SeGA: 基于偏好感知的 Twitter 异常用户检测提示自对比学习

摘要

在动态且快速发展的社交媒体世界中，检测异常用户已成为应对虚假信息和网络欺凌等恶意活动的关键任务。随着异常用户数量的增加，机器人模仿正常用户和逃避检测的能力不断提高，现有的只关注机器人检测的方法在捕捉用户之间细微差别方面效果不佳。为了应对这些挑战，提出了 SeGA (preference-aware self-contrast learning for anomalous user detection)，该方法利用 twitter 圈中异构实体及其关系，采用不同的恶意策略来检测异常用户。SeGA 利用大型语言模型的知识通过帖子总结用户偏好。此外，将用户偏好与提示信息融合为伪标签进行偏好感知的自对比学习，使模型能够从多个方面学习描述用户的行为。在 TwBNT 基准上的大量实验表明，SeGA 显著优于目前最先进的方法 (+3.5% 27.6%)，并实证验证了模型设计和预训练策略的有效性。

关键词：社交媒体；异常用户检测；偏好感知；自对比学习

1 引言

异常用户检测的研究在各个领域都具有广泛的适用性。无论是开发网络安全策略 (Bilot et al. 2023) [1]，金融交易 (Chai et al. 2022) [2]，还是社交媒体分析 (Agarwal et al. 2022; Wang and Peng 2022) [3, 4] 等等，这些场景都可以有效地建立以用户之间错综复杂的关系为特征的异常用户检测系统。Twitter 作为使用最广泛的社交媒体平台之一，近年来随着用户对内容偏好的多元化也导致了异常用户数量的快速增长。例如，异常用户经常发布恶意内容，破坏社区的正常聊天环境，这违反了社交媒体平台上健康的在线讨论原则 (Alieva, Moffitt, and Carley 2022) [5]。最常见的异常用户类型之一是机器人，在活跃用户中占 9%到 15%(Varol et al. 2017) [6]。2017 年底，美国国会披露了一份 2.7 万个被确认为俄罗斯喷子的 Twitter 账户清单 (Zannettou et al. 2019) [7]，这是一种新兴的异常用户类型。随着异常用户类型的增加，以及其对社交媒体的潜在负面影响日益明显，因此提出能够同时识别机器人和喷子的异常检测方法至关重要。之前的工作主要集中在识别机器人，并通过利用图神经网络 (GNNs) (Kipf and Welling 2017; Velickovic et al. 2018) [8, 9] 和异质信息网络 (HINs) (Feng et al. 2021c, 2022a) [10, 11] 的拓扑结构来实现对异常用户的有效检测。然而，随着异常用户的多样性和恶意活动的进化策略不断扩大，现有方法缺乏有效区分机器人和喷子等各类异常用户的能力。检测喷子和机器人的主要区别在于前者是由人类控制的，与后者相比，前者的行为与正常用户相似。为了解决这些问题，文献提出了一种偏好感知的 Twitter 异常用户检测自对比学习方法 SeGA，利用异构编码器编码

Twitter 上不同实体的异构关系，以及提出的伪标签的自对比学习，通过相应的帖子根据用户偏好来辨别用户之间的细微差异。

2 相关工作

早期的 Twitter 异常用户检测模型专注于检测具有用户特征或推文的机器人账户 (Miller et al. 2014; Cresci et al. 2016) [12, 13]。随着图神经网络的出现，越来越多的基于图的机器人检测器被提出，它们将用户及其交互表示为社交图，并利用聚合技术从邻近节点收集信息。

2.1 图神经网络

GCN (Kipf and Welling 2017) [8] 平等地聚合来自邻近用户的特征以学习表示，而 GAT (Velickovic et al. 2018) [9] 使用注意力机制对用户影响进行建模。另一方面，由于异构图能够有效地表示具有不同节点和边类型的社交网络，因此也被用于僵尸网络检测。Lv et al. (2021) [14] 采用不同的策略增强 GAT 在异构图上的表现，Feng et al.(2022a) [11] 提出关系图 transformer 对异构关系进行建模并影响用户之间的异构性。

2.2 自监督学习

近年来，自监督学习 (SSL) 已被证明是通过伪装任务学习上下文表示的一种强大而有效的方法，如自然语言处理 (Gao, Yao, and Chen 2021) [15] 和计算机视觉 (Lv et al. 2021; Bardes, Ponce, and LeCun 2022) [13, 16]。作为 SSL 的一个分支，预测学习也被应用于 bot 检测，取得了很好的结果。Feng et al. (2021a) [17] 将关注者计数作为自监督学习信号，以增强模型在 bot 检测中的性能。然而，依赖单一特征作为自监督指标忽略了异常用户之间的多样性，无法充分表示不同用户。例如，喷子用户可以构建被其他喷子用户关注的人工粉丝数。因此，可以采用 llm 获取的用户偏好和相应的帖子来总结多方面的行为 (文中的主题和情感) 来描述用户，进行自对比学习。

3 本文方法

3.1 本文方法概述

SeGA 的框架如图 1 所示，它由三个阶段组成: 节点特征编码、预训练阶段和微调阶段。SeGA 利用用户和列表作为 HIN 中的节点来捕获这些实体之间的不同关系。首先编码三种类型的节点特征 (即指示器、数值和文本) 来获取节点的特征信息。在预训练阶段，使用偏好感知的自对比学习对编码器进行预训练，以学习 llm 从帖子中总结的用户偏好。在微调阶段，对预训练模型进行微调，对异常用户进行分类。 1

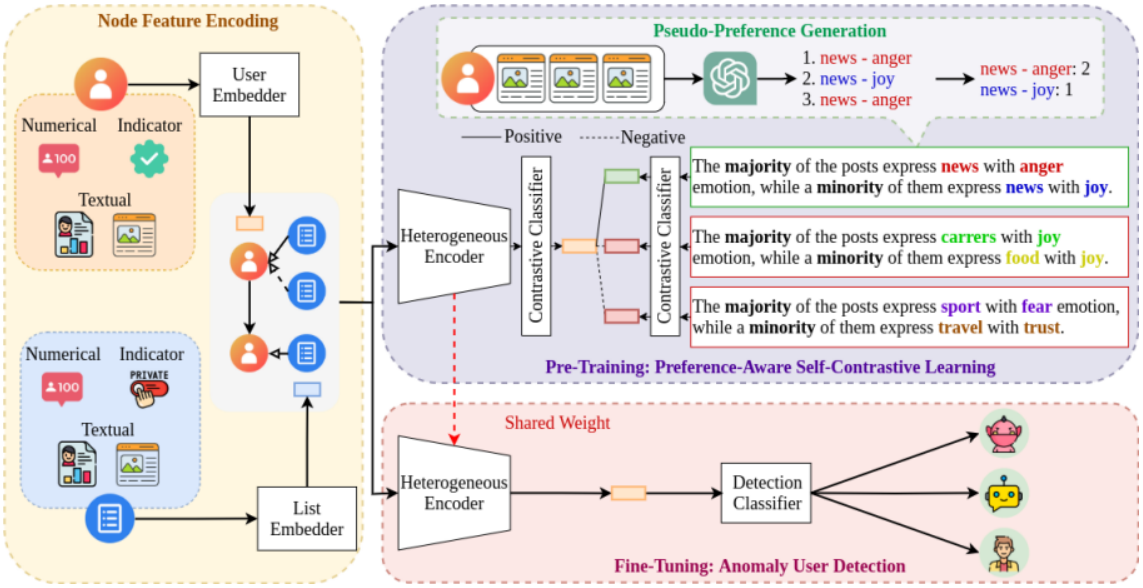

图 1. SeGA 框架示意图

3.2 节点特征编码

类似于特征编码为 (Feng et al. 2022a) [11] 的过程，我们首先将每个节点 i 的每个指示符特征与节点类型 A 连接起来:

$$x_i^{A_{ind}} = c_{i1} \oplus c_{i2} \oplus \cdots \oplus c_{ik}, \tag{1}$$

其中 $\oplus$ 是连接操作符，$x_i^{Aind} \in \mathrm{R}^k$ 是由 k 特征连接的嵌入指示器。

类似地，数值特征连接为:

$$x_i^{A_{num}} = n_{i1} \oplus n_{i2} \oplus \cdots \oplus n_{im}, \tag{2}$$

其中 $x_i^{Anum} \in \mathrm{R}^m$ 是由 m 特征连接的节点 i 的数字嵌入。

为了编码文本特征，使用预训练的 RoBERTa (Liu et al. 2019) [18] 进行 s 单词编码:

$$x_i^{A_{des}} = RoBERTa\left(\left\{w_{ij}^{des}\right\}_{j=1}^{s}\right), \tag{3}$$

其中 $x_i^{A_{ind}} \in \mathrm{R}^{d_{des}}$ 表示描述嵌入。同样，通过使用预训练的 $RoBERTa$ 对每条推文编码的所有嵌入进行平均来获得推文嵌入 $x_{i_j}^{A_{twe}} \in \mathrm{R}^{d_{twe}}$:

$$x_{i_j}^{A_{twe}} = RoBERTa\left(\left\{w_{i_{jy}}^{twe}\right\}_{y=1}^{L}\right), \tag{4}$$

$$x_i^{A_{twe}} = AVG\left(x_{i_1}^{A_{twe}}, x_{i_2}^{A_{twe}}, \cdots, x_{i_q}^{A_{twe}}\right), \tag{5}$$

然后，分别对它们进行变换。

$$x_i'^{A_F} = \sigma\left(W_x \cdot x_i^{A_F} + b_x\right), F \in \{ind, num, des, twe\} \tag{6}$$

其中 $x_i'^{A_F} \in \mathrm{R}^{d_h}, W_x, b_x$ 是可训练参数，σ 表示为 LeakyReLU。

对于每个用户或列表节点 i，通过连接指示器 $x_i'^{A_{ind}}$，$x_i'^{A_{num}}$，描述 $x_i'^{A_{des}}$ 和 tweet $x_i'^{A_{twe}}$ 嵌入来获得用户/列表嵌入 $x_i^A \in \mathrm{R}^{4*d_h}$：

$$x_i^A = x_i'^{A_{ind}} \oplus x_i'^{A_{num}} \oplus x_i'^{A_{des}} \oplus x_i'^{A_{twe}} \tag{7}$$

然后，将用户/列表嵌入 x_i^A 转换为异构编码器的初始节点嵌入 $z_i^{A,(0)} \in \mathrm{R}^{4*d_h}$：

$$z_i^{A,(0)} = \sigma\left(W_I \cdot x_i^A + b_I\right) \tag{8}$$

其中 W_I 和 b_I 是可训练的参数。

为了对各种实体及其不同重要性的不同关系进行建模，以丰富用户的嵌入，采用了 (Feng et al. 2022a) [18] 之后的关系图 transformer(RGT) 和具有激活函数的 MLP 作为异构编码器:

$$z_i^{u,(g)}, z_i^{l,(g)} = RGT^{(g)}\left(z_i^{u,(g-1)}, z_i^{l,(g-1)}\right) \tag{9}$$

$$z_i^u = \sigma\left(W_z \cdot z_i^{u,(g)} + b_z\right) \tag{10}$$

其中 $z_i^{u,(g)}, z_i^{l,(g)} \in \mathrm{R}^{d_{out}}$，是从 $g-th$ 层学习到的 $i-th$ 节点嵌入，$z_i^u \in \mathrm{R}^{d_u}$ 是 $i-th$ 用户表示，W_z 和 b_z 是可训练的参数。注意到，由于任务是对用户类别进行分类，所以采用聚合用户嵌入 z_i^u 进行偏好学习和异常检测。

3.3 基于偏好感知的自对比学习

用户偏好代表了个体的行为，反映了个体在各个方面的选择，有利于用户行为模式的检测。因此，文献引入了伪偏好生成，它是基于用户的历史帖子从 llm 中总结出来的。为了描述和学习用户偏好，设计了一个提示模板来表示用户帖子的多数主题-情感对和少数主题-情感对，从而进行偏好感知的自对比学习。鉴于在自然语言应用中利用 llm 的成功范例 (Wu, Zhang, and Huang 2023) [19]，本文纳入了 llm 的强大知识，以从帖子中检索用户偏好。具体来说，我们选择用户偏好的主题和情感来表示每个用户对相应帖子的偏好，因为异常用户可能利用它们来达到恶意目的 (Ghanem, Buscaldi, and Rosso 2020; Balasubramanian et al. 2022) [20,21]。用户 i 最近的 10 个帖子被用作 LLM 的提示，以生成主题 t 和每条推文 j 中使用的情感 e:

$$\left\{\left(t_{i_j}^u, e_{i_j}^u\right)_{j=1}^{10}\right\} = LLM\left(P\left(twe_{i_1}^u \oplus \cdots \oplus twe_{i_{10}}^u\right)\right) \tag{11}$$

其中 $P()$ 是用于从 LLMs 中获取用户偏好的指令提示(Prompt)，LLM 在文献中指的是 ChatGPT (Liu et al.2023) [22]。话题是基于 Twitter 话题 (Jim and Ann 2020) [23] 从 16 个类别衍生出来的，8 种情绪是基于普鲁契克的情绪 (Ghanem, Buscaldi, and Rosso 2020) [20]。

在获得每个用户的主题-情感对后，我们的目标是利用这些伪信息对模型进行预训练，以描述用户的偏好。一种直观的方法是预测用户每篇文章的所有主题和情感，这可以看作是一个多标签分类任务。然而，由于模型对所有标签一视同仁，没有考虑它们的相对重要性，因此无法有效捕获用户偏好的主题和情感。例如，一个用户可能主要是愤怒地发布一个新闻话题，而很少高兴地发布一个新闻话题。为解决该问题，文献提出了一种偏好感知的自对比学习方法，将偏好与提示相结合以增强学习。文献中将用户的偏好定义为使用频率最高的主题-情感对 $\left(t_{i_{\max}}^u, e_{i_{\max}}^u\right)$，和使用频率最低的主题-情感对 $\left(t_{i_{\min}}^u, e_{i_{\min}}^u\right)$ 的综合，以反映用户的兴趣和情感

行为。我们为每个用户生成伪标签 $p_i^u i$，利用设计的提示模板 $PT()$，其中填充了最常用和最不常用的主题情感对:

$$p_i^u = PT\left(\left(t_{i_{\max}}^u, e_{i_{\max}}^u\right), \left(t_{i_{\min}}^u, e_{i_{\min}}^u\right)\right) \tag{12}$$

提示模板被设计为: “大多数帖子表达 $t_{i_{\max}}^u$ 的主题带有 $e_{i_{\max}}^u$ 的情感，而少数帖子表达 $t_{i_{\min}}^u$ 的主题带有 $e_{i_{\min}}^u$ 的情感。” 然后，使用 SimCSE RoBERTa (Gao, Yao, and Chen 2021) [24] 的提示编码器对用户 i 的伪标签进行编码，该编码器从对比学习中学习句子嵌入:

$$p_i'^u = SimCSE\left(p_i^u\right) \tag{13}$$

其中 $p_i'^u \in \mathrm{R}^{d_p}$ 是用户 i 的伪标签嵌入。然后我们使用对比分类器对用户嵌入 z_i^u 和伪标签嵌入 $p_i'^u$ 进行转换:

$$\tilde{z}_i^u = W_{\tilde{z}} \cdot z_i^u + b_{\tilde{z}} \tag{14}$$

$$\tilde{p}_i^u = W_{\tilde{p}} \cdot p_i'^u + b_{\tilde{p}} \tag{15}$$

其中 $\tilde{z}_i^u \in \mathrm{R}^{d_a}$ 表示用户 i 的锚用户嵌入，$\tilde{p}_i^u \in \mathrm{R}^{d_a}$ 表示相应的正样本嵌入，$W_{\tilde{z}}$，$b_{\tilde{z}}$，$W_{\tilde{p}}$ 和 $b_{\tilde{p}}$ 是可训练参数。

为了计算对比损失 L_{pre}，我们将嵌入 $\tilde{p}_i^u$ 定义为锚用户嵌入 $\tilde{z}_i^u$ 的正对，从其他伪标签转换而来的嵌入视为锚用户的负对:

$$L_{pre} = -\sum_{i \in U} \log \left(\frac{\exp(sim\left(\tilde{z}_i^u \cdot \tilde{p}_i^u\right)/\tau)}{\sum_{j \in S(i)} \exp(sim\left(\tilde{z}_i^u \cdot \tilde{p}_j^u\right)/\tau)} \right) \tag{16}$$

其中 U 是所有用户节点的索引集合，$S(i)$ 是一个正对的集合，负对是从所有负对中随机抽样得到的。在最小化对比损失时，具有相同伪标签的用户嵌入往往更接近，同时鼓励编码器学习用户偏好的主题和情感。

3.4 异常用户分类

为了实现异常用户检测的目标，预训练模型中的用户嵌入 z_i^u 被用于具有 $softmax$ 层的检测分类器中，以预测每个用户的类别 $\hat{y}_i$:

$$\hat{y}_i = soft\max\left(W_y \cdot z_i^u + b_y\right) \tag{17}$$

其中 W_y 和 b_y 是可训练的参数。

最后，联合微调预训练的嵌入器、预训练的编码器和检测分类器，实现异常用户检测。微调过程包含交叉熵损失和 L2 正则项，如下所示:

$$L_{fine} = -\sum_{i \in U} \left[y_i \log\left(\hat{y}_i\right)\right] + \lambda \sum_{\omega \in \theta} \omega^2 \tag{18}$$

其中 U 是所有用户节点的索引集合，y_i 表示用户 i 的真实标签，λ 是一个超参数，包含所有可训练参数。

4 复现细节

4.1 与已有开源代码对比

在特征编码阶段，用户和列表节点的指标特征个数 k 分别设置为 3 和 1，数值特征个数 m 分别设置为 5 和 4。对数值特征 N 进行 $z-score$ 归一化处理。d_{des}, d_{twe}, d_p 的大小为 768；d_h 的大小为 32，d_{out} 的大小为 128，d_u, d_a 的大小为 64；Relational Graph Transformer 的层数 g 为 2。在 (Feng et al. 2022a) [11] 之后，每个用户和列表的最大推文数量 q 被设置为 20，表示每个用户和列表最近的 20 条推文。描述 s 和每条推文 L 的最大长度设置为 50 个单词，如果长度不足，则应用零填充。

在预训练阶段，利用 ChatGPT 生成的模板，将每条推文分为 16 个主题和 8 种情绪。由于仍然有一些结果不属于这些类别，我们将它们视为 “其他”。开始会随机抽取 100 个提示和其他伪标签作为负样本，并将温度 τ 设置为 0.1 以计算预训练损失。在微调阶段，λ 设置为 3×10^{-5}。预训练和微调阶段的训练周期分别设置为 100 和 150。dropout 比例设置为 0.3。使用了 AdamW 优化器 (Loshchilov and Hutter 2019) [25]，学习率为 0.001，*batchsize* 设置为 2048。

```
parser.add_argument("--template", type=str, default='l') # s, t, e, te
if args.template == "l" or args.template == "s" or args.template == "te":
    args.neg_num = 100
elif args.template == "t":
    args.neg_num = 50
elif args.template == "e":
    args.neg_num = 10
# Negative Pairs Sampling
neg_index = []
all_samples = self.none_zero_dict
for i in range(anchor.size(0)):
    neg_index.append(
    random.choices(all_samples[all_samples != label[i].item()],
    k=self.neg_num))
neg_embedding = trans_pretrain_embedding_dict[torch.tensor(neg_index)]
split_tensors = torch.split(neg_embedding, 1, dim=1)
```

为了分析用于自对比学习中不同提示模板的预训练效果，尝试了四种不同的提示模板：(1) 简短提示 s：多数主题情感对：$t^u_{i_{\max}} - e^u_{i_{\min}}$，少数主题情感对：$t^u_{i_{\min}} - e^u_{i_{\min}}$；(2) 主题提示 t：大部分帖子主题是 $t^u_{i_{\max}}$，小部分帖子主题是 $t^u_{i_{\min}}$；(3) 情感提示 e：多数帖子表达情感 $e^u_{i_{\max}}$，少数帖子表达情感 $e^u_{i_{\min}}$；(4) 串联提示 te：大部分帖子主题是 $t^u_{i_{\max}}$，小部分帖子主题是 $t^u_{i_{\min}}$，多数帖子表达情感 $e^u_{i_{\max}}$，少数帖子表达情感 $e^u_{i_{\min}}$，对于主题提示和情感提示，分别计算其对应的偏好频率，串联提示是主题提示和情感提示的融合，可以看作是基于独立频率的提示。采用简短提示符进行学习的效果与所提出的设计相比较差，这意味着与单纯地结构化地形成模板相比，考虑自然语言语义形成提示符对于偏好感知自对比学习至关重要。从主题提示和情感提示两方面来看，倾向性效应表明仅考虑任意一种情感主题信息来描述用户都不足以区分

正常用户、troll 用户和 bot 用户。此外，当应用串联提示时，性能有所下降，说明捕获主题和情感之间的成对关系比单独考虑它们更有意义。同时，所提方法也优于串联提示词，体现了联合考虑频率因素以及共同描述主题-情感对形成提示词的优势。

4.2 实验环境搭建

所有复现实验均在 Nvidia RTX A100 GPU 上进行，通过在服务器中创建 conda 环境，根据文献对应的 github 项目文件的 requirements.txt 文件选择导入相应版本的三方库，包括 pytorch，scikit-learn 等等。

4.3 界面分析与使用说明

模型具有两种运行方式，第一种是先运行项目中的 preprocess-sega.py 文件进行节点特征编码以生成相应的 pt 文件，再运行 main.py 文件进行模型训练和分类测试，命令行如下：

```
python preprocess-sega.py
python main.py --lst --pretrain
```

第二种运行方式是利用 github 项目文件中预处理好的 pt 文件，在服务器上设置好对应的路径直接运行 main.py 文件进行训练。

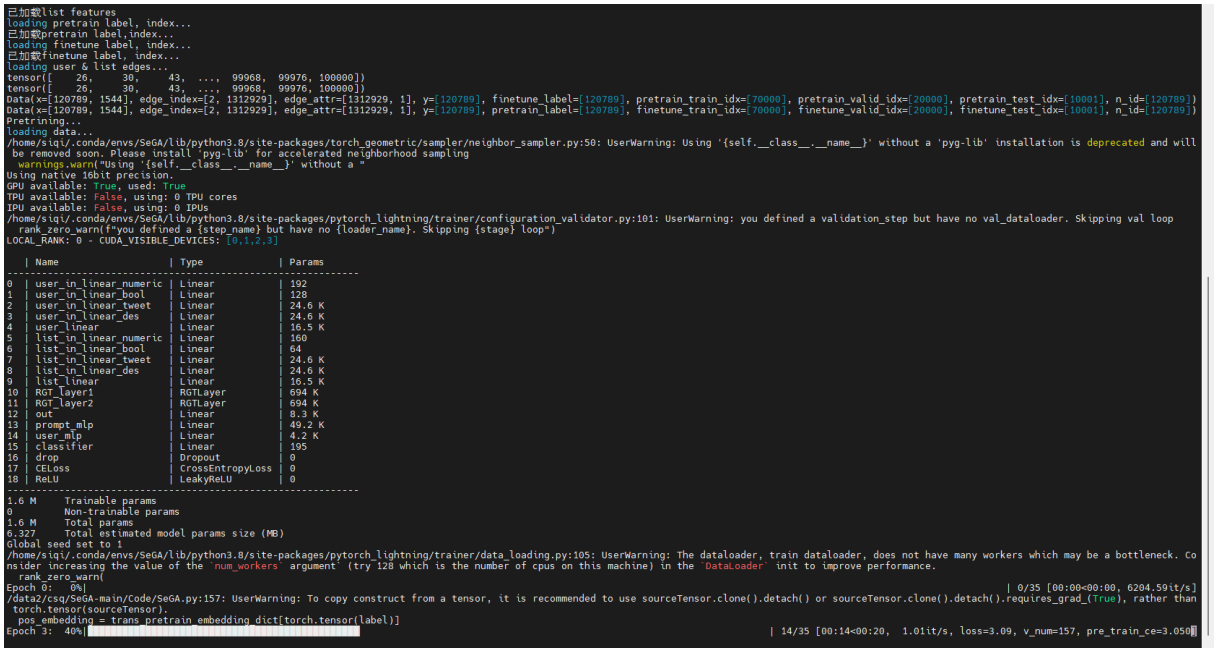

图 2. 模型训练界面示意

4.4 创新点

通过调整 Relational Graph Transformer 中的门控机制，以期望能提升网络的性能和稳定性，最终使得模型在分类准确率上有少量提升。

```
class RGTLayer(torch.nn.Module):
        self.gate = torch.nn.Sequential(
            torch.nn.Linear(in_channels + out_channels, in_channels),
```

```
            torch.nn.Sigmoid()
        )
```

5 实验结果分析

5.1 基线对比实验

由于目前还没有由喷子、机器人和正常用户组成的公开数据集，因此文献通过在机器人检测基准 Twibot-22 (Feng et al. 2022b) [26] 中添加自动喷子注释，提出了一个新的数据集 TwBNT。Twibot-22 是一个机器人检测数据集，它提供了 Twitter 网络中具有各种实体和关系的图结构，它评估了基于图的机器人检测方法。尽管如此，Twibot-22 将用户标注为机器人或人类，但排除了喷子的关键类别。为此，使用了广度优先搜索算法对之后的用户集合从 Twibot-22 中采样 (Feng et al. 2021b) [27]，以确保采样的用户包括不同类型的机器人、喷子和正常用户。然后，利用与采样用户连接的列表节点构建 TwBNT 数据集; 由于喷子由真实的人类用户控制，通过为 Twibot-22 中使用广泛认可的平台 Bot Sentinel 标记的每个用户获取喷子分数 $scr \in [0,1]$ 来自动识别它们。分数大于阈值 $scr = 0.5$ 的用户被标记为喷子，而其他用户被标记为正常用户，遵循传播意图检测 (Zhou et al. 2022) [28]。

由于所提出的异常用户检测任务缺乏区分 troll 用户的基线，文献将所提出的模型与 6 种基线的 bot 检测方法进行了比较，以验证其有效性:GCN (Kipf and Welling 2017) [8]、GAT (Velickovic et al. 2018) [9]、SimpleHGN (Lv et al. 2021) [14]、SATAR (Feng et al. 2021a) [17]、BotRGCN (Feng et al. 2021c) [10] 和 RGT (Feng et al. 2022a) [11]。

基线对比实验的复现结果如图 3 所示:

Methods	Attention Mechanism	Edge Heterogeneity	Node Heterogeneity	Self Supervision	Precision	Recall	F1
GCN					67.43±1.24	42.76±1.14	47.57±1.54
GAT	✓				66.40±2.19	45.54±0.84	50.75±0.66
SimpleHGN	✓	✓			79.84±2.54	45.68±1.19	52.22±1.54
SATAR	✓	✓		✓	55.31±7.69	44.77±2.35	46.98±1.06
BotRGCN		✓			65.85±12.55	43.51±5.20	47.94±7.20
RGT	✓	✓			75.55±1.27	52.10±1.25	58.61±0.92
SeGA	✓	✓	✓	✓	71.60	54.29	**59.99**

表 1. 基线对比实验结果示意

表 1 总结了异常用户检测方法的整体性能，表明文献提出的模型超过了所有基线，这些基线被扩展到对喷子进行分类。从数量上看，SeGA 的 F1 值比最好的基准至少提高了 3.5%。与基于同构图的方法 GCN 和 GAT 相比，SeGA 在 F1 上取得了 27.6%和 19.6%的显著提升，体现了异构图对各类异常和正常用户的分类能力。此外，SeGA 优于其他基于异构图的方法，如 SimpleHGN, BotRGCN 和 RGT，性能在 3.5%到 26.6%之间，这是因为不仅考虑了来自边的不同实体，还考虑了来自节点的不同实体。这也突出了通过带有提示的帖子利用主题-情感对作为伪标签来建模用户偏好的能力。虽然 SATAR 也将关注者数量作为自监督目标，但我们可以观察到，由于 troll 用户可以通过操纵其他 troll 用户的关注者数量来扮演正常用户的角色，SATAR 的性能大幅下降。SeGA 和 SATAR 的比较揭示了考虑用户偏好对区分异常用户的重要性，其中用户偏好不仅表现为一个与任务无关的目标，而且描述了基于相应帖子的用户轮廓。

5.2 消融实验与预训练策略对比实验

Experiment	Methods	Precision	Recall	F1
Ablation	w/o pretrain	66.34	54.44	58.62
A1	RoBERTa	64.20(-6%)	53.68(-3%)	57.73(-3.5%)
A2	Short	64.94(-1.4%)	54.94(+1.4%)	58.92(+1.1%)
	Topic	66.78(+1.1%)	53.94(-2.9%)	58.60(-1.1%)
	Emotion	64.42(-1.8%)	57.80(+4.7%)	58.32(-0.8%)
	Tandem	71.85(-1.4%)	55.53(+3.7%)	58.06(-3%)
A3	Multi-label	70.30(-2.2%)	52.89(+0.5%)	58.48(-0.4%)
Main	SeGA	71.60(+5%)	54.29(-4%)	**59.99(-1.4%)**

表 2. 消融实验与预训练策略对比实验结果示意

如表 2 所示，分别对文献的消融实验和预训练策略对比实验进行了复现（括号中的百分数表示复现结果与文献结果的误差）。消融实验中剔除了模型的预训练阶段，发现预训练过程会显著影响模型的性能。

A1: 为了进行偏好感知的自对比学习，文献将提示编码器从 SimCSE RoBERTa 更改为 RoBERTa，可以观察到用 RoBERTa 替换提示编码器性能略有降低。如图 5 所示，分别使用这些提示编码器来进一步计算提示嵌入之间的余弦相似度，以分析不同主题-情感提示之间的差异。实验结果表明，在同一个提示模板下，不同主题-情感对的表达差异更大，而不同主题-情感对的表达差异较小。此外，预训练 RoBERTa 的最小余弦相似度为 0.9795,SimCSE RoBERTa 的最小余弦相似度为 0.3083，同时使用 RoBERTa 进行的分类实验 F1 更小。这些结果表明，SimCSE RoBERTa 能够捕捉到由情绪替换引起的提示之间的细微差异，而 RoBERTa 产生了具有不同伪标签的相似表示。可区分的嵌入使自对比学习的性能得到提高，这再次提出了将其纳入基于提示的对比学习的必要性。

A2 的实验设计见 4.1 节。

A3: 为了深入研究基于帖子的主题-情感对的学习策略，文献将偏好感知的自对比学习改进为多标签分类任务 (multi-label classification task, multi-label)，旨在使用相同的模型结构预测用户所有潜在的主题-情感对。可以看到，所提出的自对比学习优于预测学习，这表明多标签分类无法捕捉与特定主题相关的用户偏好情感，因为它将所有标签视为同等重要。相比之下，偏好感知的自对比学习通过提示设计缓解了这一限制，实现异常用户检测的实质性改进。

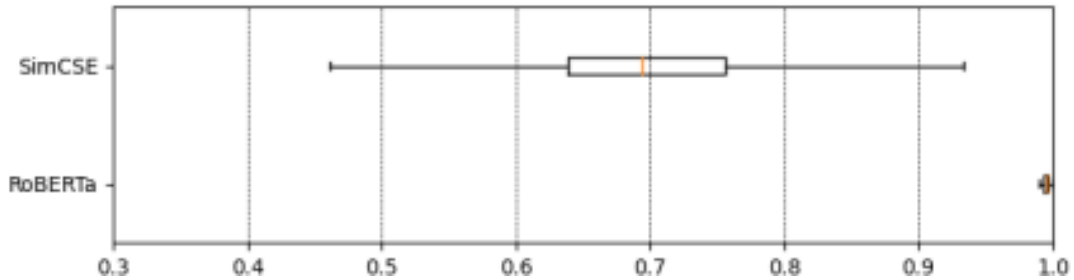

图 3. RoBERTa 和 SimCSE 的提示嵌入之间的余弦相似度的第一、第二和第三四分位数

6 总结与展望

本文介绍了 SeGA 及其复现工作，SeGA 作为一种新颖的偏好感知的伪偏好自对比学习方法，用于 Twitter 上的异常用户检测，包括更具挑战性的 troll 用户。不同于现有的仅关注机器人检测的工作，所提出的方法能够通过学习 llm 总结的用户帖子中主题-情感对的异同点来区分各种异常和正常用户，从而能够捕捉用户偏好行为。同时，提示模板的设计考虑了用户多方面偏好的上下文，避免了模型只考虑最大外观偏好的偏差。文献提出了一个用于区分 Twitter 上异常和正常用户的新基准，表明 SeGA 的性能明显优于最先进的方法，在 3.5%和 27.6%之间。SeGA 在未来也许可以作为社交媒体的通用框架，因为其设计灵活，可以从 llm 的外部知识和自对比学习方法中纳入用户偏好，并且可以在该框架中进一步探索多个有趣的方向，如用户偏好的更多元数据、少样本示例等。在复现过程中也发现文献对于利用帖子的主题和情感把握用户偏好的工作仍有创新的角度，可以考虑结合情感分析领域的其他研究成果更好地获取用户的情感偏好。

参考文献

[1] Tristan Bilot, Nour El Madhoun, Khaldoun Al Agha, and Anis Zouaoui. Graph neural networks for intrusion detection: A survey. *IEEE Access*, 11:49114–49139, 2023.

[2] Ziwei Chai, Siqi You, Yang Yang, Shiliang Pu, Jiarong Xu, Haoyang Cai, and Weihao Jiang. Can abnormality be detected by graph neural networks? In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 1945–1951. ijcai.org, 2022.

[3] Prabhat Agarwal, Manisha Srivastava, Vishwakarma Singh, and Charles Rosenberg. Modeling user behavior with interaction networks for spam detection. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2437–2442. ACM, 2022.

[4] Wei-Yao Wang and Wen-Chih Peng. Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification (short paper). In Amitava Das, Tanmay Chakraborty, Asif Ekbal, and Amit P. Sheth, editors, *Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022), Virtual Event, Vancouver, Canada, February 27, 2022*, volume 3199 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[5] Iuliia Alieva, J. D. Moffitt, and Kathleen M. Carley. How disinformation operations against russian opposition leader alexei navalny influence the international audience on twitter. *Soc. Netw. Anal. Min.*, 12(1):80, 2022.

[6] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 280–289. AAAI Press, 2017.

[7] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 218–226. ACM, 2019.

[8] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, 2018.

[9] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[10] Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM ' 21, page 236–239. ACM, November 2021.

[11] Shangbin Feng, Zhaoxuan Tan, Rui Li, and Minnan Luo. Heterogeneity-aware twitter bot detection with relational graph transformers. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3977–3985. AAAI Press, 2022.

[12] Zachary Miller, Brian Dickinson, William Deitrick, Wei Hu, and Alex Hai Wang. Twitter spammer detection using data stream clustering. *Inf. Sci.*, 260:64–73, 2014.

[13] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64, September 2016.

[14] Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking, and refining heterogeneous graph neural networks, 2021.

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

[16] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[17] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. Satar: A self-supervised approach to twitter account representation learning and its application in bot detection. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, CIKM ' 21, page 3808–3817. ACM, October 2021.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[19] Dingjun Wu, Jing Zhang, and Xinmei Huang. Chain of thought prompting elicits knowledge augmentation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6519–6534. Association for Computational Linguistics, 2023.

[20] Bilal Ghanem, Davide Buscaldi, and Paolo Rosso. Textrolls: Identifying trolls on twitter with textual and affective features. In Antonela Tommasel, Daniela Godoy, and Arkaitz Zubiaga, editors, *Proceedings of the Workshop on Online Misinformation- and Harm-Aware Recommender Systems co-located with 14th ACM Conference on Recommender Systems (RecSys 2020), Rio de Janeiro, Brazil, September 25, 2020*, volume 2758 of *CEUR Workshop Proceedings*, pages 4–22. CEUR-WS.org, 2020.

[21] Siva K. Balasubramanian, Mustafa Bilgic, Aron Culotta, Libby Hemphill, Anita Nikolich, and Matthew A. Shapiro. Leaders or followers? a temporal analysis of tweets from ira trolls, 2022.

[22] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dinggang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *CoRR*, abs/2304.01852, 2023.

[23] Jim and Ann. twitter topics –the mega list. https://inboundfound.com/twitter-topics-list/, 2020.

[24] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics, 2021.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[26] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, Xinshun Feng, Qingyue Zhang, Hongrui Wang, Yuhan Liu, Yuyang Bai, Heng Wang, Zijian Cai, Yanbo Wang, Lijing Zheng, Zihan Ma, Jundong Li, and Minnan Luo. Twibot-22: Towards graph-based twitter bot detection. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[27] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, CIKM ' 21, page 4485–4494. ACM, October 2021.

[28] Xinyi Zhou, Kai Shu, Vir V. Phoha, Huan Liu, and Reza Zafarani. "this is fake! shared it by mistake": Assessing the intent of fake news spreaders. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3685–3694. ACM, 2022.