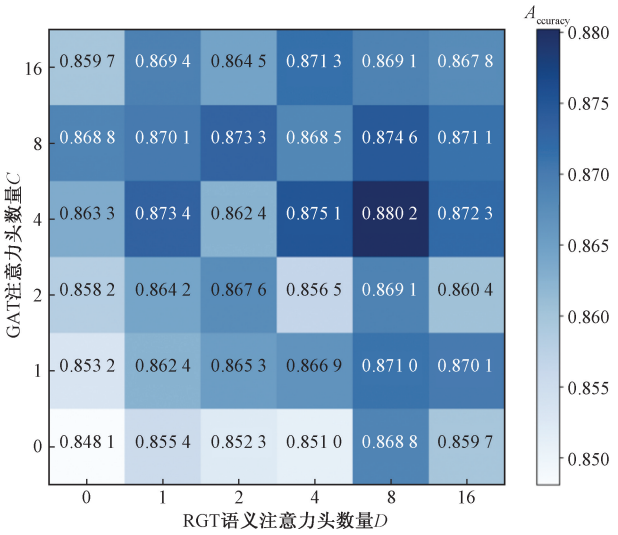

图 3　VRGAT 中的特征注意力网络和关系图转换器中的语义注意力网络的消融实验

Fig. 3　Ablation experiments of feature attention network in VRGAT and semantic attention network in relational graph transformer

因素导致数据集充斥着噪声与数据缺失,因此机器水军检测模型需要具备较好的抵抗噪声的能力,进行了噪声实验。

以 0、5%、10%、15%、20%、25%、30% 的比例为 Twibot20 数据集添加随机噪声,并通过构造特定的对抗样本来测试模型是否容易受到欺骗。对文本数据进行随机同义词替换、插入错别字和删除字符,在用户的点赞和评论数加入随机误差以反映用户随机的互动行为,机器人用户,用正常用户元数据替换它们的用户元数据以对抗热度异质性,随机替换了用户发文以对抗语义异质性,随机新增对邻居节点边的关系以对抗关系异质性。结果如图 4 所示,由于 VRGAT 融合了变分推理使得对抗机器人节点强行伪造特征并不能显著改变节点的特征分布,模型在噪声比例升高时准确率衰减相较其他模型最少,VRGAT 的抵抗能力相较其他模型最优。对 VRGAT 进行训练集大小对模型效果影响的实验,已评估模型对于新数据的适应能力和鲁棒性。在控制训练集大小等差从 2 500 至 30 000 的过程中,F_1 值快速接近模型可实现的最优分类结果,如图 5 所示。表明变分推理通过近似推理来估计复杂分布,这在处理大规模图数据时尤其有效。当训练数据量增加时,变分推理可以更准确地估计节点的潜在表示,更好地利用了其内部的复杂性和容量,模型能够学习到了更多的有用特征和关系来更好地捕捉图的全局结构和节点间的依赖关系,从而提高模型在未见数据上的泛化能力。

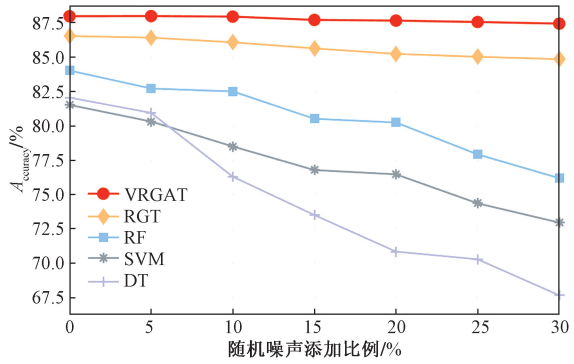

图 4 随机噪声对不同模型效果的影响

Fig. 4 Effect of random noise on different models

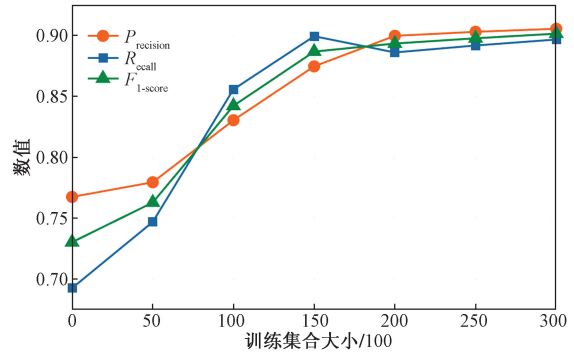

图 5 训练集大小对 VRGAT 效果的影响

Fig. 5 Effect of training set size on VRGAT

3 结论

由于机器水军不断迭代,自动养号造成的高隐蔽性使得已有的机器水军检测模型有一些缺陷:通过分析节点间局部异构影响力关系无法充分体现机器水军网络全局特征;在处理数据的复杂性、噪声问题以及跨平台应用的泛化能力上仍存在挑战。本文研究提出了一种结合变分推理的图神经网络模型——VRGAT,以检测社交网络机器人。通过深入分析和实验验证,本文模型在 Twibot-20 数据集上展示了卓越的性能,得出了以下结论。

(1)提出了一种融合变分推理的关系图神经网络模型,旨在优化机器水军检测问题中数据集的信息缺失和数据噪声问题。通过对每个节点的表示引入概率分布,模拟社交网络用户行为的不确定性和潜在的表示分布,从而更好地处理信息缺失和噪声问题。

(2)针对节点特征表示不全的问题,本文研究构造了融合节点热度、语义和异构关系特征的多元异构网络 mHIN,以更全面地表征节点特征,有效提升了节点分类的准确率。

(3)基于真实的 Twitter 社交网络环境数据进行多基线实验、消融实验和噪声实验。结果表明,在机器水军检测的准确率方面,与现有的 GNN 方法相比,VRGAT 的 $F_{1\text{-score}}$ 提升了 1.52%,验证了本文方法的优越性和鲁棒性。

综上所述,本文研究证明了结合变分推理的异构图神经网络的方法在机器水军检测领域的有效性和先进性。未来工作中需要进一步探索涉及音频、图像视频数据的复杂机器人攻击场景下的检测,以及机器水军跨平台个人资料和内容的一致性,进一步提升模型的泛化能力和可解释性。本文研究为社交网络安全领域提供了一个强大的新工具,为打击恶意机器人活动开辟了新的可能性。

参 考 文 献

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

[2] Kresovich A, Andrew H N, Chandler C C, et al. Deciphering influence on social media: a comparative analysis of influential account detection metrics in the context of tobacco promotion[J]. Social Media Society, 2024, 10(1). DOI:10.1177/20563051231224268.

[3] Wang R, Walter D, Ophir Y. Not all bots are created equal: the impact of bots classification techniques on identification of discursive behaviors around the COVID-19 vaccine and climate change[J], Social Science Computer Reciew, 2024, 42(2): 394-415.

[4] Zhang Y N, Chen F, Rohe K. Social media public opinion as flocks in a murmuration: conceptualizing and measuring opinion expression on social media[J]. Journal of Computer-Mediated Communication, 2021, 27(1). DOI:10.1093/jcmc/zmac002.

[5] Feng S B, Wan H, Wang N N, et al. TwiBot-20: a comprehensive twitter bot detection benchmark[C]//International Conference on Information and Knowledge Management. New York: ACM (Association for Computing Machinery), 2021: 4485-4494.

[6] Hays C, Schutzman Z, Raghavan M, et al. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection[J]. Computing Research Repository, 2023, 2023: 3660-3669.

[7] Wu J, Ye X S, Mou C J. BotShape: a novel social bots detection approach via behavioral patterns[J]. Computing Research Repository, 2023, 2023: 45-60.

[8] 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.

[9] Kudugunta S, Ferrara E. Deep neural networks for bot detection[J]. Information Sciences, 2018, 467: 312-322.

[10] Yang C, Harkreader R C, Gu G F. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers[J]. Recent Advances in Intrusion Detection, 2011, 6961: 318-337.

[11] Costa A F, Yamaguchi Y, Traina A J M, et al. RSC: mining and modeling temporal activity in social media[C]//ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM (Association for Computing Machinery), 2015: 269-278.

[12] Hurtado S, Ray P, Marculescu R. Bot detection in reddit political discussion[C]//Proceedings of the Fourth International Workshop on Social Sensing. New York: IEEE, 2019: 30-35.

[13] 邱桂华, 李贤阳, 程宪宝, 等. 社交网络中水军用户团队不当行为准确识别技术[J]. 科学技术与工程, 2019, 19(7): 177-182.
Qiu Guihua, Li Xianyang, Cheng Xianbao, et al. Research on accurate identification of misconduct of water user team in social network[J]. Science Technology and Engineering, 2019, 19(7): 177-182.

[14] 高东伟. 在线社交网络中用户伪装攻击检测方法研究[J]. 科学技术与工程, 2017, 17(7): 194-198.
Gao Dongwei. Online social network users in disguise attack detection method research[J]. Science Technology and Engineering, 2017, 17(7): 194-198.

[15] Varol O, Ferrara E, Davis C A, et al. Online human-bot interactions: detection, estimation, and characterization[C]//International AAAI Conference on Weblogs and Social Media. Palo Alto: AAAI Press, 2017: 280-289.

[16] Sajjad H, Alani H, Rashid A. Lexical and syntactic features for identifying abnormal behavior in social media[C]//IEEE Access. New York: IEEE, 2019. DOI: 10.1109/ACCESS.2019.2904462.

[17] 杜锐, 朱艳辉, 邓程, 等. 一种基于粗糙集的微博文本特征选择方法[J]. 科学技术与工程, 2013, 13(33): 9830-9834.
Du Rui, Zhu Yanhui, Deng Cheng, et al. Micro blog text feature selection based on rough set[J]. Science Technology and Engineering, 2013, 13(33): 9830-9834.

[18] Feng W, Trang N U. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings[C]// First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). New York: IEEE, 2019: 101-109.

[19] David D, Dominik K, Dominik S. Are you human? detecting bots on twitter using BERT[C]//International Conference on Data Science and Advanced Analytics. New York: IEEE, 2020: 631-636.

[20] 姜钰棋, 侯智文, 王一帆, 等. 社交平台不平衡文本数据处理与应用研究[J]. 计算机科学与探索, 2024, 18(9): 2370-2383.
Jiang Yuqi, Hou Zhiwen, Wang Yifan, et al. Research on processing and application of imbalanced textual data on social platforms[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(9): 2370-2383.

[21] Alhosseini S A, Tareaf R B, Meinel C. Engaging with tweets: the missing dataset on social media[C/OL]//Conference on Recommender Systems. 2020: 34-37.

[22] Alothali E, Salih M, Hayawi K, et al. Bot-MGAT: a transfer learning model based on a multi-view graph attention network to detect social bots[J]. Applied Sciences-Basel, 2022, 12(16): 8117-8117.

[23] 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[J]. arXiv:2106.13089.

[24] Feng S B, Tan Z X, Li R, et al. Heterogeneity-aware twitter bot detection with relational graph transformers[C]//AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2022: 3977-3985.

[25] Liu Y H, Myle O, Naman G, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv preprint, arXiv: 1907.11692, 2019.

[26] Graves A. Practical variational inference for neural networks[C]//Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Cambridge: MIT Press, 2011: 2348-2356.

[27] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural networks[J]. arXiv preprint arXiv: 1505.05424, 2015.

[28] Wang H, Wang Y, Leskovec J. Graph convolutional networks with SVM for node classification[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017). Cambridge: MIT Press, DOI: 10.5555/3294996.3295057.

[29] Fu H, Xie X, Rui Y. Leveraging careful microblog users for spammer detection[C]//Proceedings of the 24th International Conference on World Wide Web. New York: Association for Computing Machinery, 2015: 419-420.

[30] Zineb E, Faouzia B, Sara O, et al. Word embedding for social bot detection systems[C]//Proceedings of the 2021 5th International Conference on Intel ligent Computing in Data Sciences (ICDS). Fez: IEEE, 2021: 1-7.

[31] Seyed A A, Raad B T, Pejman N, et al. Detect me if you can: spam bot detection using inductive representation learning[C]// Companion Procedings of the 2019 World Wide Web Conference. San Francisco: Asociation for Computing Machinery, 2019: 148-149.

[32] Veličkovic P, Cucurull G, Casanova A, et al. Graph attention networks[C]//Proceedings of the 7th International Conference on Learning Representations. New York: IEEE, 2018: 1-12.

[33] Hanschmann L, Gnewuch U, Maedche A. BotOrNot: a platform for conducting experiments with undisclosed chat agents[C]// Mensch and Computer. Bonn: GI(German Informatics Society), 2022: 618-621.

[34] Feng S B, Wan H R, Wang N N, et al. BotRGCN: twitter bot detection with relational graph convolutional networks[C]//Advances in Social Networks Analysis and Mining. New York: IEEE, 2021: 236-239.