

去除主题偏差的暗网论坛作者归属分析方法

张慧妍^{1,3}, 霍艺璇^{1,3}, 赵佳鹏^{1,2,3}, 王学宾^{4,5}, 孙岩炜^{1,3}, 赵 璨^{4,5}, 时金桥^{1,3}

¹北京邮电大学 网络空间安全学院 北京 中国 102206

²浙江大学区块链与数据安全全国重点实验室 杭州 中国 310058

³可信分布式计算与服务教育部重点实验室 北京 中国 102206

⁴信息安全内容安全国家工程研究中心 北京 中国 100093

⁵中国科学院信息工程研究所 北京 中国 100093

摘要 暗网市场由于其匿名性经常被用来交易非法商品和服务。为了更好地对频繁迁移的匿名论坛用户进行作者分析和取证,可以利用用户留下的大量帖子中的文本信息来识别和分析用户,根据论坛中的帖子对匿名用户身份进行表示学习,从而链接恶意用户的不同账号,实现匿名用户的身份追踪。然而,现有的暗网论坛作者归属分析方法学到的作者身份表示会受数据集中主题分布的影响,在训练集与测试集主题分布不一致的情况下出现明显性能下降,这意味着现有的模型学习方法在训练过程中捕获到了主题分布与标签之间的虚假相关特征,使得模型在数据分布变化的场景下无法达到预期的预测结果。本文通过结构因果模型(Structural Causal Models, SCM)解释了文本主题给作者归属分析任务引入偏差的原因,并将因果推理的方法引入基于文本的暗网论坛作者归属分析工作,提出了一种因果去偏方法来缓解主题分布不一致对作者身份表示的影响。该方法通过按主题划分数据集并保持同一作者身份表示在各主题下的潜在因果效应的不变性来减少模型对主题分布信息的依赖,学习在各个主题下性能表现更稳定的作者表示。在四个不同的暗网论坛数据集上的实验结果表明,本文的方法实现了模型预测性能的有效提升并且在不同主题数据集下表现得更稳定,在 MRR 和 Recall@10 这两个评估指标上平均提升 7.25% 和 6.33%。

关键词 作者归属; 暗网论坛; 用户分析

中图法分类号 TP391.1 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.05.02

A Topic Debiasing Method for Author Attribution Analysis of Dark Web Forums

ZHANG Huiyan^{1,3}, HUO Yixuan^{1,3}, ZHAO Jiapeng^{1,2,3}, WANG Xuebin^{4,5}, SUN Yanwei^{1,3},
ZHAO Can^{4,5}, SHI Jinqiao^{1,3}

¹School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 102206, China

²The State Key Laboratory of Blockchain and DataSecurity, Zhejiang University, Hangzhou 310058, China

³Key Laboratory of Trustworthy Distributed Computing and Service (BUCT), Ministry of Education, Beijing 102206, China

⁴National Engineering Research Center for Information Content Security, Beijing 100093, China

⁵Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Abstract Darknet marketplaces are often used to trade illegal goods and services due to their anonymity. In order to better analyze and collect authors from frequently migrated anonymous forum users, the text information in a large number of posts left by users can be used to identify and analyze users, and the anonymous users' identities can be represented and learned according to the posts in the forum, so as to link different accounts of malicious users and realize the identity tracking of anonymous users. However, the authorship representation learned by the existing dark web forum author attribution analysis method will be affected by the topic distribution in the dataset, and the performance will be significantly degraded when the topic distribution of the training set is inconsistent with the test dataset, which means that the existing model learning methods capture the spurious correlation features between the topic distribution and the labels in the training process, which makes the model unable to achieve the expected prediction results in the scenario of data distribution change. This paper uses the Structural Causal Models (SCM) to explain the reasons why the topic of texts introduces bias to the author attribution analysis task, and introduces the method of causal reasoning into the text-based author attribution analysis of dark web forums, and proposes a causal debiasing method to alleviate the influence of topic distribution inconsistency on the authorship representation. By dividing the dataset by topic and maintaining the invariance of the potential causal effects under each topic represented by the same authorship, the model reduces the dependence of the model on topic distribution information, and learns the author representation with more stable performance under each topic. Expe-

通讯作者: 赵佳鹏, 博士, 博士后, Email: zhaojp9574@bupt.edu.cn。

本课题得到浙江大学区块链与数据安全国家重点实验室开放研究基金、青年科学基金项目(No. 6240071954)资助。

收稿日期: 2024-11-16; 修改日期: 2025-02-28; 定稿日期: 2026-04-15

perimental results on four different dark web forum datasets show that the proposed method achieves an effective improvement in the prediction performance of the model and performs more stably under different topic datasets, with an average improvement of 7.25% and 6.33% on the two evaluation indicators of MRR and Recall@10.

Key words authorship attribution; darkweb forum; user analysis

1 引言

暗网需要使用特殊的软件和通信协议(如 Tor2)才能访问,这提供了匿名性和不可溯源性。而由于其匿名性,暗网市场论坛经常被用于进行非法交易,对网络空间安全造成了极大的威胁。与传统社交媒体中的用户不同,暗网市场论坛的用户经常从一个市场转移到另一个市场,难以溯源,只留下大量帖子作为他们的数字足迹^[1]。在此情况下,根据论坛中的帖子对匿名用户身份进行表示学习,从而链接恶意用户的不同账号,实现对匿名用户的身份追踪,具有极大的研究意义。

为了实现对匿名用户的追踪分析,一种研究思路是通过分析未知文本的文体学特征或写作特征等来推断作者归属。作者归属任务是作者分析的一个细分领域,目标是从一组可能的作者中找到文本真正的作者。传统的作者归属方法^[2-3]主要通过提取作者的写作特征,例如单词频率、大写字母、标点符号样式、功能词等,将其输入到相应的统计或机器学习框架实现对作者的身份归属分类。然而该技术更适用于长文本场景下的作者归属任务(如书籍、小说),在短文本场景下,由于文本的长度相对较短,提取可用特征的难度较大,模型性能表现不佳。因此传统的方法不适用于暗网论坛高度匿名的短文本场景。最近基于深度神经网络^[4-5]或预训练语言模型(如BERT)^[6]的研究在较短的社交媒体文本数据集(如Blogs50等)上有了良好的表现。而在暗网市场数据集上,Pranav等人^[7]、Liu等人^[8]受到Andrews等人^[9]的启发,通过组合相近时间段用户发布的帖子文本作为作者文本片段,并以度量学习确保同一作者文本片段具有相似的嵌入来学习作者身份表示用于作者归属任务。这些方法在一定程度上解决了用户文本数据集中文本数量不平衡的问题,在暗网市场数据集上取得了一定成果。

尽管当前的方法已取得一定成效,但仍存在一些不足之处,主要在于未充分考虑现实场景中主题分布不一致的情况。在现实世界的跨平台场景中,用户往往在各个平台上表现出不同的主题偏好,比如X(原推特)和Linkedin两个平台,前者侧重于社交,后者侧重于求职,用户在不同平台上的发帖存在主题差异。且用户偏好的主题会随着时间推移有

所变化,在不同时间节点下,用户关注的主题可能会有所不同,而现有的作者身份归属方法在主题分布不一致的情况下表现不佳,无法应对主题变化的场景。图1,图2举例了暗网论坛数据集BMR中测试集内预测分数较低和较高的用户在训练集和测试集的主题分布情况,可以看出预测分数较低的用户在训练集和测试集上存在显著的主题分布不一致,即由于主题分布偏差的存在,模型错误地估计了数据集中主题特征与预测目标的相关性,导致在主题分布与训练集不一致的测试集上,模型无法做出正确的预测。因此需要针对该问题开展研究,使模型不受限于训练集中的主题分布,改进现有模型应对复杂的主题变化的现实场景的能力。

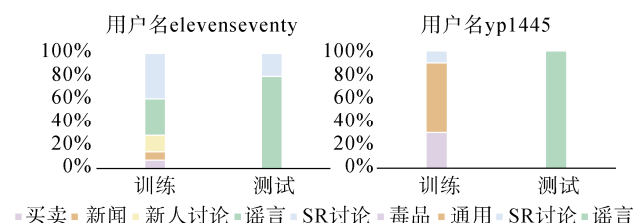


图1 BMR数据集中测试集下预测准确率较低的两名用户训练集与测试集分布情况

Figure 1 The Distribution of the Training Dataset and the Test Dataset of two Users With Low Prediction Accuracy in the Test Dataset in the BMR Dataset

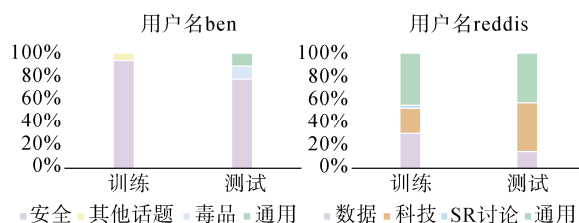


图2 BMR数据集中测试集下预测准确率较高的两名用户训练集与测试集分布情况

Figure 2 The Distribution of the Training Dataset and the Test Dataset of two Users With High Prediction Accuracy in the Test Dataset in the BMR Dataset

本文对已有文献进行综合分析的基础上,借鉴因果推断理论,从因果图的角度刻画作者归属任务,发现并缓解任务中存在的主题分布与模型预测目标之间的虚假相关性,使模型学习到更稳定的作者身份表示,改进在主题分布不一致场景下模型的预测效果。

基于此, 本文主要的研究工作如下:

1) 本文通过对暗网论坛文本作者归属任务的方法和数据集进行统计分析, 提出现有的作者身份表示方法会受到数据集主题分布的影响, 并通过因果框架形式化作者归属任务过程, 解释了作者归属任务中文本主题分布引入表示偏差的原因。

2) 本文首次将因果推理的方法引入基于文本的作者归属分析工作, 通过因果去偏方法保持同一作者身份表示在各主题下的潜在因果效应的不变性, 以此缓解主题分布不一致对作者身份表示的影响, 促使模型学习到更稳定的作者身份表示, 实现更有效的作者归属分析。

3) 在四个公开的暗网市场数据集上的实验结果表明此方法可以有效地提升模型的预测性能和模型性能表现的稳定性。

本文章节的组织结构如下: 第 1 节介绍了暗网论坛作者归属分析工作的研究背景; 第 2 节介绍了相关的研究工作; 第 3 节介绍了因果推理的先验知识; 第 4 节对当前暗网论坛作者归属任务中存在的主题偏差现象进行了分析; 第 5 节介绍了本文设计的暗网论坛文本的作者归属分析方法; 第 6 节介绍了实验中使用的数据集、评价指标以及基线; 第 7 节通过实验分析验证本文提出的方法的有效性; 第 8 节对全文进行总结, 并对未来的研究方向进行了展望。

2 相关工作

2.1 作者归属

作者归属任务旨在将未知文本归因于 k 个候选作者集中的作者 A 。近年来, 随着预训练语言模型在自然语言理解和自然语言推理任务上的成功, 越来越多的工作利用预训练语言模型为作者身份识别任务生成表示, 如 Fabien 等人^[7]证明了 BERT 在这一领域的巨大潜力。此外, Rivera-Soto 等人^[10]依靠 Siamese BERT 尝试学习作者身份不变的通用表示, 并把他们的研究拓展到零样本可迁移的学习。Ai 等人^[11]在预训练语言模型的基础上引入了对比学习的方法, 在三个短文本作者归属数据集 (IMDb62, Blogs10, Blogs50) 上取得了良好的表现。Tyo 等人^[12]的实验表明基于预训练语言模型训练的表示在短文本作者归属数据集 (IMDb62, Blogs50) 上具有更好的结果。

此外, 研究人员还研究了作者归属任务的同主题、跨主题^[13]和主题混淆^[14]设置的场景, 以确定该模型是否捕获了作者写作风格或主题的转移。对于跨主题的场景, 主要考虑训练集与测试集主题不同的

场景, 目前主要使用多任务学习、数据增强的方法与传统的分类方法相结合。但是这些研究更聚焦于学习作者的写作风格表示特征以应对主题变化的场景。Jitkatpat 等人^[15]尝试使用蒸馏框架减少对特定于主题信息的依赖。但实际上, 作者的特征不只包含风格特征, 还包括一些内容层面的特征, 如词汇特征、句法特征等, 如何让模型更充分地学习各主题下作者特征的一致性, 还需要进一步讨论。

2.2 暗网论坛用户分析

暗网论坛中包括各种非法资源交易、数据泄露、毒品交易、政治以及其他相关内容, 对网络空间安全造成了很大的威胁。而对暗网市场论坛的调查和分析表明, 当一个市场关闭时, 用户通常会迅速迁移到新市场, 只留下大量的文本数据。这使得利用暗网论坛中存在的文本数据进行用户分析成为可能。最近的工作尝试利用论坛中的文本或图像数据来进行暗网作者分析工作。一些研究使用各种技术和多模态数据来进行暗网作者分析。Wang 等人^[16]和 Jeziorowski 等人^[17]的作者使用产品的照片来链接供应商的多个账户。Wang 等人^[16]主要使用 Agora、Evolution、Silk Road 2.0 (SR2) 三个论坛的图像数据进行实验。类似的, Zhang 等人^[18]提出了一个模型, 利用毒品的图像和描述来创建作者嵌入并在 Valhalla、Dream Market、Silk Road 2.0 (SR2)、Evolution 四个数据集上取得了一定成果, 但这些方法依赖于较大规模的多模态信息源, 而这种大规模多模态信息的获取十分困难。

另外一些研究探索了利用用户帖子中的文本信息来获取作者身份表示的方法。Arabnezhad 等人^[19]展现了利用文本中的文体特征链接暗网论坛中的用户的可能性, 他们在 The Majestic Garden 和 the Dream Market forum 两个暗网论坛上进行了实验。Pranav 等人^[7]对 Munksgaard 等人^[20]研究中的四个暗网论坛数据集 Silk Road (SR)、Silk Road 2.0 (SR2)、Agora Marketplace (Agora) 和 Black Market Reloaded (BMR) 进行了文本清洗和预处理, 构建了用于进行暗网论坛作者归属研究的数据集。另外, 他们使用卷积神经网络 (CNN) 来获取作者文本信息的嵌入, 并通过异构神经网络 (HIN) 建模上下文信息, 通过文本信息、上下文信息和时间信息来学习作者身份表示。在此基础上, Liu 等人^[8]使用相同的数据集进行实验, 将卷积特征与通过自适应门机制的自注意力特征结合, 使其更适合建模同一用户不同长度的帖子。本文遵循他们研究中的实验设置继续在这四个数据集上进行分析 and 实验。

尽管当前的研究基于深度学习模型已经取得了一定的效果,但现有的模型性能受限于训练集中的数据分布,当测试数据中主题分布与训练数据中主题分布不一致时模型性能会受到明显的影响(见第4节)。而实际上,在暗网场景中,随着时间的推移,用户的主题分布是存在显著变化的,因此在暗网场景下需要设计合理的方法来缓解主题分布不一致对于作者身份表示的影响,以更好地识别作者的身份。

2.3 因果推理相关研究

因果推理^[21-22]是一种统计建模工具,用于对可能导致虚假相关性和混淆的偏差进行解释性分析,因果推理通过图模型推断变量之间的因果关系,并引入了干预分析的概念,其中 **do** 算子是其中的重要工具,通过 **do** 算子可以定义因果效应。目前因果推理已经广泛应用于各种领域,如计算机视觉^[23-24]和自然语言处理^[25]。

因果不变性是因果推理中的一个重要概念,这个想法被用于各种学习域偏差^[26]的分类任务。Wu 等人^[27]认为,通过最小化平均损失和不同环境中数据损失的方差,可以帮助模型从输入特征中学习独立于环境的特征,实现分布外泛化。Mitrovic 等人^[28]通过不变因果机制的表示学习使模型性能在 ImageNet 的鲁棒性和分布外泛化上优于其他方法,Zhang 等人^[25]将因果不变机制用于自然语言处理中的远程监督命名实体识别任务,利用因果不变性正则化项来解决远程监督命名实体识别任务中字典引入的偏差。

Tyo 等人^[12]对现有的作者归属分析工作进行了总结分析,当前相关的研究工作还没有把因果推理用于文本的作者归属分析。虽然有一些工作将因果推理引入自然语言处理任务,但这些工作主要是考虑各种环境下的降噪。本文首先提出了利用因果推理的方法分析主题给作者归属任务带来的因果效应,引入因果不变性学习降低模型对主题分布虚假线索的依赖,并利用实验验证所提方法的有效性。

3 因果推理先验知识

3.1 因果图

因果图(Cause Graph)是一种可视化因果关系的图形工具,一般使用有向无环图(Directed Acyclic Graph, DAG)来描述变量之间的因果关系。在因果图 $G = \{N, E\}$ 中,节点 N 代表变量集,边 E 代表两个变量之间的因果关系。

图3展示了一些因果图的示例,其中 X 、 M 、 Y 分别表示因果变量、中介变量和结果变量; * 表示对

变量进行干预;实线箭头表示变量间存在的因果关联;虚线箭头表示变量间的因果关联被切断。如图3(a)所示, $X \rightarrow Y$ 表示变量 X 和变量 Y 之间存在直接的因果关联,而 $X \rightarrow M \rightarrow Y$ 表示变量 X 与变量 Y 以 M 为中介节点存在间接的因果关联,这类路径被称为后门路径,此时中介变量 M 的存在为计算变量 X 和变量 Y 的因果效应引入了混杂因子。图3(b)表示进行干预后的场景,通过干预切断了变量 X 与 M 之间的联系,这样就能估计出 $X \rightarrow Y$ 的直接影响。此处通常用 **do** 算子来表示干预操作。

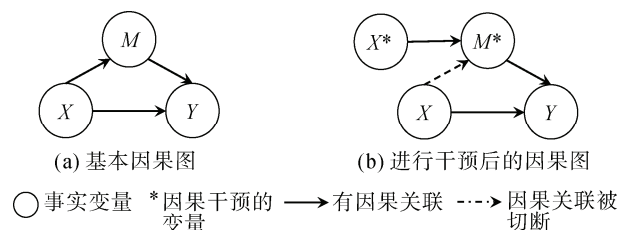


图3 因果图示例

Figure 3 Examples of Causal Graph

3.2 因果相关性

因果图中,节点表示变量,箭头表示变量之间的因果关系,如 $X \rightarrow Y$ 可以来表现节点 X 和节点 Y 之间的因果关系,在这个式子中节点 X 是因变量,节点 Y 是结果变量, $\rightarrow$ 代表 Y 与 X 存在因果相关性。因果相关性,是指在其他变量不变的情况下,两个变量所呈现出的相关性,即两个变量之间存在直接的因果关联。

3.3 虚假相关性

虚假相关性则是指观测数据中所呈现出的两个变量的相关性,它不同于两个变量之间的因果相关性。在有混杂因子存在的情况下,两个变量之间的相关性是被高估或者低估的,这部分相关性会随着数据分布偏移而变化,呈现出不稳定性。值得注意的是,虚假相关性并不代表两个变量之间不存在任何因果影响,而是说这两个变量在观测数据中最终呈现出的相关性与其直接的因果相关程度存在差异。

例如,研究发现经常服用维生素的人健康状况会改善,但实际上,经常服用维生素的人相比于不经常服用的人有更多的健康习惯,这些可能才是健康状况改善的真正原因。在这种情况下,维生素和健康状况改善之间存在虚假的相关性,维生素并不是健康状况改善的原因。

3.4 因果去偏

因果推断方法将造成偏差的原因视为混淆因子,通过考虑场景下的多种因素,分析其中的混淆因子

去除对应的因果效应, 实现去偏来提升模型的预测效果。因果推断理论将实际生活中常见的场景抽象成因果问题, 以因果图的形式明确各因素之间的因果关系, 在数据中找出引起模型预测效果下降的原因, 将其视为混淆因子, 认为该混淆因子是偏差产生的原因。这种混淆因子是现实应用中客观存在的问题, 且部分混淆因子无法直接观测到, 而模型训练时会此类因素考虑进去, 最终导致预测过程中产生偏差。因果推断通过使用干预或者反事实假设来去除混淆因子的因果效应, 从而去除偏差并提升模型的预测效果。

针对预测中产生的偏差, 主要有两种解决思路。第一种思路是在训练阶段对模型进行调整, 针对已知的虚假线索设计惩罚损失或对抗学习目标(文献[29-30]), 使模型在学习如何预测原始目标的同时, 降低对虚假线索信息的捕捉; 第二种思路在训练前阶段对样本进行调整, 主要是通过样本加权、样本补充、样本过滤等调整方法(文献[31]), 减少训练样本对虚假线索的依赖。

4 暗网论坛作者身份归属任务中的主题偏差现象分析

4.1 主题偏差现象

主题偏差指的是从数据角度而言, 主题的分布使在数据集上训练的模型对此特征产生不合理的依赖, 由此导致在主题分布不一致的测试集下模型表现效果不佳。为了衡量暗网论坛归属任务中的训练集与测试集的主题偏差, 本文采用主题分布相似度进行计算。主题分布相似度计算了用户训练集与测试集主题相同的文本数量与训练集与测试集的文本总数的比值, 公式如下:

$$s_a = \frac{\sum_{T=1}^n N_{\text{train}} \cap N_{\text{test}}}{\sum_{T=1}^n N_{\text{train}} \cup N_{\text{test}}} \quad (4-1)$$

其中, s_a 代表用户 a 的主题分布相似度, N_{train} , N_{test} 分别代表训练集和测试集中用户指定主题的文本数量, n 代表主题的个数。主题分布相似度越低则意味着同一作者训练集和测试集主题的重合度越低, 两者的分布差异越大。本文使用的四个暗网论坛数据集的主题分布相似度与各主题分布相似度区间内用户的平均预测分数关系如图 4 所示。此处平均预测分数使用 MRR 进行衡量, MRR 用于计算第一个检索到的文本为同一作者索引的倒数排名, MRR 值越高, 同一身份的作者表示越接近, 预测的效果越好。

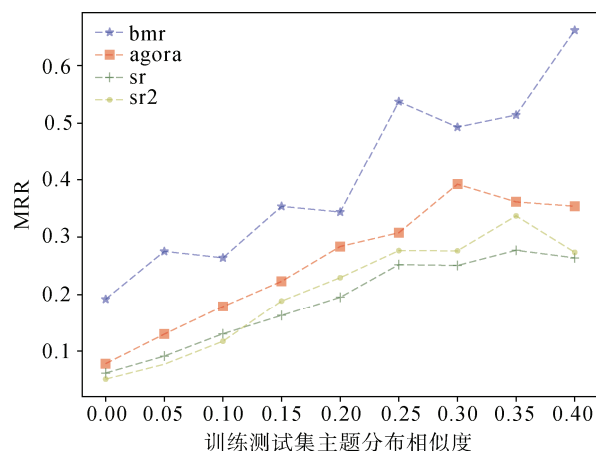


图 4 在四个暗网论坛数据集下主题分布相似度与区间内用户平均预测分数之间的关系

Figure 4 The Relationship Between the Similarity of Topic Distributions and the Average Prediction Scores of Users in the Interval Under the Four Darkweb Forum Datasets

从图中可以清晰地显示出在这四个暗网论坛数据集下的用户平均预测分数都随着主题分布相似度的提升而提升, 即用户文本的主题分布在训练集和测试集较为相近时预测分数较高, 而当训练集与测试集主题分布相似度较低时, 即两者的主题分布不一致时, 模型的性能会出现明显的下降。由此图可以看出当前模型学习的作者身份表示普遍在用户主题分布相似度低时表现不佳。

另外, 为了验证主题偏差是否是制约模型性能的主要因素, 此处考虑了可能影响模型性能的语言偏差与主题偏差因素与模型性能的关系, 如图 5 所示。

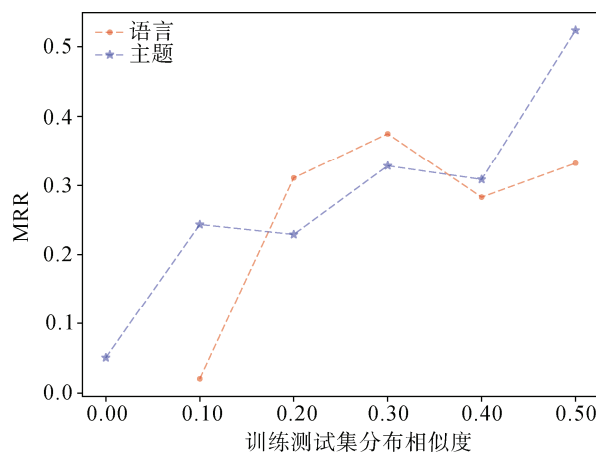


图 5 在 BMR 数据集下主题和语言分布相似度与区间内用户平均预测分数之间的关系

Figure 5 The Relationship Between the Similarity of Topic and Language Distributions and the Average Prediction Scores of Users in the Interval Under the BMR Dataset

从图中可以看出在当前的数据集下主题分布相似度与模型性能表现是呈正相关的, 而语言分布相似度较高时模型性能表现相对降低, 因此本文认为主题偏差是制约模型在暗网论坛文本归属任务表现的一个关键因素。现有模型表示方法会依赖作者文本主题分布的虚假线索学习作者的身份信息, 导致学习到的作者身份表示模型存在一定的局限性, 受限与训练集中的数据分布, 无法充分应用于作者主题变化的现实场景, 也难以推广到分布外的主题上。

4.2 因果图视角的主题偏差分析

为了探究主题偏差现象产生的原因, 本文分析了作者身份归属任务中的因果关系, 确定了各因素之间的因果关系, 以及其中造成偏差的因素, 并抽

象出一个因果图, 此处由于本文主要探究的是主题对于作者归属分析任务的影响, 因此将其他可能影响作者归属分析的因素都视为外部因素, 主要探讨主题、文本、作者特征这几个因素之间的因果关系。

4.2.1 作者身份表示的结构因果模型

图 6 显示了作者身份表示的结构因果模型, 过程中包含 5 个关键变量, A 代表作者的特征; T 代表主题; X 代表文本; R 代表模型编码出的文本表示; Y 代表下游任务。变量之间存在如下关系:

作者特征 A 会对文本 X 产生影响。对于文本, 作者经常会融入自己独特的写作风格, 比如有些作者习惯于在末尾添加特殊符号('):'), 不同作者在描述同一对象时可能会有不同的情感偏好特征, 也可能会有不同的内容特征, 如用词习惯(USA 和 US)。

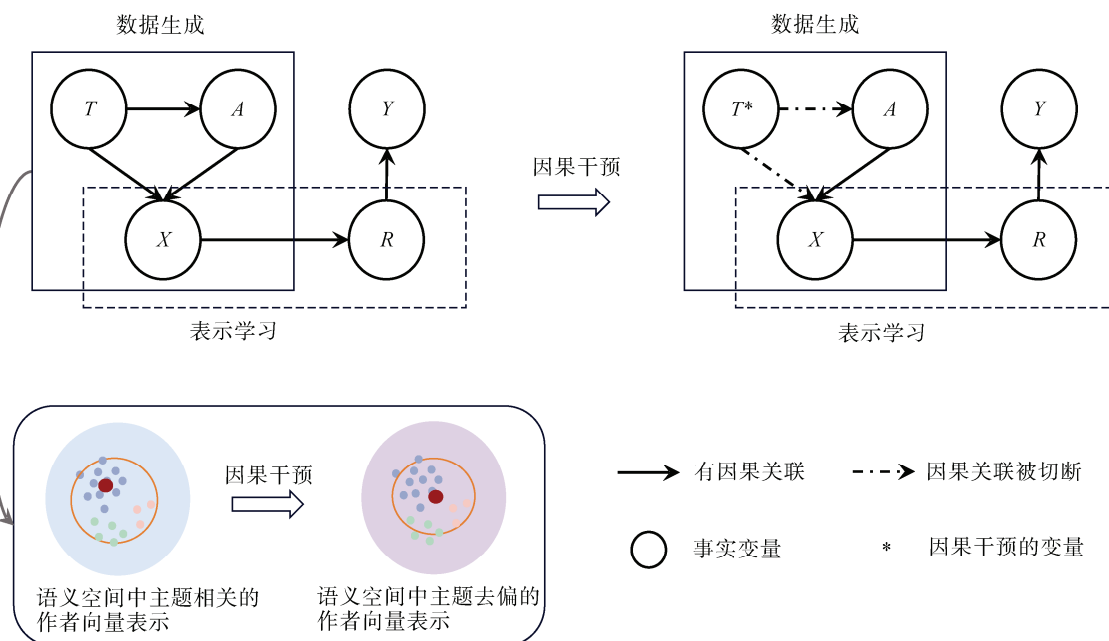


图 6 作者身份表示的结构因果模型

Figure 6 Structural Causal Model of Authorship Representation

主题 T 会对作者特征 A 产生影响, 由于每个作者都有自己的发帖主题偏好, 这种主题偏好会影响作者的写作风格和写作内容范围, 因此作者特征会受到文本主题分布的影响。

主题 T 也会对文本 X 产生影响。例如, 对于工作相关的主题, 作者的文本写作风格会更正式, 而对于游戏娱乐相关的主题, 其写作风格会更轻松。且主题会限制内容词汇表达的范围, 谈论同一个主题时作者往往会在有限的词汇中进行选择。

$X \rightarrow R$ 代表使用模型对文本 X 编码的过程, $R \rightarrow Y$ 代表下游任务, 在本文的设置中可视为作者归属任务。

4.2.2 作者身份表示主题偏差产生的原因

由于目标的下游任务是依靠文本识别相应的作者, 需要使用模型来学习作者的表示 $A \rightarrow X \rightarrow R$, 按照传统的方法, 作者归属任务通常被转化为文本分类任务, 即把待识别的文本输入模型, 将模型捕获文本与作者标签之间的推理关系并进行预测。

但是, 由于训练集中作者各主题数据分布的偏差, 直接训练的作者身份表示模型学习到的作者身份特征会受到主题分布的影响, 拟合主题分布与作者身份标签之间的虚假相关特征。另外, 作者的写作主题通常与时间、平台等因素有密切的关系, 拟合主题与表示之间的相关性会使模型在分布外的主题上

的作者身份识别表现不佳。为了促使模型能够更好地学习到更稳定的作者身份表示, 需要缓解主题引入的虚假相关性。从因果的角度分析, 这种分布不平衡会造成混淆效应并形成后门路径 $A \leftarrow T \rightarrow X$ 。这种情况下, 可以使用 do 算子和后门调整来进行干预, 去除混杂效应, 得到真正的因果效应。

5 方法

5.1 任务定义和关键思想

暗网市场论坛遵守作者归属任务中的开放世界假设, 目标作者 A 可能包含在候选的作者集合中。这项任务的核心挑战在于学习一个作者与其他作者不同的独特作者特征表示。给定数据集 D 中一个作者 a 的 N 条文本的数据 $(t_1, \dots, t_N)$, 应该最大化每条文

本数 t_i 到作者标签 a 的预测准确性, 最小化 t_i 到同一作者文本集合 $(t_1, \dots, t_N)$ 中其他文本的距离并最大化 t_i 到 D 中其他作者文本的距离。

为了解决现有方法存在的问题, 本文构建主题数据集并通过因果不变性正则化项来去除模型中存在的主题偏差, 学习更稳健的作者身份表示。本文模型基础架构图如图 7 所示。在训练阶段, 本文通过构建主题数据集将文本集按主题进行划分来进行主题特征的增强, 促使模型更充分地捕获主题片段中单个主题下作者的内容相关特征。接着将同个作者的不同主题片段表示输入因果去偏模块以学习不同主题片段下的作者的不变性特征, 使用因果不变性正则化项与交叉熵损失结合迭代更新模型进行优化完成对主题的去偏, 促使模型学习更稳定的作者身份表示。

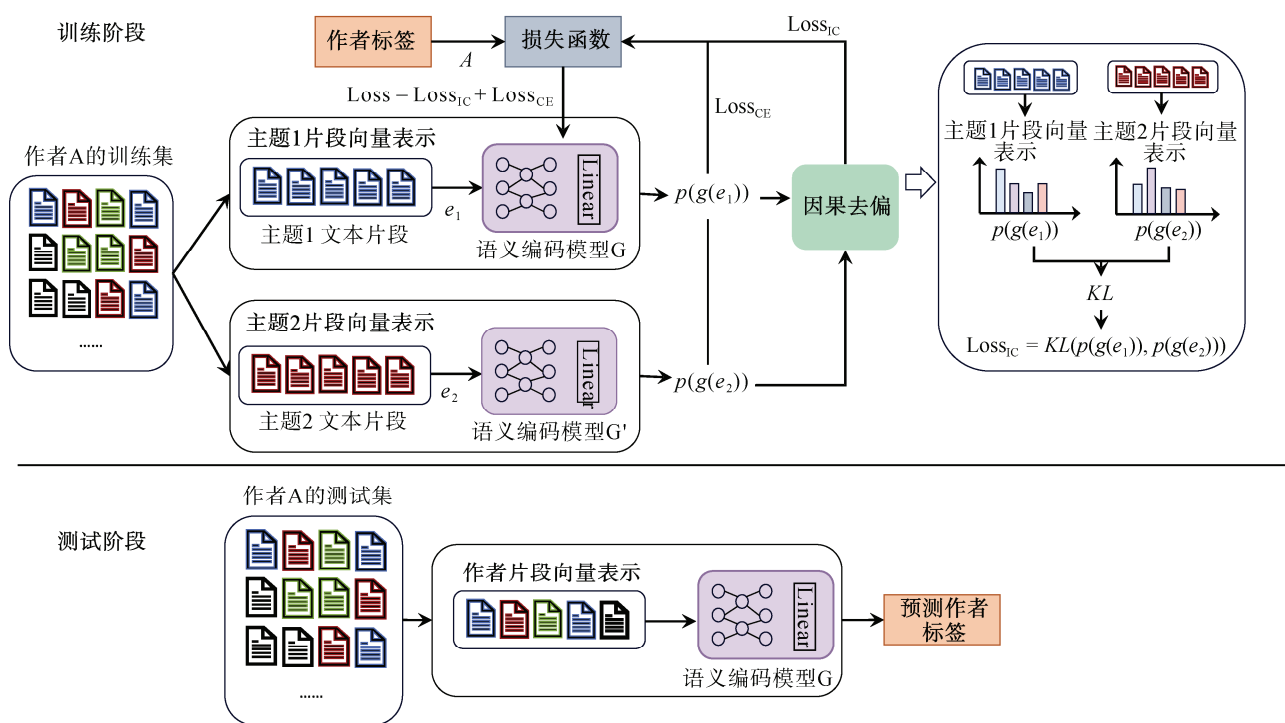


图 7 本文模型基础架构图

Figure 7 Model Structure of our Method

5.2 数据预处理

前人研究^[7-9]以片段数据集预处理方法对数据集进行处理。其基本思想是时间段连续的帖子上下文中可能包含作者身份表示的独特信息, 因此他们按时间顺序对帖子进行排序, 再按照长度 L 将排序好的数据集划分成作者片段, 确保每个片段中包含 L 条帖子文本, 即作者的每个片段集中包含特定时间段内的作者文本信息, 但这种划分方式只考虑时间连续性, 一个片段中可能包含多个主题的作者文本, 学习各片段中的作者特征时无法充分捕捉到各主题

下作者的独特内容特征。片段数据集预处理算法如下所示。

算法 1. 片段数据集预处理.

输入: 暗网论坛数据集, 数据集中包含帖子文本信息、作者信息、时间以及主题信息等。

- 1) 将输入的所有论坛文本按时间顺序排序;
- 2) 以作者名称 `auth` 为单位将帖子划分成作者-帖子集合 `auths`;
- 3) FOR `auth:authors`

4) 获取作者 **auth** 下的所有帖子文本;

5) 以 L 为单位将文本划分成片段, 即每 L 条帖子文本划分为一个片段, 放入作者-片段数据集 **episode** 中;

6) END FOR

输出: 作者-片段数据集 **episode**

由于每个主题的语义空间是有限的, 而每个作者在各主题发布的文本会有自己的用词偏好特征。按照片段数据集构造方式更有利于捕获不同主题下作者的风格特征, 难以充分捕获单个主题下作者文本独特的内容特征。因此本文提出一种按照主题划分数据集的方法, 以促使模型更好地捕获单个主题下作者的内容相关特征, 实现单个主题特征的增强, 并有助于后续通过因果去偏模块更好地学习不同主题下作者的不变特征。

主题数据集预处理方法为先将帖子按时间排序, 保证帖子在时间上的连续性, 再将排序好的帖子按主题进行分类, 最后按照长度 L 对数据集进行划分, 确保每一个主题片段数据中包含的是相近时间段下特定主题的作者文本信息, 将原始数据集转换成主题数据集, 主题数据集预处理算法如下所示。

算法 2. 主题数据集预处理.

输入: 暗网论坛数据集, 数据集中包含帖子文本信息、作者信息、时间以及主题信息等。

1) 将输入的所有论坛文本按时间顺序排序;

2) 以作者名称 **auth** 为单位将帖子划分成作者-帖子集合 **auths**;

3) FOR **auth:authors**

4) 获取作者 **auth** 下的所有帖子文本;

5) 将帖子按主题进行排序;

6) 以主题为单位获取帖子文本, 同一主题中再以 L 为单位将文本划分成片段, 即同一主题中文本帖子数量大于 L 的情况下, 每 L 条帖子文本划分为一个片段, 放入作者-主题数据集 **topic episode** 中;

7) END FOR

输出: 作者-主题数据集 **topic episode**

5.3 因果不变性正则化项

理想情况下, 一个鲁棒性更强的作者身份表示应该是与主题在数据集中的分布无关的。因为作者感兴趣的主题是会受到时间、平台等因素影响的, 即不同时间段内作者感兴趣的主题可能会发生变化, 依赖主题分布信息学习到的作者身份表示是存在一定局限性的。例如, 对于同一个作者而言, 在刚进入论坛时可能会在新人板块发帖更多, 但在一段时间

后会倾向于在其他自己感兴趣的主题下发帖。而模型学习的过程中可能仅包含前一个时间段的主题文本, 而导致学习到的特征存在一定主题偏差。即主题是作者特征和预测结果之间的混淆因子, 需要通过后门调整的方式来实现干预, 这里通过控制 T 的取值来实现对 A 的干预阻断后门路径 $A \leftarrow T \rightarrow X$ 。

当因果变量 T 取不同的值时, 由于 T 会同时作用于 A 和 X , 即 $T = t_i$, 此时变量 R 可以表示为 $R_{t,a} = R(A = a_i, T = t_i)$, 通过干预切断了 A 和 T 之间的联系, 这样可以估计出 T 对 R 的直接影响。

为了评估影响的具体大小和程度, 通常使用因果效应(Causal Effect)来计算。因果效应的计算通常是通过构建反事实世界来实现的, 即改变因变量的取值, 以推断如果因变量发生变化会有什么不同的结果, 通过比较真实世界和反事实世界的结果以计算变量之间的因果效应。在本文中 T 通过直接路径和间接路径对 Y 产生的总因果效应(Total Effect, TE)为

$$TE = Y_{a,t} - Y_{a_i,t_i} \quad (1)$$

总因果效应可以分解为 $T \rightarrow X$ 的直接因果效应和 $T \rightarrow A \rightarrow X$ 的间接因果效应, 直接因果效应(Natural Direct Effect, NDE)可以通过以下公式计算:

$$NDE = Y_{a_i,t} - Y_{a_i,t_i} \quad (2)$$

通过从总因果效应中减去直接因果效应就可以得到间接因果效应(Total Indirect Effect, TIE):

$$TIE = TE - NDE = Y_{a,t} - Y_{a_i,t} \quad (3)$$

通常来讲, 直接计算间接因果效应较难以实现, 所以通常通过这种方式来计算所需要的间接因果效应。一般来讲, 会通过 **do** 算子来进行计算, 公式如下:

$$\begin{aligned} p^{\text{do}(T=t)}(Y | f(x)) &= \sum_{t \in T} P(t)P(Y | a, x, t) \\ &= E_t P(Y | a, x, t) \end{aligned} \quad (4)$$

可以通过最小化负对数似然函数将 $P(Y | a, x, t)$ 映射到函数 $f(a, x, t)$, 所用的最小化负对数似然函数公式为

$$\min \sum -y \log(f(a, x, t)) - (1 - y) \log(1 - f(a, x, t)) \quad (5)$$

为了提高学习过程中的稳健性, 本文通过使用因果不变性正则化项减少对主题分布特征的拟合。具体而言, 以因果不变性为学习目标, 每次采样一个作者的两个不同主题片段, 保持作者表示在采样的不同主题下的不变因果效应, 公式可以表示为

$$p^{\text{do}(T=t_i)}(Y | f(x)) = p^{\text{do}(T=t_j)}(Y | f(x)) \quad (6)$$

此处 $t_i, t_j \in T$, $\text{do}(T = t_i)$ 代表着对此数据集的

主题进行干预。这代表给定不同的主题数据集, 保持从 X 到 R 的因果效应不变, 由此使 $X \rightarrow Y$ 的因果效应保持不变。因果不变性正则化项通过最小化同一作者不同主题间的表示差异来增强因果一致性, 公式表示如下:

$$L_{CI}(\theta; T; A) = \sum_{i=1}^n \sum_{x \in X} \|R_{T,A}(x; \theta) - R_{T_i,A}(x; \theta)\| \quad (7)$$

这里 $R_T(x; \theta)$ 是 T 主题下的实例 x 编码得到的表示, n 代表主题数据集的数量, $\|\cdot\|$ 代表两个分布之间的距离。此处具体使用 KL 散度来度量两个分布之间的差异程度, KL 散度越大, 则两个概率分布差异越大, KL 散度越小则两个概率分布差异越小, 计算公式如下:

$$KL(P \| Q) = \sum p(x) \log \frac{p(x)}{q(x)} \quad (8)$$

最后, 通过将交叉熵损失和因果不变损失相结合来优化模型的参数, 获得最终的模型。

$$L_{CE} = -\sum_{i=1}^n a_i \log(p(x)_{a_i}) \quad (9)$$

$$L = L_{CE} + \lambda L_{CI} \quad (10)$$

其中, λ 为需要在验证集中调整的超参数。

6 实验

6.1 数据集

暗网市场论坛是使用数字加密来隐藏其身份的各方之间进行商品和服务交换的在线论坛。本文在四个暗网市场论坛数据集上进行实验。Pranav 等人^[7]对数据集进行了基本的预处理。数据集中的信息不包括参考信息、链接、图像等。四个暗网论坛有: BMR(Black Market Reloaded)、Agora(Agora Marketplace)、SR(Silk Road)、SR2(Silk Road 2.0)。测试数据包含训练期间未见到的作者。所有数据集的统计信息如表 1 所示。这里训练数据集和测试数据集大小基本相等, 按时间顺序划分。从表中可以看出 BMR、Agora 数据集中用于训练测试的样本量与用户量相对而言较少, SR、SR2 数据集中用于训练测试的样本量与用户量相对较多。从主题上多样性角度看, BMR 数据集中主题多样性较高, 主题种类更丰富。对于单个用户而言, BMR、SR、SR2 这三个论坛数据集中平均每个用户的主题种类更为多样, Agora 数据集中每个用户发帖的平均主题种类相对较少。

四个暗网市场论坛数据集的主题分布相似度箱线图如图 8 所示。箱线图可以反映数据分布的中心位置和散布范围, 箱线图的上下底分布是数据的上

四分位数和下四分位数, 中间的线代表数据的中位数, 即箱体包含了 50% 的数据, 箱子的高度在一定程度上反映了数据的波动程度。

表 1 数据集统计信息

Table 1 The Statistics of the Datasets

	训练用户	测试用户	训练文本	测试文本	主题种类	单用户平均主题种类
BMR	855	931	30083	30474	31	4.39
Agora	3115	4209	175978	179482	17	3.7
SR	6585	8865	379382	381959	15	5.08
SR2	5346	6530	373905	380779	20	4.37

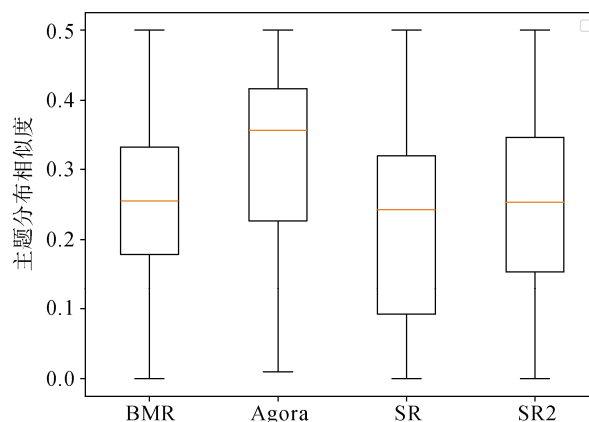


图 8 四个数据集主题分布相似度箱线图

Figure 8 Boxplots of the Similarity of the Topic Distributions of the Four Datasets

根据箱线图中位数和下四分位数可以看出在四个数据集中相对而言 SR 和 SR2 两个数据集整体主题分布相似度相比 BMR 和 Agora 更低, 特别是 SR 数据集的上下四分位数值都相对较低, 证明其存在较多的低主题分布相似度文本, 主题偏差现象表现得更为明显。并且在图中可以看出 SR 和 SR2 这两个数据集中数据会有更明显的范围波动, 即其分布更加分散。BMR 和 SR2 两个数据集中位数接近箱体中间, 而 Agora 和 SR 数据集中位数更接近箱体的上限值, 呈左偏分布。

6.2 基线

本文使用了以下基线和本文的方法进行比较分析。

SYSML^[4]: 一种基于暗网文本的作者归属模型, 他们使用 Text CNN 来编码帖子内容, 并通过异构信息网络(HIN)里建模结构信息, 将这些信息和时间嵌入一起输入 Transformer 编码器并通过平均池化生成最终的嵌入。

URM4DMU^[6]: 一种基于暗网文本的作者归属

模型, 他们的模型通过使用自适应门机制增强卷积运算符和自注意力表示。

IUR^[9]: 一种基于短文本的作者归属模型, 使用组合在线用户的多个帖子来辅助进行作者归属任务, 其使用 BERT 进行微调并通过度量学习更好地捕获有意义的作者向量空间距离。

ARR^[13]: 一种基于明网文本的作者归属模型, 使用了作者表示正则化的蒸馏框架减少对特定于主题的信息的依赖, 在跨主题开放集评估中取得了一定的性能提升。

CNN^[32]: 一种基于短文本的作者归属模型, 它使用字符 CNN^[33]来编码每个帖子, 获取帖子的特征。

BERT^[34]: 一个可以考虑输入文本上下文的双向模型。它是一个基于大规模无标注的数据训练的预训练语言模型, 这使它能够学习到高质量的文本初始化表示, 支持微调以适应各种自然语言处理任务。

Distil BERT^[35]: 通过蒸馏 BERT 模型训练而成, 它在不过分降低性能的前提下, 减少了模型的大小并提升了模型的推理速度。

RoBERTa^[36]: BERT 模型的一种变体, 它在一系列的自然语言处理任务中取得了更先进的结果。

6.3 评估指标

与先前的研究相同, 本文使用 MRR (Mean Reciprocal Rank)和 Recall 作为评价指标来计算作者身份表示的距离以检索相同身份的作者。

MRR: 第一个检索到的文本为同一作者索引的倒数排名。MRR 是 episode 样本倒数排名的平均值, MRR 越高, 同一身份的作者表示越接近。其中 r_{ij} 表示按作者表示 i 和作者表示 j 的相似度排序的 episode 表示, e_i 表示第 i 个 episode 的作者表示,

$A(\bullet)$ 代表将表示映射到原始作者标签。

$$\text{MRR} = \frac{1}{k} \sum_{i=1}^k \frac{1}{\min_j (A(r_{ij}) = A(e_i))} \quad (11)$$

Recall@K: 它表示第一个检索到的相同身份作者是否包含在前 k 个作者表示的子集中。 $I(\bullet)$ 衡量作者是否在前 k 的检索范围内。 N 是输入数据集的样本量。 Recall 值越高, 证明模型的预测结果越好, 此处 k 通常取 10, 意为在召回的前 10 位作者中是否包含目标作者。

$$\text{Recall@K} = \frac{1}{N} \sum_{i=1}^N I(A(ij) = A(\text{rep}_i), k) \quad (12)$$

6.4 实施细节

本文使用 RoBERTa-base 作为语义编码模型, 划分片段数据集的超参数 L 被设置为 5, 批次大小和学习率分别被设置为 32 和 $5e-5$ 。

7 实验结果和分析

7.1 实验结果

在四个暗网论坛数据集下使用本文提出的方法 CIL 与没有引入 CIL 的基线方法的实验结果如表 2 所示, *代表本文微调实现的方法。从表中可以看出, 在同样的设置下, 其他方法相比于本文提出的因果去偏方法 CIL 在 MRR 和 Recall@10 这两个评估指标上都优于基线工作, 这表明本文的方法相比于前人的方法更能捕获作者身份的有效特征信息, 能够取得更好的预测结果。另外, 在使用不同的预训练语言模型(BERT、Distil BERT、RoBERTa)的情况下, 本文的方法均获得了稳定的模型性能提升, 这证明了本文提出的因果去偏方法的模型无关性。

表 2 各模型在四个数据集上的实验结果
Table 2 Experimental Results of Each Model on Four Datasets

方法	BMR		Agora		SR		SR2	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
CNN(2017)	0.07	0.165	0.126	0.214	0.036	0.073	0.082	0.131
IUR(2019)	0.223	0.409	0.114	0.218	0.109	0.19	0.126	0.223
*BERT(2019)	0.337	0.545	0.282	0.411	0.152	0.263	0.198	0.318
*RoBERTa(2019)	0.492	0.684	0.394	0.527	0.29	0.418	0.338	0.463
*Distil BERT(2019)	0.414	0.598	0.403	0.501	0.244	0.355	0.335	0.454
SYSMIL(2021)	0.320	0.533	0.152	0.279	0.157	0.266	0.123	0.210
ARR(2022)	0.439	0.612	0.344	0.472	0.261	0.386	0.367	0.51
UMR4DMU(2023)	0.342	0.547	0.280	0.464	0.254	0.410	0.250	0.425
*BERT+CIL	0.426	0.615	0.368	0.474	0.201	0.301	0.201	0.32
*Distil BERT+CIL	0.437	0.634	0.361	0.482	0.239	0.352	0.317	0.468
*RoBERTa+CIL	0.553	0.729	0.454	0.559	0.358	0.487	0.439	0.570
*RoBERTa+ARR+CIL	0.552	0.710	0.375	0.506	0.379	0.505	0.467	0.589

另外, 相比于之前用于明网数据集的主题正则化去偏方法 ARR, 在单独使用两种主题去偏方法的情况下, 本文的因果去偏方法整体而言取得了更好的预测效果, 在四个暗网论坛数据集上的表现都优于 ARR。

在 BMR 和 Agora 两个数据集上, 单独使用 CIL 方法取得了最好的效果, 这可能是由于这两个数据集文本数量相对较少, 且主题分布相似度相对较高, 即主题偏差现象表现得相对不那么明显, 混合采用两种去偏方法会导致模型过拟合而损失了一些文本原始的特征。而在 SR 和 SR2 这两个数据集上, 混合使用 ARR 和 CIL 两种主题去偏方法取得了更好的效果, 这可能是由于 SR 和 SR2 这两个数据集文本数量更多, 主题分布相似度更低, 主题分布不一致的情况表现得更明显, 因此混合使用两种主题去偏方法更能够使模型充分学习作者自身的特征, 取得更好的预测效果。整体而言, 通过使用本文提出的 CIL 方法, 模型能够更充分地学习到同一作者在不同主题下不变的内容和风格特征, 并捕获不同作者在同一主题下的特征差异, 从而实现了对作者身份特征的良好建模, 在作者归属任务下取得良好的预测结果。

7.2 消融实验

7.1 节中的实验结果已经充分证明了本文提出的 CIL 方法的有效性, 但目前还不能明确不同模块在训练过程中起到的作用。因此, 本文继续对前文提出的不同模块进行消融实验, 比较模型在不同模块设置的条件下的性能差异, 以清晰地展示出各个模块的作用。实验的结果如表 3 所示。

1) 使用数据集的策略: 首先对两种构建数据集的策略进行了比较, 其中 (E) 代表使用片段数据集的结果, (T) 代表使用主题数据集的结果。从实验结果可以看出使用本文提出的构建主题数据集对基础模型直接进行微调的效果相对之前的数据集构造方法表现更佳, 在四个数据集上的 MRR 和 Recall@10 实现了平均 4.38% 和 3.18% 的性能提升。这说明使用按主题划分的数据集能够促使模型充分学习同一作者在各主题下的内容特征和风格特征, 捕获到更多作

者的独特特征。

2) 使用因果去偏模块: 这里比较了使用因果去偏模块在两类数据集下的结果, CIL 标识在基础模型上加入因果去偏模块。从结果上看, 使用因果去偏模块后, 模型的预测性能得到了显著提升, 而相比于在片段数据集上的预测结果, 在主题数据集上使用 CIL 方法后, 模型的预测结果表现得更好, 模型在四个数据集上的 MRR 和 Recall@10 实现了平均 7.25% 和 6.33% 的性能提升, 这证明了本文提出的因果去偏模块有效地促进模型学习同一作者不同主题语义空间的不变的身份特征, 使模型减少受到数据中虚假相关性的影响, 捕获更有效标识作者身份的特征, 最终实现良好的预测效果。同时也说明将主题视为混淆因子, 考虑主题分布不一致造成的影响对提升模型的性能是有价值的。通过使用主题数据集和因果去偏模块, 模型的性能得到了有效的提升。

7.3 稳定性实验

另外, 为了进一步验证本文的方法是否提升了模型在不同主题数据集下预测性能的稳定性。此处选取了十个典型的主题, 选择在对应主题下发帖数量最多的用户作为测试用户数据集, 以验证本文所提出的方法在去除其偏好主题的数据集训练前后预测性能是否更具有稳定性, 选择的测试数据集具体情况如表 4 所示。表 4 统计了所选的十个主题中各主题文本数量, 以及在对应主题下发帖数量最多的用户数量, 并将该数量作为测试用户数量。从表中可以看出从整体上看主题文本数量越多, 对应的测试用户数量越多。其中新人讨论、科技、安全、销售这几个主题对应的测试用户数量相对较少, 特别是新人讨论主题的数据集表现得较为明显。

图 9 比较了测试用户在使用去除对应偏好主题情况下数据集相比于使用完整数据集进行训练时模型的预测性能方差情况。

从图中可以看出从整体上来看使用本文的因果去偏方法之后在选取的十个测试主题下去除用户偏好主题数据集相比于使用完整数据集进行训练时模型的预测性能方差更小, 即去除偏好主题数据后模

表 3 消融实验结果

Table 3 Ablation Experimental Results

方法	BMR		Agora		SR		SR2	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
(E)RoBERTa	0.492	0.684	0.394	0.527	0.29	0.418	0.338	0.463
(T)RoBERTa	0.526	0.700	0.427	0.545	0.319	0.448	0.417	0.526
(E)RoBERTa +CIL	0.536	0.709	0.423	0.547	0.336	0.458	0.417	0.543
(T) RoBERTa +CIL	0.553	0.729	0.454	0.559	0.358	0.487	0.439	0.570

表 4 稳定性实验中测试数据集统计信息

Table 4 The Statistics of the Test Datasets in the Stability Experiment.

	文本数量	测试用户数量
通用讨论	8802	337
德语	4264	152
新人讨论	1778	14
荷兰语	1446	48
毒品	1077	41
谣言	973	38
科技	734	8
安全	559	11
销售	498	18
武器	386	34

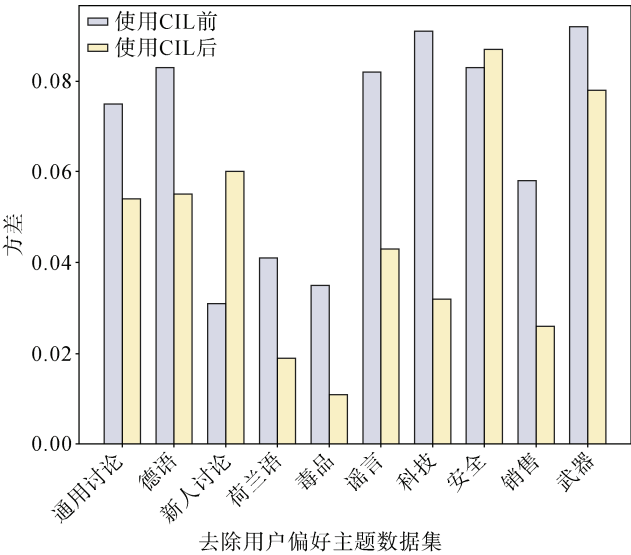


图 9 去除用户偏好主题数据集相比于使用完整数据集进行训练时模型使用 CIL 前后的预测性能方差图

Figure 9 The Variance of the Model's Predictive Performance Before and After CIL When the Model is used for Training with the User Preference Topic Dataset Compared to the Full Dataset for Training

型的预测性能相比于使用 CIL 方法前稳定性显著提升。这证明了预测的结果相较于使用 CIL 方法前受

主题的影响更小, 模型受训练集内主题分布的影响相较于之前更小, 在不同的主题分布条件下模型都能取得更稳定和良好的预测结果, 这充分证明了本文提出的去偏方法的有效性。

在选取的十个主题测试集中, 绝大部分情况下使用 CIL 后的预测性能更稳定, 而去除新人讨论主题数据集时, 使用 CIL 前的模型性能预测更稳定, 此处主要原因可能是新人讨论主题的文本数量很多, 而偏好此主题的用户较少, 即该主题下测试集中测试用户量相对较少, 在测试样本不足的条件 CIL 模型的测试表现存在一定的不稳定性。

7.4 泛化性实验

为了验证本文的方法有助于捕获作者的主题无关特征, 本文划分了跨主题测试数据集, 旨在验证训练集观察到的主题和学习到的作者信息特征表示具有最小相似度的设置下模型的有效性。这里采样了三类跨主题数据集, 分别为武器、数据泄露和毒品, 采样时保证训练集的数据不包含指定三类主题数据集的文本, 而在测试集仅使用这三类主题数据集的文本对模型性能进行验证。通过这样的方式, 确保模型在指定的主题数据集下训练, 在从未见过的主题数据集下进行测试, 以验证本文提出的去偏方法的有效性和分布外泛化性, 如表 5 所示。

此处由于考虑的是跨主题场景, 仅在除测试主题外的其他主题的文本训练集上训练模型, 训练文本量相对减少, 因此整体模型的表现结果相比表 2 有所下降。但从表中结果可以看出在使用了本文的因果去偏方法后, 在三个训练阶段从未见过的主题测试集下的各项评估指标都取得了更好的结果。这表明使用本文的因果去偏方法训练出的模型对数据集中的主题线索信息依赖更少, 模型倾向于捕获与主题分布无关的作者身份特征, 捕获不同主题下作者的不变特征, 这促使模型减少对于特定单个主题信息的依赖, 使模型能够更有效地泛化到未知主题的场景。

表 5 泛化性实验结果

Table 5 Generalization Experimental Results

测试主题	方法	MRR	Recall@1	Recall@2	Recall@5	Recall@10
武器	RoBERTa	0.478	0.397	0.483	0.543	0.647
	RoBERTa +CIL	0.486	0.405	0.467	0.595	0.655
数据泄露	RoBERTa	0.419	0.350	0.398	0.515	0.592
	RoBERTa +CIL	0.477	0.388	0.476	0.583	0.650
毒品	RoBERTa	0.273	0.217	0.271	0.325	0.379
	RoBERTa +CIL	0.339	0.289	0.336	0.397	0.430

7.5 主题分布相似度低的用户去偏结果分析

为了探究本文的去偏方法对主题分布相似度低的用户的影响, 本节对四个暗网论坛数据集中主题分布相似度小于 0.1 的用户进行分析, 同时为

了排除训练集测试集数量差距较大的极端案例干扰, 选取其中训练集测试集文本数量分布接近 1: 1 的用户, 对比其使用 CIL 方法前后的性能差异并进行分析, 如图 10 所示。

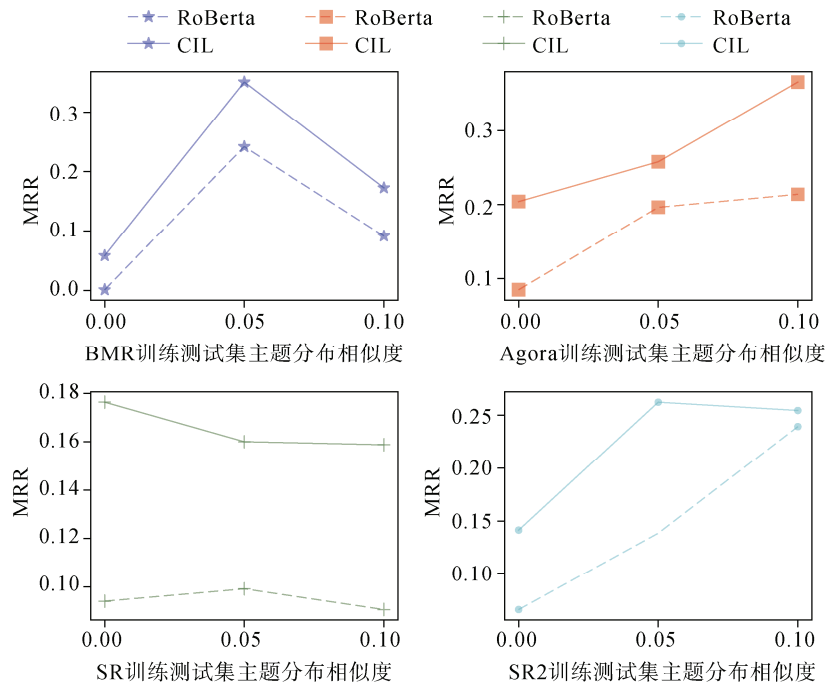


图 10 主题分布相似度小于 0.1 的用户在使用 CIL 前后的性能比较

Figure 10 The Performance of Users with Similarity of Topic Distribution Less Than 0.1 Before and After Using CIL Method

从图中可以清晰地看出使用因果去偏方法后, 模型在四个暗网论坛数据集下的预测性能都有显著的提升, 这证明了使用本文方法进行主题去偏的作者身份相比于原方法减少了对主题分布的依赖, 在主题分布相似度低的用户中取得了预测性能的提升。其中 BMR 数据集由于整体文本数量相比其他数据集更少, 主题分布相似度低的用户数量相比也更少, 因此表现具有一定的不稳定性, 波动较为明显。

另外, 图中存在一些使用去偏方法前后性能变化不大的情况, 如 SR2 数据集中主题相似度为 0.1 的部分用户, 本文分析了数据集中此类去偏效果不佳的用户, 并从中抽取了一些例子, 总结难以正确识别作者身份的帖子类型并分析其原因, 如表 6 所示。

通过分析发现这些帖子可识别度较低的主要原因如下: 暗网论坛中存在一些多语言数据, 具体而言, 有部分作者的帖子大多为英语, 少部分为其他语言如意大利语、法语、德语等, 现有的方式无法捕捉作者在不同语言间的相似特征, 当用户的

文本存在多语言混杂现象, 即用户的文本分布在不同语言的主题上时, 模型难以学习到有效的作者身份特征, 导致难以做出正确的预测, 最终表现的效果不佳。

表 6 识别度低的帖子案例分析		
Table 6 Case studies of low-recognizable posts		
难以正确预测的原因	帖子 1	帖子 2
多语言数据	Nice one! Hope this will safe some Users from being scammed.	Dacht ich mir schon. Danke für die Rückmeldung :]
	[QUOTE] True Story.	[QUOTE] agree with you

另外, 文本过短时往往不具备明显的特征, 特别是一些引用他人帖子的文本, 本文使用的暗网论坛数据集在预处理阶段使用[QUOTE]标记标识被帖子引用的其他帖子信息, 这种预处理方式由于无法获取引用文本, 原文的语义信息丢失, 从这种短文本中难以获取充分的作者身份特征信息供模型学习。训练集中存在大量这样的帖子会导致模

型难以学习良好的作者表示, 使用去偏方法前后模型的预测结果相近。

7.6 超参数实验

为了设置合适的超参数 λ , 图 11 在 BMR 验证集下比较了超参数设置在 0.1 到 1 区间内模型性能的变化情况, 从图中可以看出, 当超参数设置为 0.3 时模型的性能表现最佳, 因此, 本文的实验设置均在超参数 λ 设置为 0.3 的条件下完成。

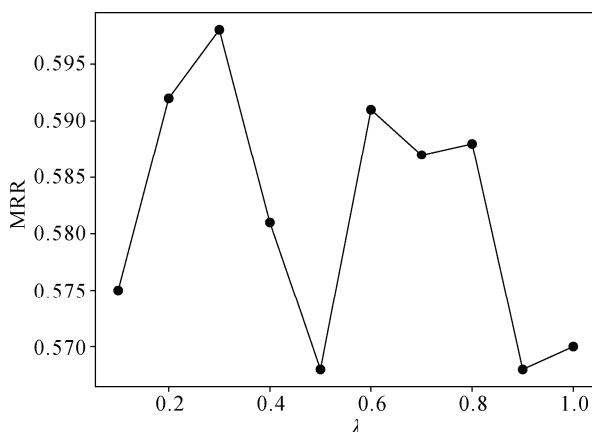


图 11 BMR 验证集下模型性能随超参数 λ 的变化情况
Figure 11 The Variation of Model Performance with Hyperparameter λ Under the BMR Validation Dataset

8 总结和未来工作

总的来说, 本文提出了一个暗网论坛中作者归属任务的因果去偏方法。本文首先从因果的视角解释了主题分布不一致给作者归属分析任务引入偏差的原因, 通过按主题划分数据集并使用因果不变性正则化项缓解训练集测试集主题分布不一致对作者表示造成的影响, 以此构建更稳健的作者身份表示, 更好地实现作者归属任务。在四个暗网市场数据集下的实验结果证实了本文方法相比于其他方法的有效性。

未来的工作主要会从以下几个方面进行, 首先是对主题的划分, 现有的主题划分方法是按子论坛的主题进行了粗粒度的划分, 因此其中存在一些主题重叠或混乱的现象, 即部分主题中混杂着其他主题, 存在一些主题不明确的现象, 这些情况造成的影响还需要继续进行研究, 后续考虑使用主题模型或引入大规模语言模型尝试进一步提取主题进行研究。

针对短文本特征不足的问题, 可能可以从其他特征的角度考虑。比如从语言的因素考虑, 尝试捕获短文本中的多语言特征, 改进短文本中主题特征不明显, 难以捕获有效特征的问题。对于多语言场景下识别度较低的情况, 可以进一步考虑使用多语言语

义编码模型作为基础模型来解决帖子中存在的多语言文本识别困难的问题, 进一步提高作者归属任务的识别准确率。

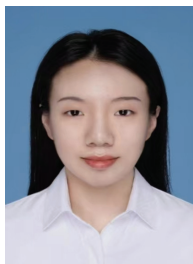
另外, 未来的工作会考虑继续完善作者归属任务的因果图, 当前的因果图主要是针对单暗网平台的场景, 将主要制约模型性能的主题偏差因素作为混淆因子进行讨论, 这在多平台跨域场景下可能存在一定的局限性。未来会继续尝试探讨多平台的跨域场景下影响模型性能的因素, 比如考虑将平台的风格等因素加入因果图进行讨论。

致谢 这项工作得到了浙江大学区块链与数据安全国家重点实验室开放研究基金和青年科学基金项目(No. 6240071954)的支持。This work was supported by the Open Research Fund of The State Key Laboratory of Blockchain and Data Security, Zhejiang University.

参考文献

- [1] Barbalau A, Brad F, Ionescu R T, et al. VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web[C]. *Advances in Neural Information Processing Systems* 35, 2022: 15574-15588.
- [2] Madigan D. Bayesian Multinomial Logistic Regression for Author Identification[J]. *AIP Conference Proceedings*, 2005, 803(1): 509-516.
- [3] Hay J, Doan B L, Popineau F, et al. Representation Learning of Writing Style[C]. *The Sixth Workshop on Noisy User-generated Text*, 2020: 232-243.
- [4] Maneriker P, He Y T, Parthasarathy S. SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis[C]. *The 2021 Conference on Empirical Methods in Natural Language Processing*, 2021: 6844-6857.
- [5] Saedi C, Dras M. Siamese Networks for Large-Scale Author Identification[J]. *Computer Speech & Language*, 2021, 70: 101241.
- [6] Ordoñez J, Soto R R, Chen B Y. Will longformers PAN out for authorship verification[J]. *Working Notes of CLEF*, 2020.
- [7] Fabien M, Villatoro-Tello E, Motlicek P, et al. BertAA: BERT Fine-Tuning for Authorship Attribution[C]. *The 17th International Conference on Natural Language Processing*, 2020: 127-137.
- [8] Liu H M, Zhao J P, Huo Y X, et al. URM4DMU: An User Representation Model for Darknet Markets Users[C]. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023: 1-5.
- [9] Andrews N, Bishop M. Learning Invariant Representations of Social Media Users[C]. *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019: 1684-1695.
- [10] Rivera-Soto R A, Miano O E, Ordonez J, et al. Learning Universal Authorship Representations[C]. *The 2021 Conference on Empiri-*

- cal Methods in Natural Language Processing*, 2021: 913-919.
- [11] Ai B, Wang Y C, Tan Y, et al. Whodunit? Learning to Contrast for Authorship Attribution[C]. *The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2022: 1142-1157.
 - [12] Tyo J, Dhingra B, Lipton Z C. On the State of the Art in Authorship Attribution and Authorship Verification[EB/OL]. 2022: arXiv: 2209.06869. <https://arxiv.org/abs/2209.06869>.
 - [13] Riddell A, Wang H N, Juola P. A Call for Clarity in Contemporary Authorship Attribution Evaluation[C]. *The Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, 2021: 1174-1179.
 - [14] Altakrori M, Cheung J C K, Fung B C M. The Topic Confusion Task: A Novel Evaluation Scenario for Authorship Attribution[C]. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021: 4242-4256.
 - [15] Sawatphol J, Chaiwong N, Udomcharoenchaikit C, et al. Topic-Regularized Authorship Representation Learning[C]. *The 2022 Conference on Empirical Methods in Natural Language Processing*, 2022: 1076-1082.
 - [16] Wang X W, Peng P, Wang C, et al. You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces[C]. *The 2018 on Asia Conference on Computer and Communications Security*, 2018: 431-442.
 - [17] Jeziorowski S, Ismail M, Siraj A. Towards Image-Based Dark Vendor Profiling: An Analysis of Image Metadata and Image Hashing in Dark Web Marketplaces[C]. *The Sixth International Workshop on Security and Privacy Analytics*, 2020: 15-22.
 - [18] Zhang Y M, Fan Y J, Song W, et al. Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network[C]. *The World Wide Web Conference*, 2019: 3448-3454.
 - [19] Arabnezhad E, La Morgia M, Mei A, et al. A Light in the Dark Web: Linking Dark Web Aliases to Real Internet Identities[C]. *2020 IEEE 40th International Conference on Distributed Computing Systems*, 2020: 311-321.
 - [20] Munksgaard R, Demant J. Mixing Politics and Crime – the Prevalence and Decline of Political Discourse on the Cryptomarket[J]. *International Journal of Drug Policy*, 2016, 35: 77-83.
 - [21] Pearl J. Causality[M]. Cambridge, U.K. New York: Cambridge University Press, 2009.
 - [22] Pearl J, Mackenzie D. The book of why: the new science of cause and effect[M]. Basic books, 2018.
 - [23] Qi J X, Niu Y L, Huang J Q, et al. Two Causal Principles for Improving Visual Dialog[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10857-10866.
 - [24] Heinze-Deml C, Meinshausen N. Conditional Variance Penalties and Domain Shift Robustness[J]. *Machine Learning*, 2021, 110(2): 303-348.
 - [25] Zhang W K, Lin H Y, Han X P, et al. De-Biasing Distantly Supervised Named Entity Recognition via Causal Intervention[C]. *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021: 4803-4813.
 - [26] Gong M M, Zhang K, Liu T L, et al. Domain Adaptation with Conditional Transferable Components[C]. *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016: 2839-2848.
 - [27] Wu Q T, Zhang H R, Yan J C, et al. Handling Distribution Shifts on Graphs: An Invariance Perspective[EB/OL]. 2022: arXiv: 2202.02466. <https://arxiv.org/abs/2202.02466>.
 - [28] Mitrovic J, McWilliams B, Walker J, et al. Representation Learning via Invariant Causal Mechanisms[EB/OL]. 2020: arXiv: 2010.07922. <https://arxiv.org/abs/2010.07922>.
 - [29] Elazar Y, Goldberg Y. Adversarial Removal of Demographic Attributes from Text Data[C]. *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018: 11-21.
 - [30] Faal F, Yu J, Schmitt K. Domain Adaptation Multi-Task Deep Neural Network for Mitigating Unintended Bias in Toxic Language Detection[C]. *The 13th International Conference on Agents and Artificial Intelligence*, 2021: 932-940.
 - [31] Ball-Burack A, Lee M S A, Cobbe J, et al. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection[C]. *The 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021: 116-128.
 - [32] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. *The 2014 Conference on Empirical Methods in Natural Language Processing*, 2014: 1746-1751.
 - [33] Shrestha P, Sierra S, Gonzalez F, et al. Convolutional Neural Networks for Authorship Attribution of Short Texts[C]. *The 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017: 669-674.
 - [34] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[C]. *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019: 4171-4186.
 - [35] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. *arXiv: Computation and Language*, 2019.
 - [36] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL]. 2019: arXiv: 1907.11692. <https://arxiv.org/abs/1907.11692>.



张慧妍 于 2022 年在北京邮电大学信息安全专业获得学士学位。现在北京邮电大学网络空间安全专业攻读硕士学位。研究领域为暗网威胁情报分析。Email: zhang_hy@bupt.edu.cn



霍艺璇 于 2023 年在北京邮电大学获得学士学位。现在北京邮电大学网络空间安全专业攻读硕士学位。研究领域为网络威胁情报分析。Email: huoyixuan@bupt.edu.cn



赵佳鹏 于 2022 年在中国科学院大学获得博士学位。现任北京邮电大学博士后, CCF 会员。研究领域为网络空间安全、匿名通信与暗网、自然语言处理。Email: zhaojp9574@bupt.edu.cn



王学宾 于 2022 年在中国科学院信息工程研究所获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为匿名通信与暗网、区块链安全、去中心化社交网络。Email: wangxuebin@iie.ac.cn



孙岩炜 于 2018 年在中国科学院大学获得博士学位。现任北京邮电大学讲师。研究领域为网络威胁发现、态势感知、攻击溯源、威胁情报运营。Email: sunyw@bupt.edu.cn



赵璨 于 2017 年在中国科学院信息工程研究所获得硕士学位。现任中国科学院信息工程研究所工程师。研究领域为匿名通信与暗网内容监测分析。Email: canzhao@iie.ac.cn



时金桥 于 2007 年在哈尔滨工业大学获得博士学位。现任北京邮电大学网络空间安全学院教授、博士生导师。研究领域为网络测量、匿名通信与隐私保护、区块链安全、信息智能处理与大数据安全分析、网络模拟仿真等。Email: shijinqiao@bupt.edu.cn