

北京交通大学

硕士学位论文

基于图表示学习的谣言检测方法研究

Research on Rumor Detection Methods Based on Graph
Representation Learning

作者：乔禹涵

导师：贾彩燕

北京交通大学

2024 年 6 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名：乔禹涵

导师签名：贾新燕

签字日期：2024年 6月 1日

签字日期：2024年 6月 1日

学校代码: 10004

密级: 公开

北京交通大学

硕士学位论文

基于图表示学习的谣言检测方法研究

Research on Rumor Detection Methods Based on Graph
Representation Learning

作者姓名: 乔禹涵

学 号: 20120400

导师姓名: 贾彩燕

职 称: 教授

学位类别: 工学

学位级别: 硕士

学科专业: 计算机科学与技术

研究方向: 谣言检测

北京交通大学

2024 年 6 月

致谢

首先感谢下列基金项目对本研究工作的支持：国家自然科学基金（61876016）、国家重点研发计划，“新一代人工智能”重大项目（2018AAA0100302）、百度松果基金项目。

感谢我的导师贾彩燕老师，贾老师是我科研之路的启蒙导师，在我研一入学时就为我指引了科研方向，在自己两个研究工作中贾老师也给予了我莫大的帮助，给我提供了新颖的研究思路，并为我的小论文仔细修改和把关。贾老师为人谦和，在生活上也帮我解决了一些困难。我深深感恩和贾老师的这段师徒缘分！

感谢同组的全体同学，每当我的学习状态有所松懈，看到同学们每周在组会上准备的精彩汇报都会使我受到鞭策与激励。感谢冉宏艳、郑忆美师姐对我科研和学习上的帮助，特别感谢王奕滢师妹和崔超群师弟对我第二个研究工作所做的贡献！

感谢我的父亲和母亲，在读研究生期间父母给我提供了经济上的支持，让我吃饱穿暖，为我排除了学习以外的一切顾虑。感谢他们对我的爱与牵挂！

感谢北京交通大学，疫情三年，在交大读研阶段所经历的种种困难让我的身心得到磨练和成长，使我在以后的科研和生活中遇到困难时不再畏惧，而是迎难而上。在交大读研这些年，通过不断地“思”与“行”，我逐渐感悟到交大“知行”校训的意义，明白了“致良知”和“知行合一”的道理，并在生活中践行。“千里之行，始于足下”，在今后漫长的科研路上，希望自己每天能够勤奋自律，不断反思并提升自己，心无旁骛地专注于自己未来的科研方向！

最后感谢各位论文评审老师、答辩老师对本篇论文所提的宝贵建议！

摘要

国内外的网络社交媒体平台已成为大众获取信息的主要渠道。然而，便捷的信息获取方式也为虚假信息的传播提供了有利条件。大众对于谣言和虚假新闻的识别能力相对有限，盲从传播将会对社会安定和公众利益造成损害。因此，高效准确地进行谣言检测至关重要。

近年来，基于深度学习的谣言检测方法不断涌现。其中，大量谣言检测方法利用了谣言的传播结构信息，并借助具有强大表达能力的图神经网络提取谣言的高层语义表示和判别特征，获得了出色的性能表现。然而，这些谣言检测模型的成功也伴随着巨大的成本。由于谣言数据的标注耗时耗力，有标注的谣言数据难以大量获得。在有标注的谣言训练样本不足的情况下，各种精心设计的谣言检测模型可能面临过拟合的风险，表现出较差的鲁棒性与泛化性。

本文针对以上问题展开研究，主要研究内容及取得的成果如下：

(1) 提出了一种基于图自监督对比学习的谣言检测方法 RD-GCSL。该方法将谣言检测视为图分类问题，采用图自监督对比学习方法，针对谣言特点设计了三种图数据增强策略。该方法对原图进行两种不同类型的数据增强得到扰动图，再将其输入到图编码器得到图的表示向量；并通过判断两个扰动图是否来源同一原始图，构建正负样本对，建立自监督对比损失，并将其作为有监督分类主损失的正则项以提升模型整体的泛化性能。借助该对比学习方法，RD-GCSL 将有监督分类和自监督对比学习任务统一到一个框架下进行端到端的联合训练，以捕获谣言更趋向本质的特征。三个谣言公开数据集上的实验结果显示：RD-GCSL 在有监督谣言检测任务中比基准方法取得了更高的准确率，在一定程度上缓解了过拟合的负面影响，提高了模型的泛化性与鲁棒性。

(2) 提出了一种自监督预训练辅助的消偏自训练半监督谣言检测方法 RDST。该方法采用了自训练的策略，在有限有标注训练样本的条件下增强了半监督图表示学习谣言检测模型的性能。具体地，该方法首先利用对比式或生成式图自监督学习策略，使用大量无标注谣言传播结构图对谣言图编码器进行自监督预训练。然后，将其作为自训练的初始模型，以减少自训练初始阶段错误伪标签的生成。并且，针对自训练过程中的伪标注样本选择问题，设计了一种自适应阈值伪标签选择策略，以提升伪标签的选取质量。四个公开数据集上的实验结果显示：RDST 在半监督谣言检测任务中的表现大幅超过了其他基线模型，并在有标注数据极少的情景下仍具有良好表现，验证了所提消偏自训练方法的有效性。

关键词：谣言检测；图表示；自监督学习；半监督学习；自训练

ABSTRACT

Domestic and international social media platforms have become the main channels for the public to access information. However, the convenient access to information has also created favorable conditions for the spread of false information. The public's ability to identify rumors and fake news is relatively limited, and blind dissemination may cause harm to social stability and public interests. Therefore, it is crucial to efficiently and accurately detect rumors.

In recent years, there has been a continuous emergence of rumor detection methods based on deep learning. Among them, a large number of rumor detection methods utilize the structural information of rumor propagation and leverage graph neural networks with powerful expressive capabilities to extract high-level semantic representations and discriminative features of rumors, achieving outstanding performance. However, the success of these rumor detection models comes with significant costs. In particular, the annotation of rumor data is time-consuming and labor-intensive, making it hard to acquire a large amount of labeled rumor data. Faced with a shortage of labeled rumor training samples, various carefully designed rumor detection models may face the risk of overfitting, exhibiting poor robustness and generalization.

This paper addresses the aforementioned issues and presents the following research contributions and achievements:

(1) A rumor detection model RD-GCSL based on graph self-supervised contrastive learning is proposed. The model considers rumor detection as a graph classification problem, employing the method of graph self-supervised contrastive learning and designing three graph data augmentation strategies considering the characteristics of rumors. The method first applies two types of data augmentation to the original graph to obtain perturbed graphs, which are then input into a graph encoder to obtain graph representations. By judging whether two perturbed graphs originate from the same original graph, positive and negative sample pairs are constructed to establish a self-supervised contrastive loss. With the aid of this contrastive learning method, the RD-GCSL model integrates supervised classification and self-supervised contrastive learning tasks into a unified framework for end-to-end joint training, aiming to capture more intrinsic features of rumors. Experimental results on three public rumor datasets demonstrate that the RD-GCSL model achieves better performances than other baseline methods, alleviating overfitting to some extent and improving the model's generalization and robustness.

(2) A debiased self-training framework with graph self-supervised pre-training aided for semi-supervised rumor detection is proposed. This framework adopts a self-training approach to enhance the performance of semi-supervised graph representation learning methods for rumor detection under limited annotated samples. Specifically, the framework initially utilizes contrastive or generative graph self-supervised learning methods to perform self-supervised pre-training on the rumor graph encoder using the propagation graph structures of a large amount of unlabeled rumor data. Then, it utilizes this pre-trained model as the initial model for self-training to reduce the generation of erroneous pseudo-labels in the initial stages of self-training. Additionally, to address the issue of pseudo-label sample selection in self-training, a pseudo-labeling strategy based on self-adaptive thresholds is designed to improve the quality of pseudo-label sample selection. Experimental results on four public datasets demonstrate that RDST significantly outperforms other baseline models in semi-supervised rumor detection tasks and maintains good performance even in scenarios with very few labeled data, validating the effectiveness of the proposed debiased self-training framework.

KEYWORDS: rumor detection; graph representation; self-supervised learning; semi-supervised learning; self-training

目录

摘要	iii
ABSTRACT	iv
1 引言	1
1.1 研究背景及意义	1
1.2 谣言检测研究现状	3
1.2.1 基于特征工程的传统机器学习方法	3
1.2.2 基于谣言文本内容和时间序列的深度学习的方法	4
1.2.3 基于谣言传播结构的深度学习的方法	5
1.3 现有研究存在的问题	6
1.4 主要研究内容	8
1.5 本文组织架构	9
2 相关理论及工作	10
2.1 基于图表示学习的谣言检测方法	10
2.1.1 图神经网络	10
2.1.2 文本表示学习方法	11
2.1.3 代表性工作及经典模型	12
2.2 图自监督学习方法	13
2.2.1 相关工作	13
2.2.2 图对比学习方法	14
2.2.3 图自编码器方法	16
2.3 半监督学习方法	18
2.3.1 相关工作	18
2.3.2 自训练方法	18
2.4 数据集与评价指标	20
2.4.1 Twitter15 和 Twitter16 数据集	20
2.4.2 Weibo 和 DRWeibo 数据集	20
2.4.3 PHEME 数据集	21
2.4.4 评价指标	22
2.5 本章小结	23

3	基于图自监督对比学习的谣言检测方法	24
3.1	研究动机	24
3.2	模型方法	25
3.2.1	问题描述	25
3.2.2	数据增强	26
3.2.3	图编码器	27
3.2.4	对比学习损失	28
3.2.5	谣言分类器	28
3.3	实验设计与结果分析	29
3.3.1	数据集与参数设置	29
3.3.2	基线模型	30
3.3.3	实验结果与分析	31
3.3.4	消融实验	33
3.4	本章小结	36
4	自监督预训练辅助的消偏自训练谣言检测方法	37
4.1	研究动机	37
4.2	模型方法	38
4.2.1	问题描述	38
4.2.2	模型初始化	40
4.2.3	自适应阈值伪标签选择方法	43
4.2.4	训练策略	45
4.3	实验设计与结果分析	46
4.3.1	数据集	46
4.3.2	基线模型	47
4.3.3	实验设置与参数选择	48
4.3.4	实验结果与分析	49
4.3.5	消融实验	51
4.3.6	参数分析	53
4.3.7	特征可视化	56
4.4	本章小结	57
5	总结与展望	59
5.1	工作总结	59

5.2 未来展望	60
参考文献	61
作者简历及攻读硕士学位期间取得的研究成果	66
独创性声明	67
学位论文数据集	68

1 引言

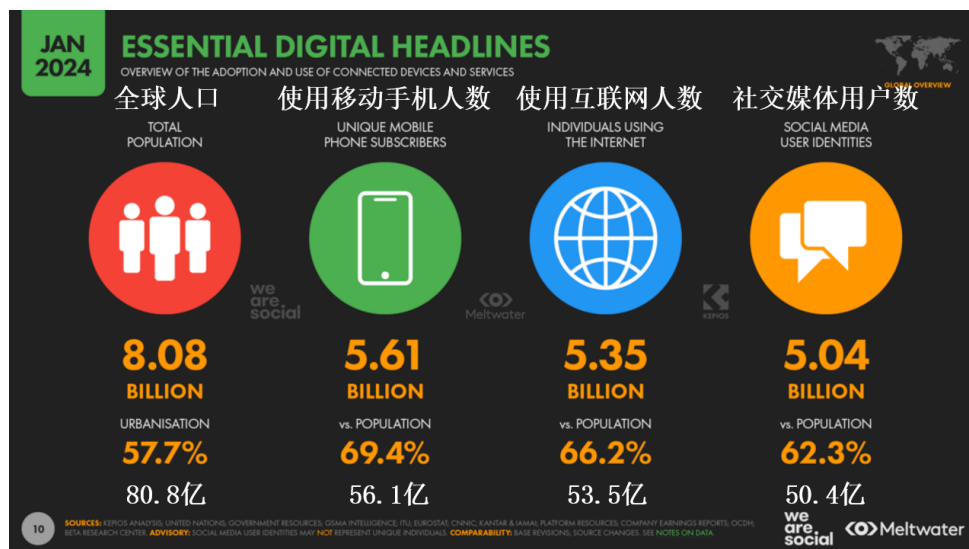
本章将主要阐述谣言检测的相关背景及研究意义，概述国内外谣言检测领域的研究现状和现有研究中所存在的问题，并引出本文重点展开研究的内容。最后将对本文各章节的内容和组织结构进行简要概述。

1.1 研究背景及意义

随着互联网和移动通信技术的迅速进步，社交媒体得以快速普及和发展。高速的网络连接、智能手机的普及使得人们能够随时随地与他人进行即时沟通和信息分享。近年来社交网络的用户数量迅猛增加，据 Digital2024 的报告显示^[1]，截至 2024 年 1 月，全球互联网用户总数已达 53.5 亿，目前全球已有超过 66% 的人使用互联网。其中，活跃的社交媒体用户数量已经超过 50 亿，相当于全世界人口的 62.3% (如图1-1所示)。在过去一年中有 2.66 亿新用户开始使用社交媒体，增长了 5.6%。这个数字意味着全球平均每秒就有 8.4 个新增的社交媒体用户。在使用时间上，“典型”的社交媒体用户平均每天花在社交媒体上的时间为 2 小时 23 分钟。在对人们使用互联网的目的调查中，近 61% 的受访者表示，“寻找信息”是他们使用互联网的主要原因之一，而“与朋友和家人保持联系”以 56.6% 排在第二位。在全世界最热门的几个社交媒体平台中，Facebook 的每月活跃用户达到了 30.5 亿，Instagram 的活跃用户达到了 16.5 亿，国内的新浪微博也已有 6.05 亿的活跃用户。显然，社交媒体已成为信息传播和社会互动的重要平台。得益于社交媒体的便捷性，每个人都可以轻松地分享自己的观点和生活，进行各种社交活动，乃至参与到社会各种热议话题的讨论中。

然而，便捷的信息获取方式也为虚假信息的传播提供了有利条件。由于在社交媒体发布、传播信息所需的低成本，许多恶意用户以各种目的来制造、传播各类谣言、虚假新闻等虚假信息。而大众对于谣言和虚假新闻这类虚假信息的辨别能力相对有限，盲目传播虚假信息将会对社会安定和公众利益造成损害。举例来说，在新冠肺炎期间，一则“喝高浓度酒精能使自身对新冠病毒免疫”的谣言造成了至少 800 人死亡、5000 人住院、60 人永久性受伤¹。短短一则谣言却产生了如此强的破坏力，虚假谣言不仅危害了个人的生命健康，也影响了社会的安定和发展。近年来，“生成式大模型”技术火遍全球，各种由 AI 伪造的文本、图片、视频层出不穷，为虚假信息检测带来了新的挑战，如何自动高效地进行谣言检测变得更加

¹<https://www.bbc.co.uk/news/world-53755067>

图 1-1 互联网和社交媒体的使用情况统计^[1]Figure 1-1 Statistics on the usage of the Internet and social media^[1]

至关重要。

社会心理学相关文献^[2]通常将谣言 (rumor) 定义为一个广泛传播的未经证实或故意捏造的事件。而谣言检测的任务目标就是对社交媒体中用户所发布的源帖子 (source posts) 的真假性进行分类。对于不同的数据集, 谣言具有不同的分类标准, 常用的中文数据集 Weibo^[3] 是一个二分类数据集, 它将所有源帖子分为: 非谣言 (non-rumor, NR) 和谣言 (rumor, R) 两大类, 在此分类方式下 “rumor” 被理解为人们常说的 “谣言” (通常为假事件), 因此, 在此情景下谣言检测任务仅需预测一个事件是真或假。而英文数据集 Twitter15^[4] 和 Twitter16^[4] 提供了更细粒度的谣言划分方式, 它们将源帖子的真假性分为了四类, 在此情景下, “rumor” 的真假性并不确定, 因此与人们常说的 “谣言” 有所不同, “rumor” 一词理解成 “流言” 更加恰当。四种谣言类别中: true rumor (TR) 表示源帖子发布的内容经验证后其信息是真实的, false rumor (FR) 则表示源帖子经验证后其信息是虚假的, unverified rumor (UR) 是指源帖子的内容还未经过验证, 其真假性未知, 而 non-rumor (NR) 是指源帖子的内容并不具有传播性, 可能为用户抒发个人情感的帖子, 不具有社会影响力。与谣言检测任务类似, 假新闻检测 (fake news detection) 也是虚假信息检测中的一个重要领域, 其任务目标大多为二分类, 即判断新闻事件的真假性。谣言检测任务与假新闻检测任务的一个不同之处在于, 谣言检测通常会利用源帖子的相关帖子 (转发、回复) 信息来进行真假性判断, 而多数假新闻检测的数据集并没有提供与事件相关的回复评论信息。本文将主要对谣言检测任务展开研究。如图1-2所示, 为新浪微博平台一则谣言的实例, 包含了源帖子及其相关 (转发、回复) 帖子。

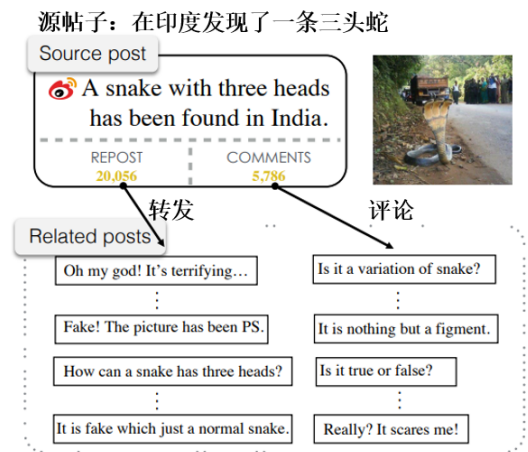


图 1-2 新浪微博平台一则谣言的实例^[5]

Figure 1-2 An instance of a rumor on the Sina Weibo platform^[5]

实际上，对一则谣言进行真假性判断通常需要耗费大量的人力财力，通过聘请领域相关的专家，收集大量相关证据和各种权威性的报告，并对谣言内容进行全面分析，最终才能确定谣言的真假性^[6]。面对互联网上浩如烟海的各种信息，按照上述步骤进行人工检测并不现实，如何自动高效地进行谣言真假性分类是近年谣言检测任务的重点研究内容。

1.2 谣言检测研究现状

谣言检测是自然语言处理领域的一个子任务方向，随着机器学习和人工智能领域知识的快速进步，近年来已有大量相关工作对自动谣言检测进行研究。根据这些工作的不同发展阶段和所利用的不同信息，大致可将国内外的谣言检测相关研究分成以下三大类：（1）基于特征工程的传统机器学习方法；（2）基于谣言文本内容和时间序列的深度学习方法；（3）基于谣言传播结构的深度学习方法。在后续各子节中将分别对以上各类方法进行详细介绍。

1.2.1 基于特征工程的传统机器学习方法

早期的谣言检测研究主要在谣言的文本内容、传播模式、用户信息等方面来提取能够判别出谣言真假性的特征，随后利用传统的机器学习方法（随机森林、决策树、支持向量机等）对谣言真假性的类别进行判断。Castillo 等人^[7]在 Twitter 数据集上从谣言文本、用户、主题、传播模式四个方面设计了评测推文可信性的相关特征，并采用决策树分类器确定推文内容的可信度水平。Qazvinian 等人^[8]探究了基于帖子内容、基于网络、基于特定帖子三种类别特征的有效性，构建了多个贝叶

斯分类器来对谣言进行预测，并有效地识别出帮助谣言传播的用户。Yang 等人^[9]采用了两种新的特征：基于客户端程序的特征和用户发布帖子的位置特征，并用支持向量机对谣言进行分类。Kwon 等人^[10]通过考察谣言传播的时间、结构和语言三个方面来识别谣言的特征，使用随机森林和逻辑回归模型来量化选取那些最具信息量的特征。Ma 等人^[11]提出了一种动态序列-时间结构模型，探究了各种社会背景特征随时间的变化，然后用线性支持向量机进行检测。Liu 等人^[12]在谣言的语言、用户、传播和其他元特征之上，使用大众的信念（beliefs of the crowd）和传统的调查性新闻特征将谣言识别为可能由一个或多个相互冲突的微博组成的事件。

上述早期传统的基于特征工程的机器学习方法虽然在谣言的自动检测上取得了前所未有的进步，但是以上研究方法过于依赖人工设计的特征。这些人工特征需要细致、精心地挑选，设计起来费时费力，并且极易忽略掉某些不易捕获的潜在谣言特征。此外，以上工作所用数据集过于单一，针对某一特定数据集所选取的特征并不具有良好的泛化能力和迁移能力。所用模型的参数量也较少，模型在表达能力上有所匮乏。而深度学习模型虽然在可解释性上有所欠缺，但却能捕获大量潜在的特征，同时深度网络模型也具有更强的表达能力。随着近年来深度学习在自然语言处理各领域取得的重大成功，谣言检测的相关研究已从传统的基于特征工程的方法演变为深度学习方法。

1.2.2 基于谣言文本内容和时间序列的深度学习方法

在上一子节中提到，传统人工提取特征的检测方法泛化能力差，特征的设计和选取费时费力。Ma 等人^[13]开创性地将深度模型循环神经网络（Recurrent Neural Networks, RNN）应用到谣言检测任务上。此方法将社交媒体中的谣言源帖子（source post）及其回复帖子（comments）建模成时间序列，而基于 RNN 的方法天然适合捕获长序列的文本特征表示，使模型同时学习了谣言的时间表征和文本表征。与基于特征工程的传统方法相比，该方法的检测效果获得了大幅提升。随后，各种基于深度学习的谣言检测方法开始不断涌现，Yu 等人^[14]指出基于 RNN 的方法存在一些不足，即不能很好地实现虚假信息的早期检测，并且对网络最新的输入存在偏差。为解决以上问题，此工作提出了一种基于卷积神经网络（Convolutional Neural Networks, CNN）的检测方法，可以灵活地提取分散在输入序列中的关键特征，并形成重要特征之间的高级交互，来有效识别虚假信息，实现谣言的早期检测。Liu 等人^[15]发现在谣言传播的早期，用户的回复和观点信息不足，谣言早期检测准确性很低，为此构建了一个时间序列分类器，该分类器结合了 RNN 和 CNN，分别捕获沿传播路径的用户特征的全局和局部变化，以进行谣言的早期检测。Guo

等人^[5]将谣言事件建模为包含不同语义级别信息的分层时间序列,然后将结构化的事件输入到分层的双向长短期记忆网络 (Long Short-Term Memory, LSTM),获得谣言的高层级特征表示。Khoo 等人^[16]提出了基于谣言文本和时序特征的经典方法 (Post-Level Attention Network, PLAN),模型在谣言帖子的时间序列上使用了具有自注意力 (self-attention) 机制的 Transformer 架构^[17]建模,捕获了任意两个帖子对之间的信息交互,与之前的研究方法相比取得了更好的检测效果。

1.2.3 基于谣言传播结构的深度学习方法

基于谣言文本和时序的深度学习方法与手工提取特征的方法相比虽然已有明显的性能提升,但是它们忽略了谣言在传播过程中潜在的结构化信息。尽管源帖子及其回复是按照时间顺序线性排列的,但根据帖子之间回复与被回复的关系,所有的帖子可以构建为一个完整的谣言传播图 (rumor propagation graph),如图1-3所示。Zubiaga 等人^[18]提出,帖子可以在不同用户分享观点、猜测和证据时“自我纠正”一些不准确的信息,这是因为回复帖子和回复关系中包含与事件话题相关的丰富的语义信息,对谣言真假性判断具有帮助作用。因此,近年来大量谣言检测的相关工作倾向于利用谣言的传播结构层级信息,捕获谣言的结构性的高层级特征。



图 1-3 来源于 Twitter 平台的一则谣言的传播图

Figure 1-3 A rumor propagation graph from Twitter platform

Ma 等人^[19]提出的递归神经网络 (Recursive Neural Network, RvNN) 是首次利用谣言传播的结构信息进行检测的工作。如图1-4所示,该方法基于谣言的双向传播树,分别建立了自底向上和自顶向下的两种递归神经网络,同时从帖子的文本和传播结构两方面获取谣言的特征表示。Khoo 等人^[16]在 PLAN 模型的基础上,显式地融合了谣言的传播结构信息,发展出了其变体模型 (Structure Aware Post-Level

Attention Network, StA-PLAN)。Bian 等人^[20]开创性地将具有强大表达能力的图神经网络 (Graph Neural Networks, GNN) 应用到谣言检测任务中, 该方法使用双向图卷积网络 (Bi-Directional Graph Convolutional Networks, BiGCN) 和根节点特征增强策略提取谣言特征。Zhang 等人^[21,22]将谣言的传播结构建模为一组独立的传播路径, 其中每个路径代表不同对话上下文中的源帖子, 然后将所有路径聚合以获得整个谣言传播图的表示。韩雪明等人^[23]提出了一种基于传播树的结点及路径双注意力谣言检测模型 DAN-Tree, 该模型使用 Transformer 架构学习传播路径中帖子间的隐式语义关系, 并利用注意力机制学习路径中节点的重要度。Wei 等人^[24]探究了谣言传播结构中的不确定性, 提出了一种边增强的贝叶斯图卷积网络 (Edge-enhanced Bayesian Graph Convolutional Network, EBGCN) 来捕获更具鲁棒性的谣言结构特征。Lin 等人^[25]将谣言传播结构建模为无向图, 提出了一种源帖子事件 (claim) 指导的层级化图注意力网络 (Claim-guided Hierarchical Graph Attention Network, ClaHi-GAT), 以充分利用源帖子中的信息, 同时在帖子级别 (post-level) 和事件级别 (event-level) 捕捉了多层次的谣言判别性特征。Ran 等人^[26,27]提出了一个多通道图注意力机制与事件共享模块的谣言检测模型, 融入了之前多源异构信息融合检测方法中所忽略的源推文-回复推文之间的传播结构, 构建了源推文-回复推文-词-用户多源关系异构图, 并用一个多通道图注意力网络充分学习三个通道子图的语义信息和传播模式。

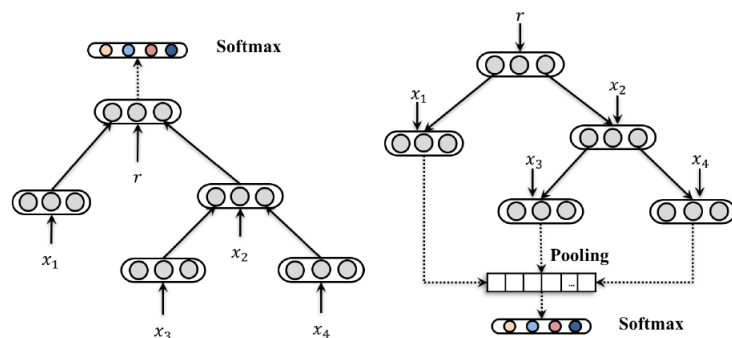


图 1-4 基于谣言传播结构的典型模型 RvNN^[19]

Figure 1-4 The typical model RvNN based on the structure of rumor propagation^[19]

1.3 现有研究存在的问题

近年来, 现有的各种深度谣言检测模型已经取得了良好的表现, 比如在四分类任务的 Twitter15 和 Twitter16 数据集上, 经典的 BiGCN 模型已能达到 88.6% 和 88.0% 的准确率, 在二分类任务的 Weibo 数据集上也实现了高达 96.1% 的准确率。然而, 现有的谣言检测相关研究还存在以下几个问题:

(1) 现有谣言数据集规模小。在谣言数据集的构建过程中, 由于国家对网络虚假信息的严格监控, 多数虚假信息一经网络平台的管理员发现就会被删除而无法收集, 这为谣言数据集的构建造成了极大的困难。此外, 对谣言进行标注也是极为复杂的一项工程, 通常对谣言数据进行标定需要在人力物力上进行大量的投资, 通过聘请谣言事件对应领域的相关专家, 并且收集各种相关背景的资料、官方的权威报告, 最后综合上述信息进行分析才能最终确定谣言的真假性类别标签^[6]。在谣言检测领域, 常用的公开数据集 Twitter15 和 Twitter16 仅有 1490 和 818 条谣言事件 (claim), 常用的 Weibo 也只有 4664 条谣言事件。而视觉领域著名的 ImageNet 数据集^[28] 有近 1400 万的数据样本。在实验过程中, 数量本就不多的谣言样本还要按照一定比例分成训练集、验证集和测试集。大量现有工作^[20,24,25] 甚至不再划分验证集, 匮乏的训练数据样本不可避免会造成模型过拟合、泛化性能差等问题。

(2) 现有模型泛化能力差、鲁棒性能不足。与近期发展迅速的大模型不同, 现有的谣言检测模型并没有巨大的参数量, 也没有利用大规模的谣言数据集进行训练。近年来, 大量的谣言检测工作根据谣言传播的特点、模式, 设计了复杂的模型结构来捕捉高层级的谣言特征^[20,24,25]。在谣言训练样本并不充足的情况下, 模型的鲁棒性和泛化性能不佳。如图1-5所示, 图1-5 (a) 是一则谣言实例的原始传播结构图, 现有模型能对谣言的真假性做出正确的预测, 然而, 谣言传播者通过将提供证据的帖子节点删除, 并添加了另一条迷惑性的评论, 恶意破坏了谣言传播结构, 致使模型的检测出现了错误, 如图1-5 (b) 所示。如何利用有限的有标注谣言数据来训练得到一个更具鲁棒性与泛化性的谣言检测模型, 具有重要的研究意义。

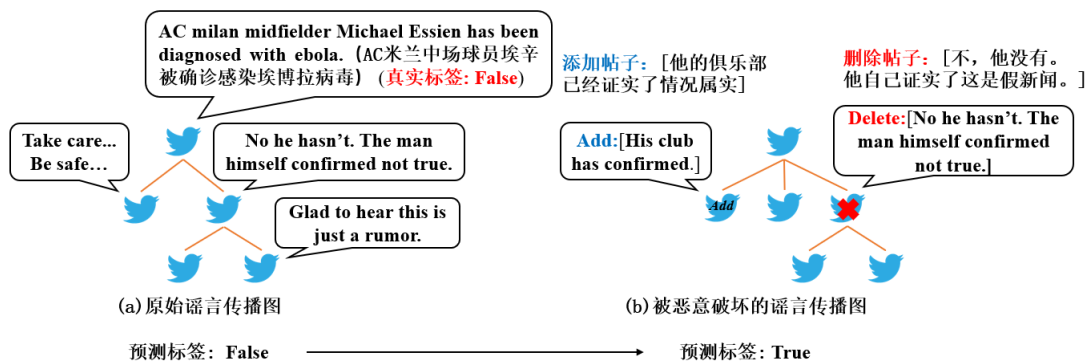


图 1-5 现有模型分类错误的实例

Figure 1-5 An instance of rumor detection model making mistakes

1.4 主要研究内容

为解决上一小节所提出的现有谣言检测研究中存在的问题，本文将围绕以下两个研究点展开：

(1) 基于图自监督对比学习的谣言检测方法 RD-GCSL。在上一节中提到，现有的谣言数据集规模小，针对谣言传播特点精心设计的模型存在过拟合的风险、模型的鲁棒性与泛化性欠佳。而自监督学习方法不依赖于数据的标签信息，通过建立代理任务，在大量的无标注数据中训练，学习到数据更本质、更具泛化性的特征。对比学习作为一种主流的自监督学习方法，通过数据增强，构建正负样本对并在特征空间中进行对比，使正样本靠近、负样本远离，提高了模型的鲁棒性与泛化性能。然而在谣言检测领域，自监督对比学习方法的应用还十分匮乏，其在谣言检测任务上是否有效仍有待研究。为此，本文将利用谣言的传播图结构，将谣言检测任务建模为图（graph）数据的分类问题。在第一个研究点中，RD-GCSL 模型建立了图自监督对比学习的辅助任务，并根据谣言的传播特点设计了三种谣言传播图的数据增强策略，将经过数据增强的谣言传播图输入到图编码器，随后将得到的图表示投影到特征空间计算对比损失，RD-GCSL 模型将有监督分类和自监督对比学习任务统一到一个框架下进行端到端的联合训练，使谣言图编码器的泛化能力和鲁棒性得到增强，并在一定程度上缓解了过拟合。在三个公开的数据集 Twitter15、Twitter16 和 PHEME^[29] 进行的实验表明，所提模型在完全监督和仅使用部分有标注数据的实验设置下，获得了比基线方法更高的准确率和 F1 值。验证了自监督对比学习方法应用在谣言检测任务上的有效性，并验证了所提模型不依赖于特定的谣言图编码器，能作为一个通用的框架提高现有谣言检测模型的性能。

(2) 针对半监督谣言检测的图自监督预训练辅助的消偏自训练方法 RDST。现有谣言检测的相关方法大多是在完全监督的实验设置下进行的。而有标注的谣言数据获得成本很高，一旦有标注训练数据不足，现有谣言检测模型具有较高的过拟合风险。为了减少对有标注数据的依赖，半监督学习（Semi-supervised Learning, SSL）技术利用少量的有标注数据和大量的无标注数据同时训练，在计算机视觉领域各任务的应用上已取得了巨大成功。然而目前还缺少相关研究将半监督学习这一天然适合于谣言检测任务情景的方法与谣言检测任务结合进行应用。为此，本文提出了一个针对半监督谣言检测的图自监督预训练辅助的消偏自训练方法 RDST。自训练是一种主流的半监督学习方法，它利用模型对无标注数据的预测结果为其指派伪标签，并选取可信样本作为有标注数据加入到下一次模型的训练中。通过自训练的迭代，模型在不断重复的训练中得到强化。然而自训练不可避免地会引入噪声，并将随迭代过程不断积累。为解决此问题，所提方法首先利用大量无标注谣言数据的传播图结构进行图自监督预训练，来增强图编码器的泛化性能，并

将其用作自训练的初始模型，以在自训练早期减少错误伪标签的生成。在伪标签选择阶段，此方法设计了自适应阈值的伪标注策略来进一步提高选取伪标注样本的质量。在 Twitter15、Twitter16、Weibo、DRWeibo^[30] 四个公开数据集进行的实验结果显示，所提模型在半监督设置下大幅超过了所有基线模型的表现，并在有标注数据极少的情景下仍保持了良好性能，验证了所提消偏自训练方法在半监督谣言检测任务上的有效性。

1.5 本文组织架构

本文共分为五个章节，各章节的组织架构安排如下：

第一章引言将首先对谣言检测任务的相关背景及研究意义进行介绍。随后将对国内外谣言检测研究工作的发展情况进行总结与分析，并指出现有研究中所存在的主要问题。最后将提出针对以上问题的解决方法，引出本文研究的主要内容并梳理全文的组织结构。

第二章将介绍与本文研究相关的基础理论，包括图神经网络、自监督学习、半监督学习、自训练等技术的原理及经典方法。最后将对本文研究所用的数据集和各类评价指标进行介绍。

第三章将介绍本文第一个研究内容：基于图自监督对比学习的谣言检测方法 RD-GCSL。本章是针对现有谣言检测模型存在的鲁棒性和泛化性能差的问题提出的第一个研究点。本章将对所提模型的各模块进行详细介绍，包括三种针对谣言传播图的数据增强策略、对比学习正负样本对的构建方法、图编码器和投影头的结构等。最后将在实验部分对比所提方法与基线方法的性能表现，在消融实验中探究模型各模块的作用，并在泛化性能实验中验证所提方法在鲁棒性、泛化性能上的提升。

第四章将介绍本文第二个研究内容：针对半监督谣言检测的图自监督预训练辅助的消偏自训练方法 RDST。本章是针对现有有标注谣言数据少、模型倾向于过拟合的问题所提出的第二个研究点。本章将对自训练方法的各部分模块及整体流程进行阐述，其中关于消偏的方法将重点介绍，包括图自监督预训练初始化、自适应阈值伪标注策略、分布对齐等。在实验部分将根据不同的数据规模对比所提方法与其他基线方法的性能表现，并设计相关实验展示所提消偏自训练方法与传统自训练方法相比在消除噪声上的效果，并对实验中重要的超参数进行定量分析。

第五章将对本文所提方法进行总结，分析本文所提方法的局限性，并对谣言检测领域在未来仍有价值进一步研究的方向进行展望。

2 相关理论及工作

本章将首先对本文主要研究方向：基于图表示学习的谣言检测方法的相关理论和现有的代表性工作进行简要介绍，随后将引出本文研究所使用的图自监督学习方法和半监督学习方法的相关理论和经典模型，最后将说明实验中所使用的数据集和采用的评价指标。

2.1 基于图表示学习的谣言检测方法

第一章1.2节已对现有谣言检测的各类相关工作做出了简要介绍，本文研究的重点是基于图表示学习的谣言检测方法，它是一种基于谣言传播结构的方法。自Bian 等人^[20]开创性地将谣言检测建模为图分类问题并将图神经网络这一强大的图表示学习方法进行应用后，各种基于图神经网络的谣言检测模型开始大量涌现。研究者们结合谣言传播特点，设计了各类不同结构的图神经网络来捕获谣言高层级特征，并取得了良好的表现。本节将对这一类方法的基础理论和具有代表性的工作及其经典模型进行详细介绍。

2.1.1 图神经网络

传统的深度学习模型在提取欧氏空间数据的特征上已取得了重大成功，然而现实中很多数据是非欧氏空间的，比如图数据。图的复杂性使现有的深度学习方法（如卷积操作）在图上很难直接进行。于是有大量研究者重新定义和设计了专门用于处理图数据的神经网络模型^[31-33]。图神经网络的核心原理是通过消息传递机制，迭代地聚合网络中邻居节点的信息，不断更新当前节点的特征表示。在这个过程中，每个节点的特征表示将包含其邻居和更远节点的信息。

具体的, 对于一个 K 层的 GNN, 第 k 层的更新公式为:

$$a_v^{(k)} = \text{AGGREGATE}^{(k)}(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}) \quad (2-1)$$

$$h_v^{(k)} = \text{COMBINE}^{(k)}(h_v^{(k-1)}, a_v^{(k)}) \quad (2-2)$$

其中, $\text{AGGREGATE}^{(k)}(\cdot)$ 和 $\text{COMBINE}^{(k)}(\cdot)$ 分别表示节点的聚合和拼接操作。 $h_v^{(k)}$ 是节点 v 在第 k 层的特征向量, $\mathcal{N}(v)$ 是节点 v 所有邻居的集合。在不同的图神经网络变体中, $\text{AGGREGATE}^{(k)}(\cdot)$ 和 $\text{COMBINE}^{(k)}(\cdot)$ 的具体操作方法有所差异。以经典方法 GraphSAGE^[34] 为例, AGGREGATE 聚合操作可表示为:

$$a_v^{(k)} = \text{MAX}(\{\text{ReLU}(W \cdot h_u^{(k-1)}), \forall u \in \mathcal{N}(v)\}) \quad (2-3)$$

其中 W 是一个可学习的矩阵, MAX 代表元素级的最大池化。GraphSAGE 中的 COMBINE 操作是一个线性映射 $W \cdot [h_v^{(k-1)}, a_v^{(k)}]$ 的拼接。如图2-1所示为 GraphSAGE 的消息传递过程示意图。

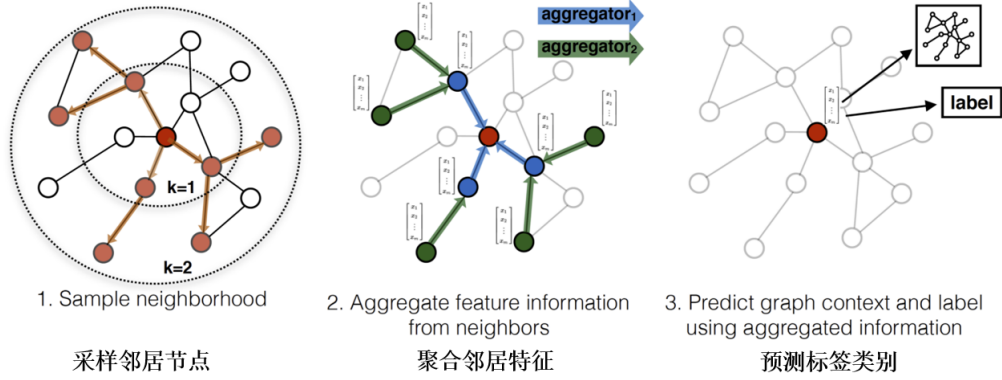


图 2-1 GraphSAGE 的工作原理示意图^[34]

Figure 2-1 The illustration of the working principle of GraphSAGE^[34]

对于节点分类任务, 网络最后一层节点特征表示 $h_v^{(K)}$ 可用来做最终的预测。而对于整图分类任务, 通常使用一个 READOUT 函数来聚合网络最后一层的节点特征, 得到整张图的表示 h_G :

$$h_G = \text{READOUT}(\{h_v^{(K)} \mid v \in G\}) \quad (2-4)$$

其中 READOUT 函数的选择可以为简单的变换函数如平均池化 (Mean-Pooling)、最大池化 (Max-Pooling)、求和池化 (Sum-Pooling) 或更复杂的图级池化函数。

2.1.2 文本表示学习方法

文本表示学习是自然语言处理任务中最基础和最重要的工作, 它的目标是将自然语言文本转换成计算机能够运算处理的数字或向量形式。在谣言检测任务中, 帖子中的文本内容是可利用的最重要信息之一, 如何准确、高效地获得文本特征表示将对后续检测的表现起到至关重要的作用。

现有谣言检测工作中所使用的文本表示学习方法主要分为以下三大类:

(1) 词袋模型 (Bag-of-Words, BOW), 它的基本思想是将整个文档视为一个无序的词汇集合, 完全忽略词汇出现的顺序或语法结构, 而只关注词汇出现的频率。具体而言, BOW 模型将每个文档用一个向量来表示, 向量的每一个元素代表词汇表中某个词汇在整个文档中的出现次数。词袋模型使用简单, 但是也存在丢失了词汇之间顺序和上下文信息的局限性。

(2) Word2Vec 模型^[35], 它是一种经典的词嵌入 (embedding) 方法, 即将文本中各个单词从原空间映射到新的多维空间中。它采用了无监督学习的方式, 从大量文本语料库中学习语义知识。Word2Vec 模型分为两种实现方式: 连续词袋模型 (Continuous Bag of Words, CBOW) 和跳字模型 (Skip-gram)。CBOW 模型的工作原理是根据某个词的上下文来预测这个词, 而 Skip-gram 模型的训练方式则是输入一个词, 预测上下文。Skip-gram 在学习低频词方面的效果更好, 但训练时间比 CBOW 更长。本文研究采用了 Skip-gram 的训练方式, 尽管其所需训练时间更长, 但是当模型第一次训练完成之后, 可以将模型参数保存, 在之后需要用到文本表示时直接加载模型参数进行使用。

(3) 预训练语言模型 BERT^[36] (Bidirectional Encoder Representations from Transformers), 它是近年提出的一种基于 Transformer^[17] 架构的双向编码器表示技术, 其效果表现刷新了 11 项自然语言处理任务的记录。然而在现有谣言检测的研究工作中, 许多使用 BERT 来提取文本表示的方法并没有带来性能的明显提升。究其原因, 考虑到 BERT 的预训练是在维基百科、图书语料库等大规模语料库上预训练得到。而在谣言检测任务情景下, 社交媒体中很多帖子的文本内容并不具有规范性, 通常包含用户的口头语、俚语、“文字梗”等非规范性语言, 这也就导致提取的文本表示不准确。

综合上述分析, 本文研究采用了前两种文本表示学习方法: BOW 和 Word2Vec 模型。

2.1.3 代表性工作及经典模型

在1.2节已对谣言检测相关研究的各类方法进行了梳理和介绍, 本小节将选取两个具有代表性的基于图表示学习的谣言检测研究工作: BiGCN^[20]、ClaHi-GAT^[25] 进行详细介绍。选取的两个经典模型也将作为本文研究的基模型在后续实验中进行对比。

(1) BiGCN 模型。如图2-2所示, 为 Bian 等人^[20] 在 2020 年提出的经典模型 BiGCN, 这是首个将谣言检测建模为图分类问题并应用图神经网络予以解决的工作。该方法将源帖子 (source post) 和回复 (replies) 关系所构成的谣言传播结构建模成双向的谣言传播图, 利用两个结构相同的图神经网络作为图编码器分别提取它们的图级表示, 随后将两种图表示拼接输入到全连接层做预测。此外, 为了强调源帖子信息的重要性, 此方法还融合了根节点增强模块, 即在每层图神经网络得到的隐层表示都要与前一层的根节点特征拼接, 再输入到下一层图神经网络。为了缓解过拟合, 模型还会将谣言传播图的连边以固定概率随机丢弃。实验结果显示, BiGCN 模型得益于强大的图神经网络, 其性能表现大幅超越了之前基于传

播结构的方法，达到了新的 SOTA 表现。在 Twitter15、Twitter16、Weibo 三个数据集上分别达到了 88.6%、88.0% 和 96.1% 的准确率。

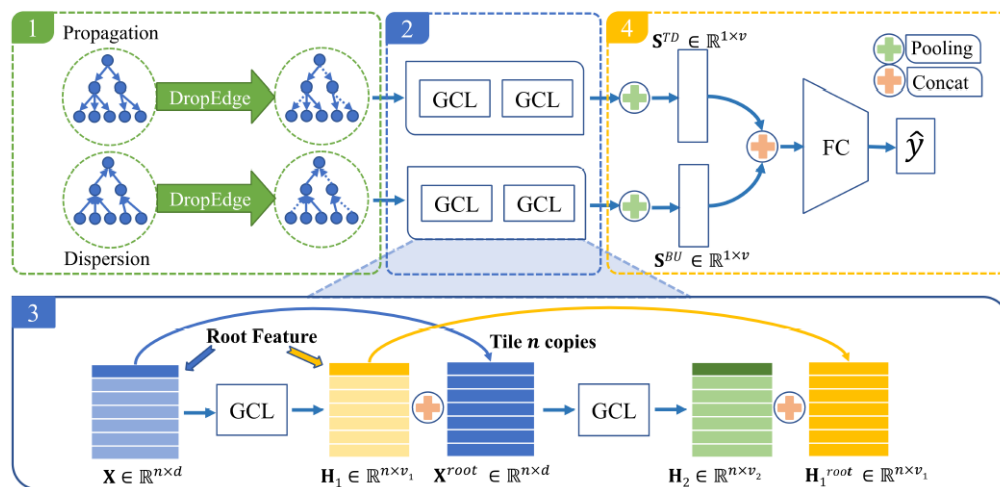


图 2-2 BiGCN 模型图^[20]

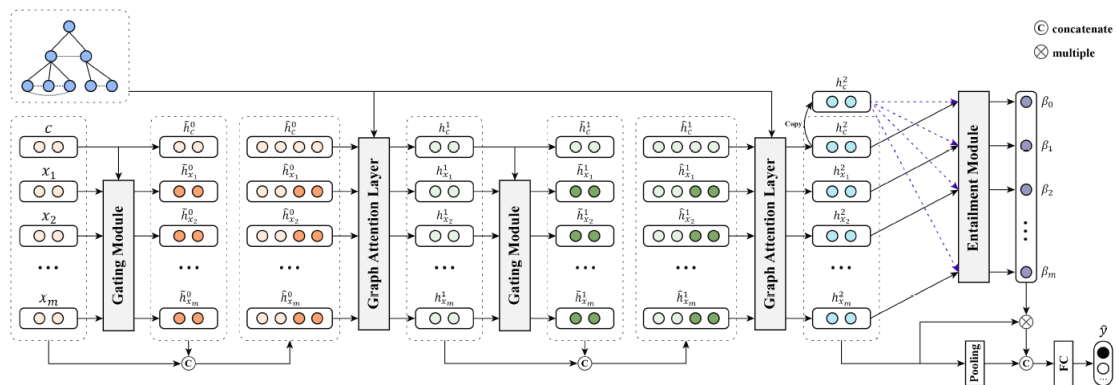
Figure 2-2 The model architecture of BiGCN^[20]

(2) ClaHi-GAT 模型。如图2-3所示，为 Lin 等人^[25]在 2021 年提出的 ClaHi-GAT 模型。该方法认为之前研究的谣言传播图结构过于简单，在兄弟节点之间也应存在信息的交互。因此，此方法在原始的谣言传播图基础之上将兄弟节点之间也添加了连边，并将节点之间所有的连边视为无向边。该方法选取了图注意力网络^[32] (Graph Attention Networks, GAT) 作为谣言传播图的图编码器，鉴于 GAT 在聚合邻居帖子时，能够为其指派不同的权重，而图卷积网络^[31] (Graph Convolutional Networks, GCN) 只能将所有邻居帖子分配相同的权重。在具体的网络结构设计上，模型设计了帖子级 (Post-level) 的注意力机制来控制相关帖子与源帖子信息的交互。在从节点表示获取全图表示的池化操作上，与 BiGCN 所使用的平均池化操作不同，此方法认为不同帖子节点在判断整个谣言的真假性上具有不同的重要性。为此，模型设计了事件级 (Event-level) 的注意力机制来为不同帖子节点分配不同的权重。实验结果显示，所提方法在 Twitter15、Twitter16、PHEME 三个数据集分别达到了 89.1%、90.8%、85.9% 的准确率，达到了新的 SOTA 表现。

2.2 图自监督学习方法

2.2.1 相关工作

自监督学习是一种利用预定义的代理任务从大规模的无监督数据中挖掘自身监督信息来进行学习的框架。通过这种构造的监督信息对网络进行训练，可以学

图 2-3 ClaHi-GAT 模型图^[25]Figure 2-3 The model architecture of ClaHi-GAT^[25]

习到对下游任务有价值的特征表示。自监督学习在计算机视觉和自然语言处理领域已有广泛的应用，在图表示学习的背景下，也已有研究者提出了大量的自监督学习方法，现有研究主要可分为对比式和生成式的自监督方法。对比学习是一种对比式模型，首先兴起于视觉领域。Chen 等人^[37]提出的 SimCLR 探究了对比学习不同的数据增强组合方式，提高了视觉表示的质量。He 等人^[38]提出的 Momentum Contrast 方法利用 Memory Bank 存储负样本，大大增加了对比学习中负样本的数量，缓解了显存不足的问题。Hjelm 等人^[39]提出 Deep Infomax (DIM) 来最大化一张图片的局部和全局上下文的互信息。随后，对比学习开始在图结构数据上被大量应用。Veličković 等人^[40]将 DIM 扩展到图领域，即最大化全局图表示与局部节点表示之间的互信息。Sun 等人^[41]进一步提出最大化图级表示与不同尺度的子结构表示之间的互信息。Hassani 等人^[42]通过建立多视角的对比来最大化不同视图的互信息。You 等人^[43]将经过数据增强的图级表示进行对比学习，提高了图编码器的鲁棒性和泛化能力。Hou 等人^[44]则是采用生成式的自监督方法，使用图自编码器重构掩盖节点特征，实现了与对比学习方法相当的性能表现。

2.2.2 图对比学习方法

图对比学习是一种主流的自监督学习方法，本小节将选取两类具有代表性的对比学习方法详细介绍其工作原理，包括基于图表示和节点表示互信息最大化的方法 InfoGraph^[41]和基于图级实例间 InfoNCE 对比损失^[45]的方法 GraphCL^[43]。本文研究也将把这两种重要的对比学习方法应用到谣言检测任务，并以此为基础做进一步改进与创新。

(1) 基于图和节点表示互信息最大化的方法 InfoGraph。如图2-4所示，其核心思想是通过最大化图级表示和局部子结构表示的互信息来获得更具表达能力的图

表示，具体地：对于图 $G = (V, E)$ ，使用参数为 ϕ 的 K 层 GNN 进行编码，通过拼接 GNN 在不同层级的特征向量表示，得到局部子结构（local patch）的表示 $h_\phi^v (v \in V)$ ：

$$h_\phi^v = \text{CONCAT} \left(\{h_v^{(k)}\}_{k=1}^K \right) \quad (2-5)$$

随后，对其进行 READOUT 池化操作就可以得到全局图表示 $H_\phi(G)$ ：

$$H_\phi(G) = \text{READOUT} \left(\{h_\phi^v\}_{v=1}^{|V|} \right) \quad (2-6)$$

其中 $|V|$ 是图 G 的节点数量。对于互信息的计算，使用定义在全局-局部对（global-local pairs）的 Jensen-Shannon 互信息估计器：

$$I_{\phi, \psi} (h_\phi^v (G); H_\phi(G)) = \mathbb{E}_{\mathbb{P}} [-sp(-T_\psi(h_\phi^v(G), H_\phi(G)))] \\ - \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}} [sp(T_\psi(h_\phi^v(G'), H_\phi(G)))] \quad (2-7)$$

其中 $\mathbb{P}$ 是训练样本在输入空间的经验概率分布， G 是采样自 $\mathbb{P}$ 的输入， G' 是采样自 $\tilde{\mathbb{P}} = \mathbb{P}$ 的一个负实例， T_ψ 是由 ψ 参数化的一个判别器（神经网络）， $sp(z) = \log(1 + e^z)$ 是 softplus 激活函数。在正负样本对的构建上，与全局图表示来源于同一个原图的所有节点表示被视为正样本，在同一个 batch 中来源于其他原图的所有节点表示被视为负样本，对比学习损失可以表示为：

$$\mathcal{L}_{css} = -\frac{1}{N} \sum_{G \in \mathbb{G}} \sum_{v \in V} I_{\phi, \psi} (h_\phi^v (G); H_\phi(G)) \quad (2-8)$$

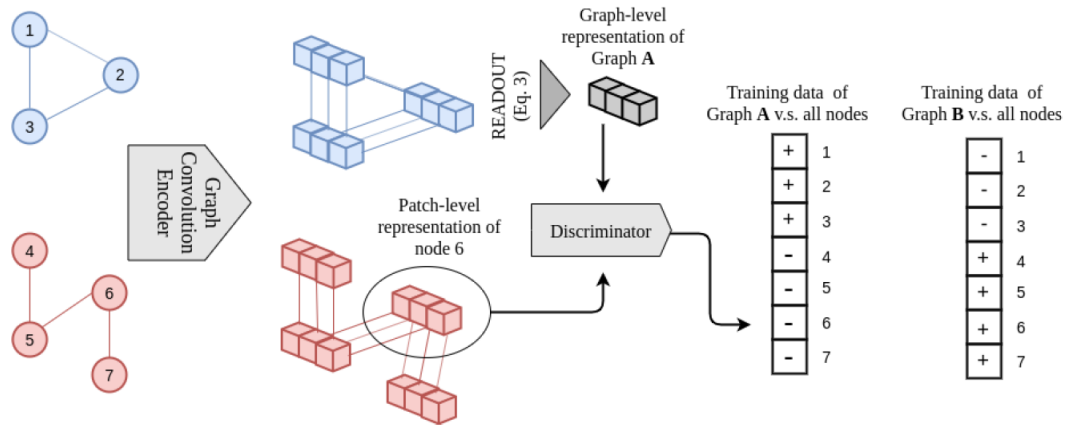


图 2-4 InfoGraph 模型图^[41]

Figure 2-4 The model architecture of InfoGraph^[41]

(2) 基于图级实例间 InfoNCE 对比损失的方法 GraphCL。如图2-5所示，其核心思想是通过对原图进行不同方式的数据增强，在特征空间中构建正负样本对，使图编码器捕获图的更本质的特征。具体地，对原图 G ，此方法设计了四种图级数

据增强方式：丢弃边、丢弃节点、特征掩盖、采样子图。通过图数据增强，得到两个互相关的视角 $\hat{\mathcal{G}}_i, \hat{\mathcal{G}}_j$ 。利用一个基于 GNN 的图编码器 $f(\cdot)$ 提取图 $\hat{\mathcal{G}}_i, \hat{\mathcal{G}}_j$ 的图级特征表示，得到 h_i 和 h_j 。随后图表示被投影头（多层感知机）投影到特征空间得到 z_i 和 z_j ，来进行对比损失的计算。

在自监督训练过程中，每个 minibatch 的 N 个图经过数据增强生成了 $2N$ 个扰动图，为了更清晰地表示对比损失的计算，在此重新将第 n 个图的投影表示 z_i, z_j 表示为 $z_{n,i}, z_{n,j}$ 。在对比学习正负样本对的构建上：来源于同一个原图的两个增强视图被当作正样本，同一个 minibatch 中其他的增强图都被视为负样本，采用 Oord 等人^[45] 提出的 InfoNCE 对比学习损失，对于第 n 个图，对比损失可定义为：

$$\ell_n = -\log \frac{\exp(\text{sim}(z_{n,i}, z_{n,j})/\tau)}{\sum_{n'=1, n' \neq n}^N \exp(\text{sim}(z_{n,i}, z_{n',j})/\tau)} \quad (2-9)$$

其中 τ 是温度系数， $\text{sim}(z_{n,i}, z_{n,j}) = z_{n,i}^\top z_{n,j} / \|z_{n,i}\| \|z_{n,j}\|$ 代表余弦相似度的计算。

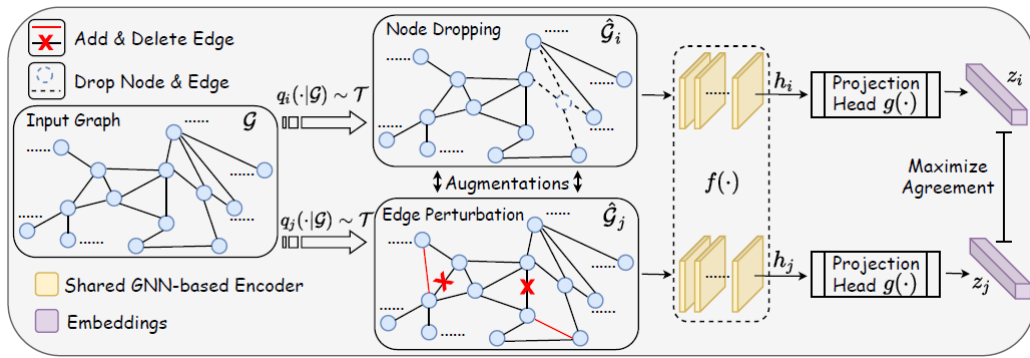


图 2-5 GraphCL 模型图^[43]

Figure 2-5 The model architecture of GraphCL^[43]

2.2.3 图自编码器方法

图自编码器（Graph Auto-encoder, GAE）是一种生成式的图自监督学习方法，它由一个编码器和一个解码器组成。GAE 通过编码并在重构损失的监督下重构输入数据来学习图表示。近年来采用掩码重构的自编码方法取得了与对比学习方法相当的性能表现。本小节选取 2022 年 Hou 等人^[44] 的代表性工作 GraphMAE 进行详细介绍。

对于以 X 为特征矩阵、以 A 为邻接矩阵的图 $G = (V, E)$ ，随机采样节点的一个子集 $\tilde{V} \subset V$ 并用 [MASK] 字符（token）来掩盖其特征，比如一个可学习的向量

$x_{[M]} \in \mathbb{R}^d$, d 表示特征的维度。于是, 节点 $v_i \in V$ 在掩码特征矩阵 $\tilde{X}$ 中的特征表示 $\tilde{x}_i$ 可被定义为:

$$\tilde{x}_i = \begin{cases} x_{[M]} & v_i \in \tilde{V}, \\ x_i & v_i \notin \tilde{V}. \end{cases} \quad (2-10)$$

随后, 掩码特征矩阵 $\tilde{X}$ 经图编码器 f_E 编码得到节点的隐层状态表示:

$$H = f_E(A, \tilde{X}) \quad (2-11)$$

为了进一步增强图编码器学习图表示的能力, 在对节点的隐层表示 H 解码前, 用一个特殊字符 [DMASK] 再次对已被掩盖的节点进行替换, 比如 $h_{[M]} \in \mathbb{R}^{d_h}$ 。随后, 重掩码的节点表示 $\tilde{h}_i$ 在重掩码特征矩阵 $\tilde{H} = \text{REMASK}(H)$ 中可以被表示为:

$$\tilde{h}_i = \begin{cases} h_{[M]} & v_i \in \tilde{V}, \\ h_i & v_i \notin \tilde{V}. \end{cases} \quad (2-12)$$

在解码过程中, 重掩码的节点隐层表示 $\tilde{h}_i$ 被解码来重构原始输入:

$$Z = f_D(A, \tilde{H}) \quad (2-13)$$

最后, 放缩余弦误差被用作损失函数来度量输入特征 X 和重构输出 Z 的重构误差:

$$\mathcal{L}_{sce} = \frac{1}{|\tilde{V}|} \sum_{v_i \in \tilde{V}} \left(1 - \frac{x_i^\top z_i}{\|x_i\| \cdot \|z_i\|} \right)^\eta, \eta \geq 1 \quad (2-14)$$

其中, η 是可调整的放缩参数因子。此处需注意, 在推理阶段, 原图直接输入到编码器中, 而不需要做任何掩码。此外, 解码器仅用于预训练阶段, 在推理阶段将会被直接丢弃。

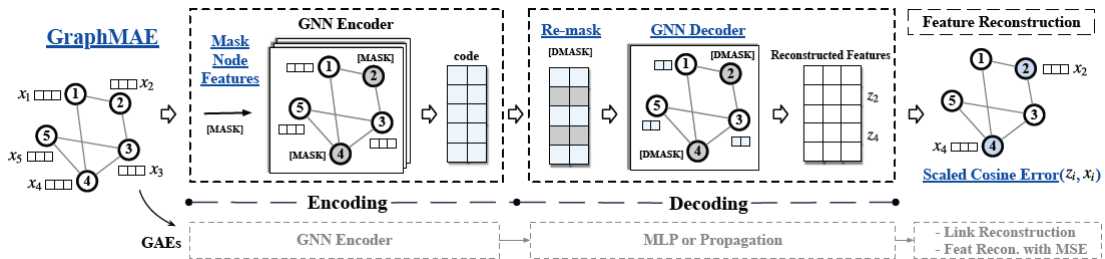


图 2-6 GraphMAE 模型图^[44]

Figure 2-6 The model architecture of GraphMAE^[44]

2.3 半监督学习方法

2.3.1 相关工作

现实中许多领域的有标注数据都难以大量获得,为了减少对有标注数据的依赖,研究者提出了同时利用大量无标注数据与少量有限的有标注数据联合训练的半监督学习 (Semi-supervised Learning, SSL) 方法。由于无标注数据容易获取, SSL 通常能以较低的成本实现性能提升。在计算机视觉领域已涌现了大量 SSL 的相关研究工作^[46-55],而在谣言检测领域,如前文1.3节所述,很难获得大规模的谣言数据集以在完全监督的设置下训练有效的分类模型。结合 SSL 方法,利用有限的有标注数据和大量无标注数据联合学习是谣言检测任务的一种自然场景,然而,目前仍缺乏以半监督方式进行谣言检测的相关研究。

现有主流的 SSL 方法大致可分为两类:一类是自训练方法 (self-training)^[46-50],也称为伪标注方法 (pseudo-labeling)。其核心思想是利用模型对无标注数据的预测结果给它们指派伪标签,然后利用这些伪标注样本迭代训练这个模型。Lee 等人^[46]提出选择具有最高预测概率的类别作为伪标签,并将其视为真实标签。Iscen 等人^[47]提出伪标签也可以基于邻域图分配给未标记的样本。尽管自训练简单高效,但 Arazo 等人^[48]认为其存在确定性偏置 (confirmation bias),即噪声积累的问题。为解决此问题,Rizve 等人^[50]提出了一种基于不确定性的伪标签标注方法,以提高伪标签的质量。Cascante-Bonilla 等人^[49]将课程学习的思想应用到伪标注策略中,以一种自定进度 (self-paced) 的方式来选取无标注样本。

另一大类是一致性正则化 (consistency regularization) 的方法^[51-56]。其核心思想是使模型对抗动的无标注的输入数据输出相似的预测结果。近些年来,基于一致性正则化的方法逐渐超过了基于自训练方法的表现。然而,基于一致性正则化的方法通常需要在有预先设定好的数据增强策略上进行^[50]。对于视觉模态数据,比如图像,各种数据增强策略已经被大量研究并应用。大多数基于一致性正则化的方法都需要对无标注样本进行弱数据增强(如裁剪、翻转)和强数据增强(如 Cutout^[57]、RandAugment^[58]、CTAugment^[59]、AutoAugment^[60])来生成扰动输入。然而对于图模态的数据,数据增强策略需要特别精心的设计,否则一致性正则化方法的有效性将得不到保证。

2.3.2 自训练方法

本小节将介绍现有半监督研究工作中典型的自训练方法,包括传统基于固定置信度阈值的自训练方法以及 2021 年 Cascante-Bonilla 等人^[49]提出的具有代表性

的基于课程学习^[61] (Curriculum-learning) 的自训练方法。

(1) 如图2-7所示, 传统的自训练方法在初始阶段先利用少量有限的有标注数据训练得到一个相对较弱的分类器, 作为自训练迭代的初始模型。在每轮自训练迭代中, 分类器为无标注数据预测来生成伪标签。在根据一个预先定义的固定阈值对可信的伪标签进行筛选后, 将这些样本加入到原始的有标注训练集。随后分类器由更新的有标注训练集重新训练, 这个迭代过程将不断重复直到没有更多可信的伪标注样本可被选择。

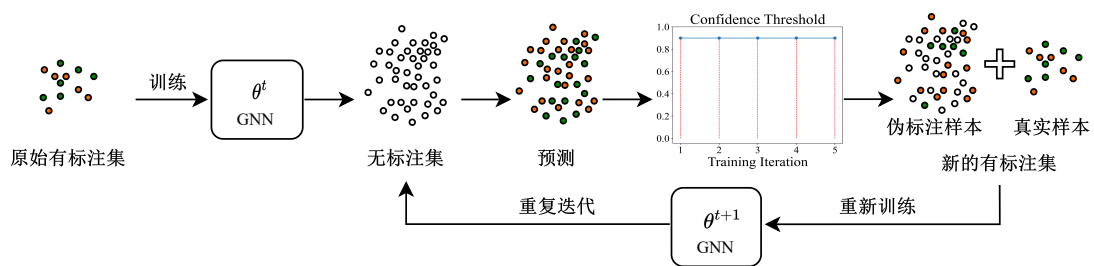


图 2-7 传统基于固定阈值的自训练方法示意图

Figure 2-7 Illustration of traditional self-training framework based on fixed thresholds

(2) 与传统使用固定阈值的伪标注方法不同, Cascante-Bonilla 等人^[49] 提出的基于课程学习的自训练方法使用了自定进度的动态阈值来选取可信的无标注样本。该方法在每轮伪标签选择的过程中首先将模型对所有无标注样本的预测概率分布按大小排序, 在第一轮选择预测概率最大的前 20% 的样本加入到原始的有标注训练集, 并重新训练模型, 随后每轮迭代以 20% 为步长, 依次增加选择的比例, 直到所有无标注数据加入到有标注训练集。使用这种课程学习方法的原因在于: 在自训练早期阶段, 模型的分类能力较弱, 这时只选择最有把握的伪标签, 随自训练进行, 模型趋于稳定, 分类能力逐渐增强, 逐渐增加选取的比例可以利用更多的无标注样本。

此外, 基于课程学习的自训练方法^[49] 还在自训练过程中使用了两个简单高效的技巧: 在每次将伪标注样本加入到有标注训练集后, 模型的参数都要初始化为最初的参数, 以防止误差的积累。其次, 每次根据动态阈值选取无标注样本时, 是从初始所有无标注样本中按相应百分比选择, 这使得之前被错误选择的伪标注样本在后续自训练迭代过程中有机会被纠正。

2.4 数据集与评价指标

2.4.1 Twitter15 和 Twitter16 数据集

Twitter15 和 Twitter16 两个数据集是 Ma 等人^[4] 将 2015 年和 2016 年构建的二分类 Twitter15^[12] 和 Twitter16^[13] 数据集进行改进而构建的四分类谣言数据集, 研究者收集了具有较高热度的源推文 (source tweets) 及其完整的传播结构, 即包括源推文的转发和回复 (retweets and replies) 信息, 并通过谣言揭穿网站 (如 snopes.com, Emergent.info) 利用谣言事件 (events) 的标签为每个源推文进行标定。每个谣言源推文包含四种类型的标签: non-rumor (非谣言), false rumor (经验证真实值为假的谣言), true rumor (经验证真实值为真的谣言), unverified rumor (未经验证的谣言)。Twitter15 和 Twitter16 两个数据集的四种类别标签对应的源推文具有大致相似的数量, 是类别平衡的数据集, 数据集的详细统计信息如表2-1所示:

表 2-1 Twitter15 和 Twitter16 数据集的统计信息

Table 2-1 Statistics of Twitter15&Twitter16 datasets

数据集	Twitter15	Twitter16
用户数	276663	173487
源推文数	1490	818
non-rumor (NR)	374	205
false rumor (FR)	370	205
true rumor (TR)	372	205
unverified rumor (UR)	374	203
事件平均时间长度	1337 小时	848 小时
事件平均帖子数	223	251
事件最多帖子数	1768	2765
事件最少帖子数	55	81

2.4.2 Weibo 和 DRWeibo 数据集

Weibo 数据集是 2016 年 Ma 等人^[3] 从新浪社区管理中心¹收集构建的中文数据集。对于每则谣言事件, 利用 Weibo 的 API 爬取了源帖子及其所有转发和回复的相关帖子, 同时研究者也爬取了相似数量的真实事件 (non-rumor events) 的源帖子和相关回复帖子, 整个 Weibo 数据集包含 2313 个谣言事件源帖子 and 2351 个真实事件源帖子, 是一个类别平衡的二分类数据集。

¹<http://service.account.weibo.com>

DRWeibo 数据集是 Cui 等人^[30] 在 2023 年构建的谣言中文数据集，收集了时间跨度从 2012 年到 2022 年期间微博平台的谣言和真实事件源帖子共 6037 条，包括 3185 条真实事件的源帖子和 2852 条虚假谣言事件的源帖子，同样，每个谣言事件包含源帖子及其相关回复帖子。由于近年来各大网络平台对虚假信息的管控更加严格，大量虚假信息已经被平台管理员删除，因此 DRWeibo 数据集虽然在时间跨度上更广，但谣言数据的样本数并没有与之相对应的大规模增加，相反，每则谣言事件具有更少的回复数量，与 Weibo 数据集高达 804 的平均回复数相比，DRWeibo 的源帖子平均回复数仅有 62，这也为基于传播结构的谣言检测方法带来了更多挑战。

如表2-2所示，为 Weibo 和 DRWeibo 数据集的详细统计信息：

表 2-2 Weibo 和 DRWeibo 数据集的统计信息

Table 2-2 Statistics of Weibo&DRWeibo datasets

数据集	Weibo	DRWeibo
语言	中文	中文
源帖子数	4664	6037
rumor (R)	2313	2852
non-rumor (NR)	2351	3185
事件平均帖子数	804	62

2.4.3 PHEME 数据集

PHEME 数据集是 Zubiaga 等人^[29] 在 Twitter 平台上收集的英文谣言数据集，包含与九个突发性新闻相关的源推文共 6425 条。这九个新闻事件分别为：《查理周刊》枪击事件、悉尼咖啡馆劫持事件、弗格森市民抗议活动、渥太华枪击事件、德国之翼飞机坠毁事件、普京失踪事件、歌手普林斯秘密演出事件、伯尔尼美术馆事件、球员埃辛感染埃博拉病毒谣言事件。PHEME 数据集的谣言类别划分方式与 Twitter15、Twitter16 数据集一致，但其各类别样本数量不平衡，属于非谣言（non-rumor）的源推文数量明显多于另外三个谣言类别，这种类别分布更符合现实中的谣言分布情景，即非谣言事件占多数，谣言事件占少数。这种类别分布不平衡的数据集也为谣言检测任务带来了新的挑战。

表 2-3 PHEME 数据集的统计信息
Table 2-3 Statistics of PHEME dataset

新闻事件	NR	FR	TR	UR	源推文数
Charlie Hebdo	1621	116	192	149	2079
Sydney Siege	699	86	382	54	1221
Ferguson	859	8	10	266	1143
Ottawa Shooting	420	72	329	69	890
Germanwings	231	111	94	33	469
Putin missing	112	9	0	117	238
Prince to play in Toronto	4	222	0	7	233
Gurlitt	77	0	59	2	138
Essien has Ebola	0	14	0	0	14
合计	4023	638	1067	697	6425

2.4.4 评价指标

为了直观地反映出各种谣言检测模型的性能优劣，需要选取相关的实验评价指标，本小节将对选取的实验评价指标进行介绍。通常，谣言检测被视为分类任务，因此常被选择的评价指标有：正确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值。

如表2-4所示，为二分类情况下分类结果的混淆矩阵，由此矩阵可定义如下评价指标：

表 2-4 分类结果混淆矩阵
Table 2-4 Confusion matrix of classification results

真实标签	预测结果	
	正例（positive）	反例（negative）
正例（positive）	真阳性（TP）	假阴性（FN）
反例（negative）	假阳性（FP）	真阴性（TN）

准确率（Accuracy）：准确率为所有评价指标中最为直观的评价标准，表示正确分类的样本所占总样本数量的比例：

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (2-15)$$

精确率与召回率：准确率只能反映所有类别整体的预测准确程度，而为了更清楚地显示出每个类别的预测情况，需要使用精确率和召回率来进行衡量。精确

率表示所有被预测为正例的样本中，实际也为正例的比例，其代表模型对某一个类别预测结果的准确程度，也称为查准率：

$$Precision = \frac{TP}{TP + FP} \quad (2-16)$$

召回率则表示所有实际为正例的样本中，被预测为正例的比例，其代表模型对某一个类别的识别能力，也称为查全率：

$$Recall = \frac{TP}{TP + FN} \quad (2-17)$$

通常，精确率和召回率并不能同时兼顾，精确率高代表模型更不倾向于“冒险”，只在很有把握时才将样本分为正例，而召回率高代表模型更倾向于“冒险”，会将很多负样本也分类为正样本。为了综合考量评价模型的有效性，需使用 F1 值作为评价指标，它是精确率和召回率的调和平均：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2-18)$$

2.5 本章小结

本章首先对本文主要研究方向：基于图表示学习的谣言检测方法的基础理论和两个具有代表性的工作 BiGCN 和 ClaHi-GAT 模型进行了简要介绍。随后介绍了本文研究所应用的图自监督学习的两大类方法：图对比学习和图自编码器，详细阐述了各自的工作原理并介绍了相关经典模型。之后介绍了本文研究所应用的另一关键技术：半监督学习方法。最后介绍了在实验中所使用的数据集和用于衡量各谣言检测模型性能表现优劣的评价指标。

3 基于图自监督对比学习的谣言检测方法

虽然现有的谣言检测模型在有标注数据充足的情况下能够有效对谣言真假性进行分类,但常用的谣言数据集规模小,有标注的谣言数据相对匮乏,现有模型往往面临过拟合的问题。与此同时,现有模型鲁棒性欠佳,当有谣言传播者采用各种手段蓄意破坏谣言原始传播结构,模型的分类结果将会出现错误。为解决以上问题,本章提出了一种基于图自监督对比学习的谣言检测模型(A Rumor Detection Model Based on Graph Contrastive Self-supervised Learning, RD-GCSL),结合谣言传播的特点设计了相应的图级数据增强策略来模拟对原图的扰动,根据所构建的正负样本对来计算对比学习损失。自监督对比任务和有监督任务端到端地联合训练,使模型学习到了谣言传播图更趋于本质的特征,在一定程度上缓解了过拟合,提升了模型的泛化性和鲁棒性。本章在三个常用的谣言公开数据集上进行了实验,验证了 RD-GCSL 模型的有效性。

3.1 研究动机

近年来,大量谣言检测相关研究结合谣言在传播过程中的结构信息,将其建模为图数据的分类问题,根据谣言传播的特点,以图神经网络为基础,精心设计了各种复杂的模型架构^[20,24,25]。通常在有标注谣言数据充足的情况下,这些模型能够有效地对谣言真假性分类并能获得较高的准确率。但有标注的谣言数据不易大量获得,如1.3节所述,常用的谣言数据集规模小,样本数量少,以上针对谣言特点精心设计的谣言检测模型存在过拟合的风险。此外,现有谣言检测模型的鲁棒性、泛化性欠佳,如图1-5所示,网络中恶意助长谣言传播的用户可能为虚假谣言信息提供支持的回复帖子,或将其他提供证据戳穿了虚假谣言的回复帖子删除。这种情况下,现有的谣言检测模型的分类结果将会出现错误。

自监督对比学习方法不依赖于数据的标注信息,通过预先定义的代理任务,从数据自身挖掘监督信息,将构建的正例样本和负例样本在特征空间中进行对比,使正样本相互靠近、负样本互相远离,学习到数据更本质的特征表示,提高了模型的泛化性与鲁棒性。在图表示学习领域中,已涌现了大量自监督对比学习的相关方法^[40,41,43],然而在谣言检测领域对其的研究和应用却依旧匮乏。为此,本章提出了基于图自监督对比学习的谣言检测模型 RD-GCSL,结合谣言传播特点设计了三种图的扰动方式,将经过扰动(数据增强)的两个扰动图输入到图编码器得到各自全图的特征表示,投影到特征空间后,通过判断两个扰动图是否来源同一原图

这一目标来建立自监督对比学习的辅助任务，并和有监督任务联合训练，以增强谣言图编码器的泛化性和鲁棒性，并缓解过拟合问题。

3.2 模型方法

如图3-1为所提 RD-GCSL 的具体模型架构，模型主要由五个部分组成：数据增强模块、图编码器、投影头、对比学习模块和谣言分类器。首先，将谣言源帖子和所有回复帖子根据它们的回复关系建模为谣言传播图，对每一张谣言传播图，进行两种不同的图级数据增强操作，随后将这些扰动的传播图输入以 GNN 为架构的图编码器来获取图的特征表示，并利用投影头将它们投影到特征空间，进行对比损失的计算。同时，在监督学习任务中，原图不经过数据增强而直接输入到图编码器和谣言分类器，计算真实标签和预测结果的交叉熵损失。整个模型的训练过程中，自监督对比学习任务和有监督分类任务端到端地联合训练。

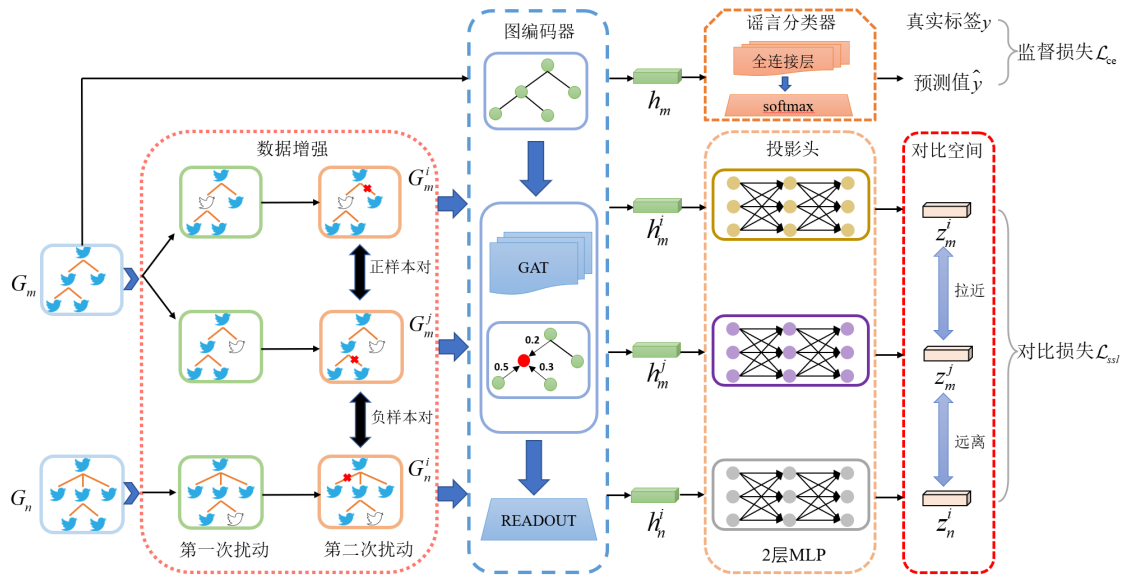


图 3-1 RD-GCSL 模型架构图

Figure 3-1 The model architecture of RD-GCSL

3.2.1 问题描述

对于一组谣言事件的集合 $C = \{C_1, C_2, \dots, C_n\}$ ，用 C_i 来表示集合中第 i 个谣言事件， n 代表集合中谣言事件的数量。 $C_i = \{r_i, x_1^i, x_2^i, \dots, x_{m-1}^i, G_i\}$ ，其中 r_i 为第 i 个谣言事件的源帖子 (source post)， x_j^i 为第 i 个谣言事件的第 j 条回复帖子， m 代表第 i 个谣言事件中所有帖子的数量。虽然一个谣言事件的所有帖子是

以线性序列排列的，但是根据帖子之间的回复关系，可以将整个谣言事件建模成一张具有传播关系的谣言传播图。用 $G_i = (V_i, E_i)$ 来表示第 i 个谣言事件的传播图， V_i 是以源帖子 r_i 为根节点的所有帖子节点的集合， E_i 是所有边的集合。举例来说，如果 x_2^i 是对 x_1^i 的回复帖子，则存在一个直接的连边 $x_1^i \rightarrow x_2^i$ 。 $\mathbf{X} \in \mathbb{R}^{m \times d}$ ， $A \in \{0, 1\}^{m \times m}$ 分别表示谣言传播图的特征矩阵和邻接矩阵。

本研究中谣言检测任务是学习一个分类器 $f: C_i \rightarrow Y_i$ ， Y_i 表示谣言的类别标签，根据不同的数据集，谣言的类别划分有所差异。

3.2.2 数据增强

数据增强的目的是在不改变数据原始语义的情况下，对原数据做一定程度的变换，生成新的可用数据以扩充样本数量。Wei 等人^[24]指出，谣言的传播结构通常具有不确定性，例如，一些谣言的制造者与传播者经常蓄意为虚假信息发布支持的帖子或删除反对其的帖子。另一方面，谣言传播图自身也存在一部分噪声信息。为增强谣言检测模型的鲁棒性与泛化性能，对谣言传播图进行数据增强，来模拟噪声扰动。具体地：对谣言事件的原始传播图 G ，进行两次扰动，得到两张扰动图 $\hat{G}_i, \hat{G}_j$ 。在现有的图表示学习的研究工作^[43]中，各种图级数据增强策略在图分类任务中已被证明简单有效。如图3-2所示，本章结合谣言传播的具体特点，设计了三种相应的图级数据增强策略：移除边（Edge Removing, ER）、移除节点（Node Dropping, ND）、掩盖节点特征（Feature Masking, FM）。

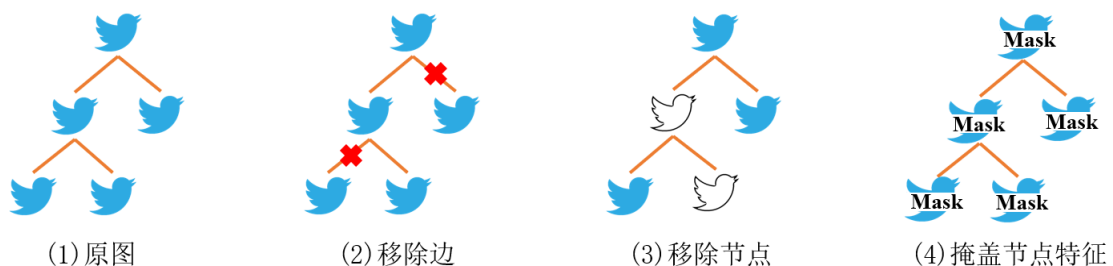


图 3-2 不同的图数据增强策略

Figure 3-2 Various graph augmentation strategies

第一种数据增强策略是移除边。在谣言传播图中，某个帖子及其回复帖子可能不存在直接的关联。例如，一些社交网络用户没有按照严格的回复关系，而是将其帖子随机地回复在谣言传播图中的任意帖子节点，造成了谣言传播图结构中边的不确定性。为建模这类情况，采用随机丢弃谣言传播图连边的数据增强策略，具体地：对以特征矩阵为 X ，邻接矩阵为 A 的谣言传播图 $G = (V, E)$ ，以概率 r 对原始边的集合随机采样并丢弃。

第二种数据增强策略是移除节点。在实际的谣言传播过程中，某些谣言制造者或恶意助长其传播的用户会蓄意为虚假信息提供支持的回复帖子，或将为戳穿虚假信息而提供了证据的帖子删除，以欺骗普通大众用户。此外，社交网络中的普通用户也可能随时将其发布的回复帖子删除，这造成了回复信息的缺失。为建模以上情况，采用随机丢弃谣言传播图节点的策略，具体地：以概率 r 对原始节点的集合随机采样，随后移除采样到的节点及其对应的连边。

第三种数据增强策略是掩盖节点特征。社交媒体平台中的大多数用户在发布的帖子中，所使用的语言通常不具有高度的规范性，往往存在噪声或歧义，例如，拼写错误、特殊字符、俚语等，使原始的语义信息具有一定噪声或偏置。为建模这类情况，采用随机掩盖节点特征的数据增强策略，具体地：以概率 r 对节点特征矩阵 X 的 d 个维度随机采样，并将特征矩阵 X 中所对应采样到的维度置零。

作为对比学习最关键的模块，数据增强策略的选择将直接影响对比学习的质量。对原数据所做的变换扰动过少会使对比学习任务过于简单，导致图编码器无法学习到谣言传播图的本质特征。对原数据所做的扰动过多，会导致有效信息丢失过多。为使对比学习的过程更加高效，采用上述两种不同数据增强方法的组合进行连续两次扰动。

3.2.3 图编码器

本章使用的图编码器以图神经网络为基础，其作用是提取输入图的特征表示。在图编码器具体架构的选择上，结合谣言传播的特点，对于某个谣言帖子，其所有回复帖子对其重要程度并不相同。图注意力网络^[32]在聚合邻居节点（回复帖子）信息时，为不同的邻居节点分配不同的权重，而图卷积网络^[31]在聚合时将所有邻居节点分配相同的权重。因此，为了提高帖子表示的质量，减少噪声信息的权重，采用 L 层的图注意力网络作为图编码器。 $H^{(l)} = [h_r^{(l)}, h_{x_1}^{(l)}, \dots, h_{x_m}^{(l)}]^T$ 表示网络第 l 层的节点隐层状态表示，其中 $H^{(0)} = X$ 。注意力系数的计算公式如式3-1所示：

$$\alpha_{i,j}^{(l)} = \frac{\exp\left(\phi\left(a^T \left[W^{(l)} h_{x_i}^{(l)} \parallel W^{(l)} h_{x_j}^{(l)}\right]\right)\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\phi\left(a^T \left[W^{(l)} h_{x_i}^{(l)} \parallel W^{(l)} h_{x_j}^{(l)}\right]\right)\right)} \quad (3-1)$$

其中， $\alpha_{i,j}^{(l)}$ 代表帖子 x_j 对帖子 x_i 的重要性， $\parallel$ 为拼接操作， a 和 $W^{(l)}$ 代表网络的权重参数， $\mathcal{N}_i$ 表示 x_i 帖子节点自身及其一阶邻居， ϕ 代表激活函数（如 LeakyReLU）。

节点的聚合更新公式如式3-2所示：

$$h_{x_i}^{(l+1)} = \text{ReLU}\left(\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(l)} W^{(l)} h_{x_j}^{(l)}\right) \quad (3-2)$$

随后，对 GAT 网络最后一层的节点表示进行平均池化，获得整个图的特征表示：

$$h = \text{Meanpooling} (H^{(L)}) \quad (3-3)$$

为进行自监督对比学习，分别将谣言传播图原图 G_m 和两个扰动图 $\hat{G}_m^i, \hat{G}_m^j$ 输入到参数共享的图注意力网络，得到对应的图级表示 $h_m \in \mathbb{R}^{d_1}$, $h_m^i \in \mathbb{R}^{d_1}$ 和 $h_m^j \in \mathbb{R}^{d_1}$ 。

3.2.4 对比学习损失

在进行对比学习之前需要利用投影头将图级表示映射到对比空间，投影头的结构是一个非线性变换 $g(\cdot)$ ，由两层感知机组成。具体地，为进行对比损失的计算，需将图编码器输出的两个扰动图的图级表示 h_m^i 和 h_m^j 投影到隐空间得到 $z_m^i \in \mathbb{R}^{d_2}$ 和 $z_m^j \in \mathbb{R}^{d_2}$ ：

$$z_m^i = g(h_m^i), z_m^j = g(h_m^j) \quad (3-4)$$

在每轮训练过程中，每个 minibatch 中的 N 个原图由数据增强生成了 $2N$ 个扰动图，选取一个扰动图的表示 z_m^i 作为锚节点，将与其来自同一个原图的扰动图的特征表示 z_m^j 视为正样本，除此之外的 $2N - 2$ 个扰动图的特征表示都视为负样本。通过最大化正样本的一致性（正样本相互靠近）、最小化负样本的一致性（负样本相互远离），建立自监督对比学习损失：

$$\mathcal{L}_{ssl} = -\lg \frac{\exp(z_m^i \cdot z_m^j / \tau)}{\exp(z_m^i \cdot z_m^j / \tau) + \sum_{\text{neg}} \exp(z_m^i \cdot z_{\text{neg}} / \tau)} \quad (3-5)$$

其中， τ 表示温度系数， z_{neg} 表示随机采样的负样本。

3.2.5 谣言分类器

将谣言原始图的特征表示 h_m 输入全连接层和 softmax 层：

$$\hat{y} = \text{softmax}(W_c h_m + b_c) \quad (3-6)$$

其中， $\hat{y} \in \mathbb{R}^{1 \times C}$ 是模型预测的各类别概率分布， C 表示谣言类别的数量， W_c 和 b_c 是可学习的参数矩阵。

根据谣言数据的真实标签信息，计算预测分布与真实类别分布的交叉熵，得到监督学习的分类损失：

$$\mathcal{L}_{ce} = - \sum_{i=1}^n y_i \lg(\hat{y}_i) \quad (3-7)$$

最终, 将有监督学习的分类损失和自监督对比学习损失相加作为 RD-GCSL 框架的总损失:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{ssl}} \quad (3-8)$$

其中, λ 表示自监督学习损失的权重超参数。

3.3 实验设计与结果分析

为验证所提 RD-GCSL 方法在谣言检测任务上的有效性, 本节将选取三个谣言公开数据集上展开实验。本小节将首先介绍所使用的数据集相关情况以及具体的参数设置, 并对实验所要对比的基线模型进行介绍, 随后将对各模型的实验结果进行分析, 并设计消融性实验验证所提模型各模块的有效性, 最后通过一个泛化性能验证实验来验证所提方法在鲁棒性、泛化性能上的提升以及对过拟合问题的缓解效果。

3.3.1 数据集与参数设置

为对本章所提模型的性能表现进行评估, 使用了来源于主流社交媒体平台的三个公开数据集 Twitter15、Twitter16 和 PHEME 进行实验。

Twitter15 和 Twitter16 是由 Ma 等人^[4]在 Twitter 社交媒体平台上爬取整理的谣言数据集。Twitter15 和 Twitter16 两个数据集中谣言各类别的数量相对均衡, 然而现实中虚假谣言的数量远少于真实事件的数量, 因此, 本章另选取了类别分布不平衡的 PHEME 数据集作为补充。PHEME 是一个在 Twitter 平台收集的与九个突发性新闻事件相关的英文谣言数据集。以上数据集的详细情况已在 2.4 一节进行了介绍。所有数据集的每则谣言事件都包含了源帖子和评论回复帖子, 因此每个谣言事件都可被建模为一张相应的谣言传播图。表 3-1 展示了所有数据集的详细统计信息。

实验环境使用 Pytorch¹, 在具体的参数设置上, 图神经网络的层数设置为 2, 采用 5000 维的词频 (Bag-of-words, BOW) 特征作为帖子的文本特征来初始化谣言传播图的节点特征矩阵, 网络中节点的隐层特征维度设置为 64, 图注意力网络中多头注意力的头数设置为 4, dropout 参数设置为 0.5 来缓解过拟合, 学习率设置为 0.0005, batchsize 设置为 256 (Twitter16 为 128), 两次数据扰动的比率在 $r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ 中通过网格搜索确定最佳组合, 对比损失中温度系数 τ 设置为 0.2, 自监督损失项权重 λ 设置为 1, 采用 Adam 优化器^[62]更新模型参数。

¹<https://pytorch.org>

表 3-1 数据集的统计信息
Table 3-1 Statistics of datasets

数据集	Twitter15	Twitter16	PHEME
谣言事件	1490	818	6425
非谣言 (NR)	374	205	4023
验证为假的谣言 (FR)	370	205	638
未经验证的谣言 (UR)	374	203	698
验证为真的谣言 (TR)	372	205	1067
用户数	276663	173487	48843
帖子数	331612	204820	197852

每次训练迭代 200 个 epochs，验证集的 loss 在 10 个 epochs 之内不再下降时采用早停机制。与 RvNN^[19]，BiGCN^[20] 等方法的实验设置一致，所有数据集按照 4: 1 的比例划分为训练集和测试集，采用 5 折交叉验证，并以不同的随机种子运行 10 次并汇报平均值，实验结果使用准确率 (Accuracy) 和各类 F1 值作为评价指标。

3.3.2 基线模型

本实验所对比使用的基线模型如下：

RvNN^[19]：是一个基于 GRU 单元和树结构递归神经网络的谣言检测模型，同时利用谣言帖子的文本信息和传播结构信息学习谣言的表示。

BiGCN^[20]：是一个基于图卷积神经网络的谣言检测模型，利用谣言传播的双向图，分别从自上而下和自下而上两部分提取谣言的高层特征，并使用根节点特征增强的策略来强调源帖子中的信息。

ClaHi-GAT^[25]：是一个基于图注意力网络的谣言检测模型，采用了层次化的注意力机制来充分利用源帖子中的信息。

RDEA^[63]：是一个以互信息最大化作为自监督目标的谣言检测框架，与本章所提 RD-GCSL 方法的区别在于，RDEA 采用的是自监督预训练后用有标注谣言数据微调的策略，而 RD-GCSL 是一种端到端的自监督谣言检测框架，即有监督任务和自监督任务联合训练。

SRD-PSID^[64]：是一个多视角对比学习的谣言检测框架，与所提 RD-GCSL 方法不同之处在于，SRD-PSID 没有在对比学习中显式设计数据增强策略，而是使用两个编码器（特征提取器）将传播路径与源码文本编码得到的两个表示作为两个不同视角进行对比。

UDGAT：是本章研究所使用的图编码器，以 GAT 为基础架构并将谣言图建模

为无向图，其与 BiGCN 模型相比，大幅减少了模型参数量。

RD-GCSL：是本章提出的图自监督对比学习谣言检测方法，以 UDGAT 作为图编码器来获取图的特征表示，建立了自监督对比学习的辅助任务，与有监督学习任务联合训练。

3.3.3 实验结果与分析

表3-2、表3-3和表3-4分别展示了 RD-GCSL 模型和其他基线方法在 Twitter15、Twitter16 和 PHEME 三个数据集上的性能表现。表中各指标的最优值已进行加粗显示。

从实验结果中发现，本章所提出的图自监督对比学习谣言检测方法 RD-GCSL 在 Twitter15、Twitter16 和 PHEME 三个数据集上分别达到了 88.0%，88.9%，85.6% 的准确率，为所有对比方法中的最优值，这表明 RD-GCSL 方法在谣言检测任务上与其他方法相比具有更大优势。

在所有基线方法中，RvNN 的精度与其它基于图神经网络的模型的表现相比有较大差距，这说明了图神经网络在捕获谣言传播图特征表示的高效性，以 GRU 单元和树结构递归神经网络为基础的架构利用谣言传播结构的效率欠佳。

UDGAT 模型以无向图作为输入，在模型结构上只使用了一个图注意力网络，与使用了两个图卷积网络的 BiGCN 模型相比，其参数量大幅减少，因此在实验过程中其占用显存空间与运行时间也随之减少，然而其性能表现与 BiGCN 模型相比却有略微提升。这说明了图注意力网络在消息传递时为邻居帖子分配不同权重的有效性，使用 UDGAT 作为所提对比学习框架的图编码器能够在保证精度的同时，降低模型在时间和空间上的复杂度。

在其它基于对比学习的谣言检测方法中，以图卷积网络为图编码器的 RDEA 方法在性能表现上超过了 BiGCN，证明了其所设计的基于互信息最大化的自监督对比学习方法的有效性。然而 RDEA 使用的自监督预训练再微调的方式与所提 RD-GCSL 方法端到端进行自监督任务和有监督任务联合训练的方式相比，性能表现仍有较大差距，并且在预训练阶段耗费了更多时间。基于多视角对比学习的方法 SRD-PSID 没有显式设计数据增强方法，而是分别利用两个特征提取器对谣言特征进行学习，在 PHEME 数据集达到了良好表现。

本章研究提出的 RD-GCSL 方法在之前研究的基础上，建立了一种端到端的自监督对比学习与有监督学习任务联合训练的框架，使图编码器学习到了谣言传播图更本质的特征，提高了模型的泛化性与鲁棒性，在 Twitter15、Twitter16 和 PHEME 数据集上与未使用对比学习的基模型 UDGAT 相比，分别提升了 3.4%，1.8%，1.2% 的准确率，验证了 RD-GCSL 方法的有效性。

表 3-2 Twitter15 数据集上的实验结果
Table 3-2 Experimental results on Twitter15 dataset

模型	准确率 (Acc.)	F1			
		NR	FR	TR	UR
RvNN	0.723±0.8%	0.682	0.758	0.821	0.654
BiGCN	0.843±0.4%	0.788	0.860	0.895	0.808
UDGAT	0.846±0.2%	0.792	0.849	0.906	0.829
ClaHi-GAT	0.859±0.4%	0.831	0.864	0.901	0.834
RDEA	0.855±0.6%	0.831	0.857	0.903	0.816
RD-GCSL	0.880±0.3%	0.851	0.886	0.926	0.852

表 3-3 Twitter16 数据集上的实验结果
Table 3-3 Experimental results on Twitter16 dataset

模型	准确率 (Acc.)	F1			
		NR	FR	TR	UR
RvNN	0.737±0.9%	0.662	0.743	0.835	0.708
BiGCN	0.858±0.5%	0.767	0.854	0.925	0.867
UDGAT	0.871±0.3%	0.794	0.876	0.927	0.870
ClaHi-GAT	0.882±0.4%	0.827	0.887	0.936	0.874
RDEA	0.880±0.5%	0.823	0.878	0.937	0.875
RD-GCSL	0.889±0.3%	0.833	0.882	0.949	0.886

表 3-4 PHEME 数据集上的实验结果
Table 3-4 Experimental results on PHEME dataset

模型	准确率 (Acc.)	F1			
		NR	FR	TR	UR
BiGCN	0.847±0.2%	0.910	0.634	0.655	0.500
UDGAT	0.844±0.2%	0.902	0.658	0.833	0.485
ClaHi-GAT	0.846±0.1%	0.896	0.670	0.623	0.515
SRD-PSID	0.838±0.3%	0.905	0.774	0.734	0.604
RD-GCSL	0.856±0.1%	0.915	0.669	0.607	0.530

为了进一步说明 RD-GCSL 能够缓解有标注数据不足带来的过拟合问题, 在仅使用部分有标注样本 (10%, 20%, 50%) 的实验设置下进行训练。如表3-5所示, 为 UDGAT 和 RD-GCSL 模型在不同的有标注数据规模下训练的实验结果, 表中 “ Δ ” 代表准确率的增益。从实验结果中发现, RD-GCSL 在全部三个数据集所有数据规

模下的准确率均超过了没有使用自监督对比学习的基模型 UDGAT，进一步验证了所提出的自监督对比学习方法在谣言检测任务上的有效性。

表 3-5 不同训练数据规模下的实验结果

Table 3-5 Experimental results with various scales of training data

数据集	模型	10%		20%		50%		80%	
		准确率	Δ	准确率	Δ	准确率	Δ	准确率	Δ
Twitter15	UDGAT	0.608	—	0.684	—	0.769	—	0.846	—
	RD-GCSL	0.626	$\uparrow 1.8\%$	0.702	$\uparrow 1.8\%$	0.803	$\uparrow 3.4\%$	0.880	$\uparrow 3.4\%$
Twitter16	UDGAT	0.594	—	0.723	—	0.820	—	0.871	—
	RD-GCSL	0.626	$\uparrow 3.2\%$	0.743	$\uparrow 2.0\%$	0.838	$\uparrow 1.8\%$	0.889	$\uparrow 1.8\%$
PHEME	UDGAT	0.738	—	0.766	—	0.797	—	0.844	—
	RD-GCSL	0.745	$\uparrow 0.7\%$	0.776	$\uparrow 1.0\%$	0.807	$\uparrow 1.0\%$	0.856	$\uparrow 1.2\%$

3.3.4 消融实验

(1) 不同谣言图编码器的影响。本章研究提出的 RD-GCSL 自监督对比学习谣言检测框架不依赖某一特定的谣言图编码器，能够作为一个通用的对比学习框架来提高现有谣言检测模型的性能表现。为了验证其对不同的图编码器结构普遍有效，使用三种谣言图编码器 UDGAT、BiGCN、ClaHi-GAT，结合本章所提出的自监督对比学习框架进行实验。如表3-6所示，用“-GCSL”代表使用了所提出的自监督对比学习的模型，“ Δ ”代表准确率的增益。

从表3-6中的实验结果来看，三种谣言图编码器在结合了本章所提出的自监督对比学习方法之后在三个数据集上的表现一致地获得了提升，这说明 RD-GCSL 框架不依赖于特定的图编码器，能够作为一个通用的对比学习框架来提升现有谣言检测模型的效果。

(2) 数据增强模块的影响。数据增强作为对比学习方法中最重要的模块，其扰动生成的样本质量将直接影响对比学习任务的效果。根据3.2.2节所设计三种图级的数据增强策略，可以构建样本多种扰动方式的组合。此外，每种增强策略中数据扰动的比例 r 也将影响对比学习的效果。为了探究不同数据增强方法以及扰动比例参数在 RD-GCSL 框架中的作用，设计以下实验：

1) 不同数据增强策略的影响：根据3.2.2节所设计三种数据增强策略，分别对原始谣言传播图进行单种方法扰动（移除边、移除节点、掩盖节点属性）、两种不同方法组合连续扰动、三种不同方法组合连续扰动生成扰动图。每种方法的扰动

表 3-6 不同图编码器的影响
Table 3-6 Effects of various graph encoder

模型	Twitter15		Twitter16		PHEME	
	准确率	Δ	准确率	Δ	准确率	Δ
UDGAT	0.846	—	0.871	—	0.844	—
UDGAT-GCSL	0.880	$\uparrow 3.4\%$	0.889	$\uparrow 1.8\%$	0.856	$\uparrow 1.2\%$
BiGCN	0.843	—	0.858	—	0.847	—
BiGCN-GCSL	0.881	$\uparrow 3.8\%$	0.888	$\uparrow 3.0\%$	0.850	$\uparrow 0.3\%$
ClaHi-GAT	0.859	—	0.882	—	0.846	—
ClaHi-GAT-GCSL	0.872	$\uparrow 1.3\%$	0.892	$\uparrow 1.0\%$	0.852	$\uparrow 0.6\%$

比例从 $r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ 中选取最优参数。表3-7展示了不同数据增强策略的影响，表中“ND”代表移除节点策略，“ER”代表移除边策略，“FM”代表掩盖节点属性策略，表中最优性能表现进行了加粗显示。

表 3-7 不同数据增强策略的影响
Table 3-7 Effects of various data augmentation strategies

数据增强策略	准确率 (Acc.)		
	Twitter15	Twitter16	PHEME
ND	0.871	0.883	0.851
ER	0.873	0.888	0.852
FM	0.869	0.887	0.855
ND+FM	0.880	0.888	0.856
ND+ER	0.875	0.889	0.853
ER+FM	0.873	0.887	0.855
ND+ER+FM	0.866	0.885	0.855

表3-7结果显示，不同的数据增强策略在三个数据集上的效果有所差异，但整体来看，移除边策略稍好于其他两种策略，采用两种不同增强方法连续扰动的方案效果略好于单种方法扰动和三种方法连续扰动的方案。由此可以推断，对比学习中数据扰动的方式不应过于简单，这会降低对比学习的效率，同时也不应过于复杂，因为对原始图进行过多扰动会造成有效信息的缺失。

2) 不同数据增强扰动比例 r 的影响：为探究数据增强不同的扰动比例对自监督对比学习效果的影响，在使用三种方法连续扰动策略（ND+ER+FM）的基础上，以不同的扰动比例 $r = \{0.1, 0.2, \dots, 0.8, 0.9\}$ 进行实验，实验结果如图3-3所示。

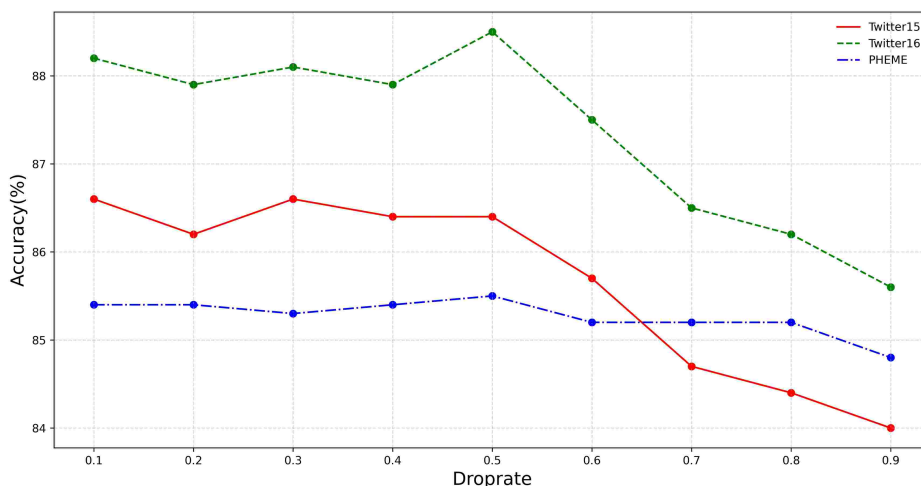


图 3-3 不同扰动比例的影响

Figure 3-3 Effect of various perturbation ratios

从图3-3的实验结果中可以发现，扰动比例分别为 0.3, 0.5, 0.5 时，模型在 Twitter15、Twitter16 和 PHEME 三个数据集上的表现达到最优。随着扰动比例继续增大，模型分类的准确率有明显下降趋势。这其中的原因在于，若对原图做过多扰动会引入过多噪声，使原图丢失掉大量有效信息，说明对比学习任务难度设置应适中，设置过难的对比学习任务并不会提升模型的性能表现。

(3) 投影头模块的影响。为验证 RD-GCSL 模型中投影头模块的作用，进行以下消融性实验，如表3-8所示，用 w/o PH (without projection head) 代表去掉投影头模块的模型。结果显示，在去除了投影头模块后，模型的性能表现有明显下降，在 Twitter15、Twitter16 和 PHEME 数据集上的准确率分别降低了 3.1%，2.3% 和 3.9%。以上结果验证了投影头模块的重要性，说明在对比学习过程中，由图编码器学习到的图级特征表示应经过投影头的非线性变换，在变换之后的隐空间中计算对比损失才能保证自监督对比学习的有效性。

(4) 泛化性能验证实验。为验证 RD-GCSL 在泛化性能上的提升以及对过拟合问题的缓解作用，设计如下泛化性能验证实验：利用3.2.2节所设计的三种数据增强策略对原始测试集中的谣言传播图以 $r = 0.1$ 的扰动比例进行两种不同类型的数据增强，所得扰动图的类标签与原图的类标签保持一致，分别将训练好的模型在原始测试集和扰动测试集上进行测试。表3-9展示了未使用对比学习的基模型 UDGAT 和自监督对比学习方法 RD-GCSL 在两种测试集上的性能表现，表中“ Δ ”代表准确率的增益。结果显示，在对原始测试集进行扰动之后，两种模型的检测准确率都有所下降，但 RD-GCSL 在扰动测试集上下降的精度明显小于未使用对比学习的基模型 UDGAT，证明了 RD-GCSL 得益于自监督对比学习任务的构建，表现出了较好的鲁棒性与泛化性能，缓解了过拟合问题。

表 3-8 投影头对模型的影响
Table 3-8 Effects of projection head

数据集	模型	准确率	F1			
			NR	FR	TR	UR
Twitter15	RD-GCSL	0.880	0.848	0.880	0.923	0.833
	w/o PH	0.849	0.822	0.841	0.901	0.817
Twitter16	RD-GCSL	0.889	0.833	0.882	0.949	0.886
	w/o PH	0.866	0.799	0.858	0.945	0.846
PHEME	RD-GCSL	0.856	0.915	0.662	0.639	0.510
	w/o PH	0.817	0.899	0.555	0.565	0.430

表 3-9 泛化性能验证实验
Table 3-9 Validation experiment of generalization performance

数据集	模型	准确率		
		原始测试集	扰动测试集	Δ
Twitter15	UDGAT	0.846	0.823	↓2.3%
	RD-GCSL	0.880	0.871	↓0.9%
Twitter16	UDGAT	0.871	0.835	↓3.6%
	RD-GCSL	0.889	0.878	↓1.1%
PHEME	UDGAT	0.844	0.830	↓1.4%
	RD-GCSL	0.856	0.849	↓0.7%

3.4 本章小结

本章针对目前谣言检测模型存在过拟合与鲁棒性不足的问题，提出了一种基于图自监督对比学习的谣言检测方法 RD-GCSL，结合谣言传播的特点设计了三种图级数据增强策略，定义了正负样本对的构建方式，建立了自监督对比学习的任务目标，并和有监督分类任务联合进行端到端地训练，使图编码器能捕获谣言更本质的图结构特征。在三个谣言公开数据集上进行的实验验证了 RD-GCSL 方法在谣言检测任务上的有效性，缓解了有标注数据匮乏造成的过拟合问题。

4 自监督预训练辅助的消偏自训练谣言检测方法

现有谣言检测模型在完全监督的实验设置下已达到了良好的性能表现，然而有标注的谣言数据难以大量获得，半监督学习方法同时利用有限的有标注数据和大量的无标注数据，以更低的成本实现了更好的性能表现，减少了对有标注数据的依赖。自训练作为一种常用的半监督学习方法，尽管使用起来简单有效，但是在迭代训练过程中不可避免地存在噪声积累。为解决以上问题，本章提出了一种自监督预训练辅助的消偏自训练谣言检测方法（A Debiased Self-training Method with Graph Self-supervised Pre-training Aided for Rumor Detection, RDST）。在 Weibo、DRWeibo、Twitter15、Twitter16 四个数据集上进行的半监督实验中，所提方法的性能表现大幅超越了其他基线方法，并在有标注数据极少的情况下获得了良好表现。

4.1 研究动机

尽管现有各种谣言检测方法^[16,24,25,64,65]已取得了良好的性能表现，但是现有研究的实验都是以完全有监督（fully-supervised）的设置进行的，即利用了全部的有标注数据进行训练。如1.3一节所述，谣言数据的标注过程需要耗费大量人力财力，对于一个谣言事件类别的标定，通常需要相关领域的专家在收集各种背景的证据、官方报告等资料后，展开综合的分析才能最终确定其真假类别。然而，无标注的谣言数据却能以较低成本大量获得，因此，利用少量有标注谣言数据和大量无标注谣言数据的组合进行半监督学习（semi-supervised learning, SSL）是谣言检测任务下的一种自然的场景。但是现在仍缺乏以半监督方式对谣言检测进行相关的工作。

现有主流的 SSL 方法可分为两大类，一种是自训练^[46,48-50]（self-training），也称为伪标注（pseudo-labeling）方法，其核心思想是利用模型对无标注数据的预测结果为其指派伪标签，并将这些伪标注样本加入到有标注训练集中，迭代训练模型。另一种是一致性正则化^[51-56]（consistency regularization）的方法，其核心思想是使模型对抗动的无标注输入样本输出相似的预测结果。近些年的研究工作中，基于一致性正则化的方法在表现上逐渐超过了基于自训练方法的表现，然而，基于一致性正则化的方法依赖于领域特定（domain-specific）的数据增强方式^[50]，比如大量研究中广泛使用的弱数据增强（weak augmentation）和强数据增强（strong augmentation）。在视觉领域，各种数据增强策略的发展与应用已比较成熟，但对于图数据来说，如何定义对应的强、弱数据增强策略以使得一致性正则化的方法有

效仍有待研究。相反，基于自训练的方法不需要进行数据增强，可以在不同领域中广泛使用。因此，本章采用自训练方法来进行半监督谣言检测的研究。

虽然自训练方法在使用上简单高效，但是它的工作机制存在一个重要的缺陷，即确定性偏置^[48] (confirmation bias)。这种偏置可以被视为对错误伪标注样本使用所造成的的误差积累：在错误伪标注样本上训练得到的模型必然会包含噪声偏置，因此会生成更多错误的预测，这造成了自训练迭代过程中噪声逐渐的累积，最终导致模型严重地退化。

针对上述问题，本章提出了一个图自监督预训练辅助的消偏自训练谣言检测方法。具体地，首先，为增强自训练初始模型的泛化性和鲁棒性，以在自训练早期生成更少错误的伪标签，利用大量无标注谣言数据的传播图结构，使用图自监督学习方法进行预训练，随后利用有标注数据微调，作为自训练框架的初始模型。在自训练过程中的伪标注样本选择阶段，为提高每轮自训练迭代所选取伪标注样本的质量，本章又设计了一种自适应阈值的伪标注策略，包括一个自定进度的全局阈值 (self-paced global threshold) 来控制整体伪标签利用的进度和一个局部阈值 (local class-specific threshold) 来关注每个类别具体的学习情况。后续4.2一节将对模型各部分模块进行详细说明。

4.2 模型方法

本节将详细介绍所提 RDST 消偏自训练谣言检测方法的各部分模块，整体架构如图4-1所示，首先利用谣言无标注数据的传播结构以图对比学习或者图自编码重构的方式来对模型（图编码器）做预训练，随后用有标注数据对预训练得到的模型做微调，将所得模型作为自训练的初始模型并对无标注数据进行预测来生成伪标签，在分布对齐模块对伪标签进行了校正之后，采用一种自适应阈值的伪标签选择方法来选取可信的伪标注样本加入到原始的有标注训练集，将模型参数重新初始化为预训练得到的模型参数，并用新的有标注集微调。整个过程迭代进行，直到没有更多可信的伪标注样本可以加入到有标注集合。

4.2.1 问题描述

为明确半监督设置下谣言检测任务的目标，采用以下公式化语言对本章所要解决的问题进行描述：

半监督设置下的谣言检测任务可按如下方式定义：利用一组由 N 个谣言训练样本组成的集合 $\mathcal{D}$ 来学习一个模型 $f_{\theta}(x)$ ，这些样本由一个有标注数据的集合 $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ (其中 $y_i \in \{0, 1\}^C$ 表示 C 个类别的 one-hot 标签) 和一个

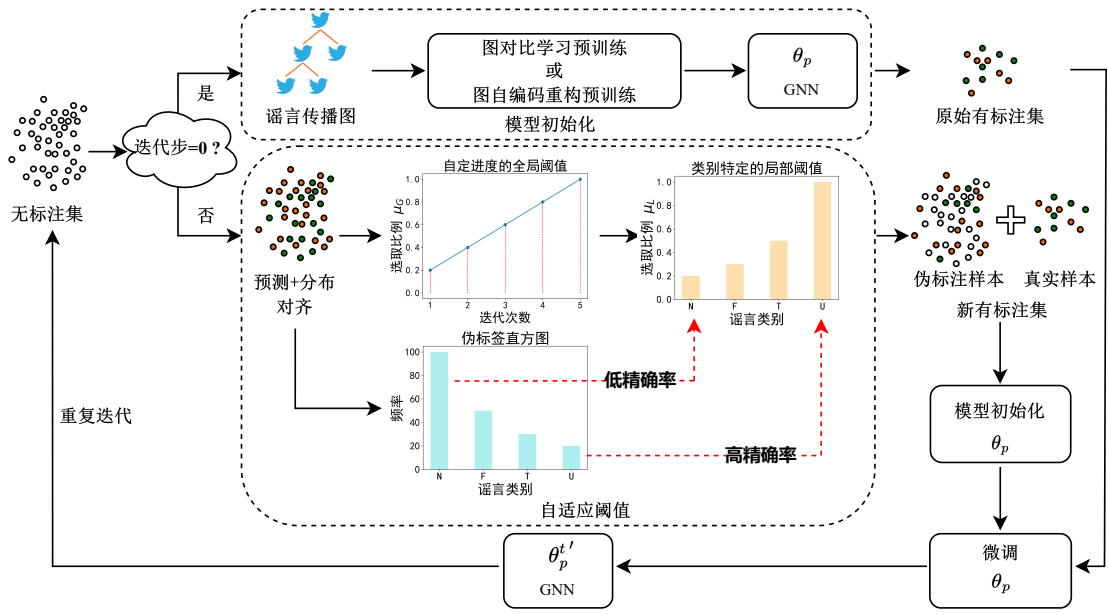


图 4-1 RDST 方法架构图

Figure 4-1 The architecture of RDST framework

无标注数据的集合 $\mathcal{D}_u = \{x_i\}_{i=1}^{N_u}$ 组成，因此有 $N = N_l + N_u$ 。鉴于本章采用自训练的方法，需要为 N_u 个无标注样本生成伪标签 $\tilde{y}$ ，于是任务目标可以进一步表达为利用 $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ （对于 N_l 个有标注样本， $\tilde{y} = y$ ）进行训练。在谣言检测的情景下， x_i 是第 i 个谣言事件的源帖子及其所有回复帖子的集合，即 $x_i = \{s_i, r_{i1}, r_{i2}, \dots, r_{im_i-1}\}$ ，其中 s_i 是源帖子， r_{ij} 是 x_i 的第 j 个回复帖子， m_i 是 x_i 中全部帖子的数量。

虽然一个谣言事件的全部帖子按照线性顺序排列，但根据它们的回复关系可以将所有帖子构建成一个谣言传播图。用 $G_i = (V_i, E_i)$ 代表第 i 个谣言事件的传播图，其中 V_i 代表以源帖子 s_i 为根节点的所有帖子节点的集合， E_i 代表 V_i 中与回复关系相对应所有无向边的集合。用 $X \in \mathbb{R}^{m \times d}$ 和 $A \in \{0, 1\}^{m \times m}$ 分别表示谣言传播图的特征矩阵和邻接矩阵， d 表示节点特征的维度。模型 $f_\theta(x)$ 是一个由 θ 参数化的图神经网络，任务的目标是优化模型使得其能够对谣言事件的真假性进行分类。根据不同的数据集，谣言的真假性有不同的分类方式。本章使用的四个公开谣言数据集按照以下两种方式进行划分：（1）二分类标签：谣言（rumor）或者非谣言（non-rumor），此情景下谣言检测任务仅需预测一个谣言事件是真或假。（2）四分类标签：非谣言（non-rumor），验证为真的谣言（true rumor），验证为假的谣言（false rumor），未经验证的谣言（unverified rumor）。在此类别划分方式下，“rumor”一词更倾向于人们熟知的“流言”，其表示一个被广泛传播讨论但未经验证真实性的事件。在此种更细粒度的分类方式下，谣言检测任务更具挑战性。

4.2.2 模型初始化

如前文2.3.2一节所介绍，传统的自训练方法通常先用有限的有标注数据训练来得到一个相对较弱的分类器，将其作为后续自训练迭代的初始模型。在随后各轮迭代中，这个分类器将对无标注数据进行预测来生成伪标签。根据一个预先定义的固定阈值筛选出可信的伪标注样本之后，将这些伪标注样本加入到原始的有标注训练集，并重新训练模型。整个过程迭代进行，直到达到期望的迭代次数或者没有更多可信的伪标注样本可被选择。

然而，利用少量有标注样本训练得到的初始模型不可避免地含有偏置，并且倾向于过拟合，这在之前工作中所设计的各种架构复杂的谣言检测模型^[20,24,25]中表现得极为明显。因此在自训练初始阶段，存在偏置的模型对无标注数据预测所生成的一大部分伪标签都是错误的，这些错误的伪标注样本在自训练后续迭代过程中被继续利用，导致偏置和误差逐渐累积，造成了严重的模型退化^[48,66]，这种现象通常被称为确定性偏置。

本小节的任务目标是在自训练的初始阶段减少误差的产生，具体地，本节希望增强自训练的初始模型的性能以在自训练早期生成更少错误的伪标签。受到近来自监督学习发展的启发^[41,67-70]，本节提出利用无标注谣言数据的传播结构来进行图自监督预训练，随之用原始的有标注训练数据微调，将得到的模型作为自训练的初始模型。采用这种方法的原因是：虽然无标注数据的谣言类别信息不可用，但是它们固有的谣言传播结构包含了丰富的上下文语义信息。在利用大量无标注数据进行自监督预训练后，图编码器能够提取谣言更本质的特征，其泛化性和鲁棒性得以增强。

本章研究采用了两种典型的图自监督学习方法：图对比学习和图自编码重构来增强图编码器的泛化性能。与之前设计了图数据增强策略的大量工作^[43,63,69]不同，本章采用的两种图自监督方法都不需要设计复杂的数据增强策略。接下来将介绍模型初始化模块的具体架构：

对于图编码器的选择，与之前方法所设计的复杂架构的 GNN 不同，本章使用一个朴素 (vanilla) 的 GNN 来提取谣言传播图的表示。它通过迭代聚合局部邻居信息来更新节点的表示。具体地，对于一个 K 层的 GNN，它的第 k 层节点表示为：

$$a_v^{(k)} = \text{AGGREGATE}^{(k)}(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}) \quad (4-1)$$

$$h_v^{(k)} = \text{COMBINE}^{(k)}(h_v^{(k-1)}, a_v^{(k)}) \quad (4-2)$$

其中， $h_v^{(k)}$ 是节点 v 在第 k 层的特征向量， $h_v^{(0)}$ 是节点的初始特征 X_v ， $\mathcal{N}(v)$ 是节点 v 所有邻居的集合。 $\text{AGGREGATE}^{(k)}(\cdot)$ 和 $\text{COMBINE}^{(k)}(\cdot)$ 为节点的聚合与拼接

操作，具体操作方法取决于所使用的 GNN 架构。在本章所提的框架中，对所使用的 GNN 架构没有特殊的限制，可以为 GCN^[31]，GAT^[32]，GIN^[33] 等。

对于图自监督预训练，采用以下两种方法：

(1) 图自监督对比学习预训练：图对比学习是图自监督学习中的一个主流方法，现有大多数对比学习方法都需要设计相应的数据增强策略^[43,63,65]。然而在不同的数据集中，最优的数据增强方式往往有所差异，在参数选择上也需要耗费大量时间来确定。为了避免数据增强方法设计与调参上的困难，采用基于图级表示和节点表示互信息最大化的图对比学习方法^[41]。

具体地，一组谣言传播图的集合可以表示为 $\mathbb{G} = \{G_1, G_2, \dots, G_N\}^1$ 。对于一个谣言传播图 $G = (V, E)$ ，将它输入到由 ϕ 参数化的 K 层 GNN 中，将节点在 GNN 不同层的表示拼接为一个特征向量作为局部子结构（local patch）表示 $h_\phi^v (v \in V)$ ：

$$h_\phi^v = \text{CONCAT} \left(\{h_v^{(k)}\}_{k=1}^K \right) \quad (4-3)$$

随后将节点表示进行池化（READOUT）操作即可得到全图的整体表示 $H_\phi(G)$ ：

$$H_\phi(G) = \text{READOUT} \left(\{h_\phi^v\}_{v=1}^{|V|} \right) \quad (4-4)$$

其中 $|V|$ 表示谣言传播图 G 中的节点数量，即此谣言事件所有相关的帖子数。

为进行互信息（mutual information, MI）的计算，使用定义在全局-局部对（global-local pairs）的 Jensen-Shannon 互信息计量器：

$$\begin{aligned} I_{\phi, \psi} (h_\phi^v (G); H_\phi(G)) = & \mathbb{E}_{\mathbb{P}} \left[-sp \left(-T_\psi (h_\phi^v (G), H_\phi(G)) \right) \right] \\ & - \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}} \left[sp \left(T_\psi (h_\phi^v (G'), H_\phi(G)) \right) \right] \end{aligned} \quad (4-5)$$

其中， $\mathbb{P}$ 是训练样本在输入空间上的经验概率分布， G 是采样自 $\mathbb{P}$ 的一个输入样本， G' 是采样自 $\tilde{\mathbb{P}} = \mathbb{P}$ 的一个负样本， T_ψ 是一个判别器（由 ψ 参数化的一个神经网络）， $sp(z) = \log(1 + e^z)$ 是 softplus 激活函数。

在对比学习正负样本的构建上，与全局图表示所属同一原图的所有节点表示被视为正样本，而所属其它图的所有节点表示被视为负样本，由此，建立以下自监督对比学习损失：

$$\mathcal{L}_{css} = -\frac{1}{N} \sum_{G \in \mathbb{G}} \sum_{v \in V} I_{\phi, \psi} (h_\phi^v (G); H_\phi(G)) \quad (4-6)$$

(2) 图掩码自编码重构预训练：图自编码器（Graph Auto-encoder, GAE）是一类生成式图自监督学习方法，通常由一个编码器和一个解码器构成。GAE 通过对输入的图数据进行编码和重构来学习图表示。本章采用近期提出的性能表现优异的掩码图自编码器^[44]进行谣言传播图自监督预训练。

¹实际上，有标注数据集 $\mathcal{D}_l$ 和无标注数据集 $\mathcal{D}_u$ 的谣言传播图都被用来进行预训练。

具体地，对于一个以 $X \in \mathbb{R}^{m \times d}$ 为特征矩阵、以 $A \in \{0, 1\}^{m \times m}$ 为邻接矩阵的谣言传播图 $G = (V, E)$ ，以随机概率对所有节点采样，得到节点的一个子集 $\tilde{V} \subset V$ ，使用一个 [MASK] 字符来掩盖其特征向量，比如一个可学习的向量 $x_{[M]} \in \mathbb{R}^d$ 。于是，对于任一节点 $v_i \in V$ 在进行掩码操作后得到的特征矩阵 $\tilde{X}$ 中的表示 $\tilde{x}_i$ 可被定义为：

$$\tilde{x}_i = \begin{cases} x_{[M]} & v_i \in \tilde{V}, \\ x_i & v_i \notin \tilde{V}. \end{cases} \quad (4-7)$$

随后，遮掩了部分节点特征的谣言传播图被图编码器 f_E 编码得到帖子节点的隐层表示：

$$H = f_E(A, \tilde{X}) \quad (4-8)$$

为进一步增强图编码器压缩和编码的能力，在对节点的隐层表示 H 解码前，用另一字符 [DMASK] 将已被遮掩特征的节点再次替换为 $h_{[M]} \in \mathbb{R}^{d_h}$ 。随后，重掩码的节点表示 $\tilde{h}_i$ 在重掩码特征矩阵 $\tilde{H} = \text{REMASK}(H)$ 中可以被表示为：

$$\tilde{h}_i = \begin{cases} h_{[M]} & v_i \in \tilde{V}, \\ h_i & v_i \notin \tilde{V}. \end{cases} \quad (4-9)$$

在解码器的选择上，与之前多数研究工作中使用多层感知机作为解码器架构不同，此方法使用单层 GNN 作为解码器，用 f_D 来表示解码器，将重掩码得到的特征矩阵 $\tilde{H}$ 输入到解码器来重构原始输入：

$$Z = f_D(A, \tilde{H}) \quad (4-10)$$

对于重构误差的衡量，采用放缩余弦误差作为输入特征 X 和重构特征 Z 的重构损失：

$$\mathcal{L}_{sce} = \frac{1}{|\tilde{V}|} \sum_{v_i \in \tilde{V}} \left(1 - \frac{x_i^\top z_i}{\|x_i\| \cdot \|z_i\|} \right)^\eta, \eta \geq 1 \quad (4-11)$$

其中， η 是在不同数据集中可调整的放缩参数因子。当自监督预训练完成后，在预测推理阶段原图将不做任何掩码直接输入到编码器中进行预测。同时，解码器在预训练完成后将直接丢弃，在推理阶段只使用图编码器来获取谣言传播图的图级表示，以供下游任务的使用。

图自监督预训练完成后，在微调阶段，用预训练得到的模型的参数 θ_p 来初始化图编码器，并用有标注集 $\mathcal{D}_l$ 中的训练样本进行微调。具体地，将有标注样本的谣言传播图输入到 GNN，在 GNN 最后一层获得了编码后的节点表示 h_v 后，使用一个 READOUT 函数聚合节点的表示来获得最终全图的特征表示：

$$h_G = \text{READOUT}(\{h_v\}_{v=1}^{|V|}) \quad (4-12)$$

随后，利用一个全连接层和一个 softmax 函数来对谣言真假性类别进行预测：

$$\hat{y} = \text{softmax}(FC(h_G)) \quad (4-13)$$

计算模型预测结果 $\hat{y}$ 和真实标签类别分布 y 的交叉熵损失：

$$\mathcal{L}_{\text{sup}}(\mathcal{D}_l) = - \sum_{i=1}^{N_l} y_i \log(\hat{y}_i) \quad (4-14)$$

最终，经过自监督预训练和微调后的图编码器，将用作 RDST 自训练框架的初始模型。

4.2.3 自适应阈值伪标签选择方法

自训练成功的关键在于伪标注样本的选择策略，因为利用错误的伪标签越少就会使自训练迭代造成的误差越小。本小节将提出一种自适应阈值的伪标注策略以在自训练过程每轮迭代中逐渐选取可信的伪标注样本。首先设定一个自定进度的全局阈值来从易到难逐渐地选择伪标注样本，为了关注每个类的学习情况，进一步引入一个类别特定的局部阈值，最终的阈值由两者共同决定。

(1) 自定进度的全局阈值：与之前研究^[53]中使用一个相对较高的固定阈值（如 0.95）不同，本章根据模型在每轮迭代对无标注样本的预测得分设定动态的全局阈值。受到课程学习^[49,54]（curriculum learning）思路的启发，采用模型对无标注样本预测概率分布最大的百分比来设定自定进度的全局阈值，即在自训练早期，只选取最有把握的伪标注样本，随着自训练进行，模型分类能力得到增强，也愈趋向稳定，增大伪标注样本选择的比例以利用更多的无标注数据。具体地，对于第 t 轮自训练迭代，全局阈值 τ_G 可定义为：

$$\tau_G(t) = \text{Percentile}(\mu_G(t)) \quad (4-15)$$

$$\mu_G(t) = \gamma \cdot t, t = 1, 2, \dots, T \quad (4-16)$$

其中 μ_G 是伪标注样本的选取比例， γ 是其每轮迭代增长的步长，于是有 $T = 1/\gamma$ 代表自训练迭代的总次数， $\text{Percentile}(\cdot)$ 是模型对无标注数据预测概率最高的前百分之 μ_G 所对应概率值的映射。注意在自训练的每轮伪标签选择步骤中，伪标注样本都是从整个无标注数据集中选择的，这使得之前错选或者漏选的伪标注样本在下一轮选择步骤中可以被纠正。

(2) 类别特定的局部阈值：全局阈值是控制所有类别伪标注样本总体利用进度的一种策略，然而它没有考虑到模型对各类别不同的学习情况。经初步实验发现，

如图4-2所示, 尽管初始的有标注和无标注样本集合是类别平衡的, 但随着自训练进行, 模型倾向于将所有无标注样本预测为某一类或某几个特定类。如果不针对各类伪标注样本的选择进行干预, 伪标注样本的类别会变得更加不平衡。以上发现与之前的研究工作^[56]所指出的问题相似, 即一个模型虽然是在类别平衡的数据集上训练与测试的, 由于数据内在的相似性, 伪标签的类别也是天然不平衡的。

为了对被预测为各类的伪标注样本的准确性进行全面分析, 在此利用了无标注数据的真实标签 (仅用作正式实验前的分析)。如图4-2为在 Twitter15 数据集每个类使用 40 个有标注样本训练的模型对无标注样本进行预测并根据全局阈值选取的伪标签质量分析结果, 从图中可以发现, 被预测为最多的类别具有最高的召回率, 但其精确率很低, 然而被预测为最少的类别具有更高的精确率、更低的召回率。这表明被预测为最少的类别的这类伪标签更有可能是准确的伪标签。受 Wei 等人^[71] 在类别不平衡自训练框架中伪标注样本选择策略的启发, 本章根据一个类别再平衡 (class-rebalancing) 的原则来设定类别特定的局部阈值: 若某个类别 c 被预测的频率越低, 则选取越多的被预测为类别 c 的伪标注样本加入到有标注样本集, 反之同理。

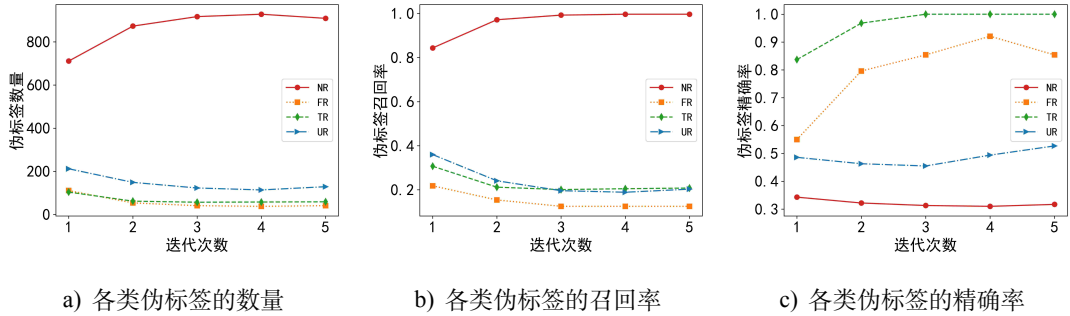


图 4-2 伪标签质量分析

Figure 4-2 Analysis of pseudo-labels quality

具体地, 首先统计在自训练第 t 轮迭代每个类别伪标签的数量:

$$\sigma_t(c) = \sum_{n=1}^{N_u} \mathbb{1} \left(\arg \max (p_{m,t}(y | u_n) = c) \right), c = 1, 2, \dots, C \quad (4-17)$$

其中 $\mathbb{1}$ 是指示函数 (indicator function), $p_{m,t}$ 是模型在第 t 轮迭代的输出概率。

随后按降序顺序对 $\sigma_t(c)$ 排序, 得到 $N_1 \geq N_2 \geq \dots \geq N_C$ 。被预测为类别 c 的伪标注样本将按照如下比例进行选取:

$$\mu_L(c) = \left(\frac{N_{C+1-c}}{N_1} \right)^\alpha \quad (4-18)$$

其中 $\alpha \geq 0$ 是调整选取比例的超参数。举例来说, 对于一个四分类的类别不平衡的伪标注样本集, 最多类和最少类数量的比率 $\delta = \frac{N_1}{N_4} = 10$, 最少类别的选取比例为

$\mu_L(4) = \left(\frac{N_{4+1}-4}{N_1}\right)^\alpha = 1$, 最多类别的选取比例为 $\mu_L(1) = \left(\frac{N_{4+1}-1}{N_1}\right)^\alpha = 0.1^\alpha$ 。当 $\alpha = 0$, 对于所有类别有 $\mu_L(c) = 1$, 即所有伪标注样本都被选取, 此时局部阈值的作用失效。在为每个类选取伪标注样本时, 根据置信度的高低, 优先选取具有更高置信度的样本。

最终的选取阈值 τ_t 由自定进度的全局阈值和类别特定的局部阈值共同决定:

$$\tau_t(c) = \text{Percentile}(\mu_t(c)) \quad (4-19)$$

$$\mu_t(c) = \mu_G(t) \cdot \mu_L(c) \quad (4-20)$$

其中 $\mu_t(c)$ 是被预测类别为 c 的伪标注样本在第 t 轮迭代的选取比例。最终, 第 t 轮迭代的无监督损失可表示为:

$$\mathcal{L}_u = \frac{1}{N_u} \sum_{n=1}^{N_u} \mathbb{1}(\max(q_n) \geq \tau_t(\hat{q}_n)) \cdot \mathcal{H}(\hat{q}_n, q_n) \quad (4-21)$$

其中 $q_n = p_{m,t}(y | u_n)$ 为模型对无标注数据的预测结果, $\hat{q}_n = \arg \max(q_n)$ 是伪标签, $\mathcal{H}(\cdot, \cdot)$ 代表交叉熵。

为了进一步提升伪标签的质量, RDST 方法进一步融入了一个分布对齐 (distribution alignment, DA) 模块。分布对齐策略是由 Berthelot 等人^[59] 首次提出, 其核心思想是将模型对无标注样本的预测分布与有标注样本的类别分布 $p(y)$ 对齐。具体地, 它通过将模型对某一无标注样本 u_n 的预测结果 $q_n = p_m(y | u_n)$ 乘以 $p(y) / \tilde{p}(y)$, 其中 $\tilde{p}(y)$ 是模型对所有无标注样本预测结果的移动平均值。在 RDST 方法中, 本节对其稍作修改, 加入了一个调节分布对齐强度的指数参数 β ($\beta \geq 0$), 最后再进行归一化来得到有效的概率分布值:

$$\tilde{q}_n = \text{Normalize} \left(q_n \left(\frac{p(y)}{\tilde{p}(y)} \right)^\beta \right) \quad (4-22)$$

$\tilde{q}_n$ 将被用作 u_n 的预测标签来替代 q_n 。需注意, 分布对齐只在伪标注阶段应用以校正伪标签, 一旦伪标注阶段结束, 分布对齐模块在随后的训练过程中将不会再被使用。

4.2.4 训练策略

在伪标注样本的选择阶段完成后, RDST 方法采用了一种新的训练策略来更高效地利用原始有标注样本集和伪标注样本集的组合: 所有原始的有标注样本将被平均分配到新的有标注集的每个 mini-batch 中。以这种方式, 每个 mini-batch 中

都能确保最少有一部分样本具有正确的标签, 进一步减少了误差的产生。对于初始有标注样本集合, 其有监督分类损失可表示为:

$$\mathcal{L}_s = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathcal{H}(y_n, p_{m,t}(y | x_n)) \quad (4-23)$$

此外, 当每轮自训练迭代结束后, 模型参数将重新初始化为自监督预训练得到的参数 θ_p , 以确保模型在后续迭代过程不会对之前可能存在偏置的模型产生噪声积累。

为了解决自训练早期分类能力较弱的模型会盲目将所有无标注样本预测为某一个类或某几个特定类的问题, 本节采用了一个在之前研究^[48]中广泛使用的公平性正则化项, 以鼓励模型以相同的频率为每个类做预测:

$$\mathcal{L}_f = \sum_{c=1}^C p_c \log \left(\frac{p_c}{\bar{h}_c} \right) \quad (4-24)$$

其中 p_c 是类别 c 的先验概率分布, $\bar{h}_c$ 是模型对被预测为类别 c 的所有样本的预测概率的平均值。鉴于本章所使用的数据集都是类别平衡的数据集, 采用均匀分布 $p_c = 1/C$ 作为每个类别的先验概率分布。

最终, RDST 总体的目标函数可以表示为:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_f \mathcal{L}_f \quad (4-25)$$

其中 λ_u 和 λ_f 分别是 $\mathcal{L}_u$ 和 $\mathcal{L}_f$ 两项损失的权重超参数。

4.3 实验设计与结果分析

为了验证本章所提自训练方法 RDST 在半监督谣言检测任务上的有效性, 本节在四个公开谣言数据集上进行了对比实验。本节将首先介绍相关数据集的详细信息, 随后将对选取的基线模型进行简要描述, 再对实验具体的参数设置进行说明, 之后将对各模型的实验结果进行分析和讨论, 并通过消融实验验证模型各部分模块的有效性, 最后对实验中重要的参数进行定量分析。

4.3.1 数据集

本章在四个公开的谣言数据集上进行了实验, 包括 Weibo, DRWeibo, Twitter15, Twitter16。Weibo 和 DRWeibo 是中文的二分类数据集, Twitter15 和 Twitter16 是英文的四分类数据集, 以上数据集的详细情况在2.4一节已进行介绍。需注意, 本研究重点关注于类别平衡条件下的半监督谣言检测任务, 因此所使用的数据集在每个

类别有大致相似的样本数量。具体的所有数据集统计信息如表4-1所示（表中 Weibo 和 DRWeibo 两个数据集统一将 R 类视为 FR 类）。

表 4-1 数据集的统计信息
Table 4-1 Statistics of datasets

数据集	Weibo	DRWeibo	Twitter15	Twitter16
语言	中文	中文	英文	英文
事件数	4664	6037	1490	818
非谣言 (NR)	2351	3185	374	205
验证为假的谣言 (FR)	2313	2852	370	205
验证为真的谣言 (TR)	-	-	372	207
未经验证的谣言 (UR)	-	-	374	201
事件平均帖子数	804	62	223	251

4.3.2 基线模型

本章实验所要对比的基线模型中，除了在第三章使用过的针对谣言传播特点精心设计的架构复杂的 BiGCN^[20] 和 ClaHi-GAT^[25] 模型之外，本章还比较了以下结构简单的 GNN 方法包括：

GIN^[33]：是一种广泛使用的 GNN 架构，在图分类任务上有良好表现。需注意，GIN 以无向的谣言传播图作为输入。

GIN-GCLP^[41]：是一个由图自监督对比学习预训练初始化（如4.2.2一节所述）的 GIN 编码器，可视为所提自训练方法 RDST-GCLP 的初始模型。

GIN-GAEP^[44]：是一个由图自编码重构预训练初始化（如4.2.2一节所述）的 GIN 编码器，此方法同时采用 GIN 架构作为自编码器的编码器和解码器，GIN-GAEP 可视为所提自训练方法 RDST-GAEP 的初始模型。

另外，本章还比较了一个典型的半监督自训练方法：

CL^[49]：是一个具有代表性的基于课程学习的自训练方法，其采用课程学习的思想在伪标注过程中逐渐增加伪标签选择的比例，然而与 RDST 不同之处在于，CL 没有具体关注每个类的学习情况，本章将此工作应用在谣言检测任务上予以复现，并使用 GIN 作为图编码器架构。

鉴于本章所提出的方法分别采用了图对比学习预训练和图自编码重构预训练来初始化自训练框架的模型，因此将所提的模型分别命名为 RDST-GCLP 和 RDST-GAEP。

4.3.3 实验设置与参数选择

在数据集的划分上, Weibo 和 DRWeibo 数据集按照 6: 2: 2 的比例划分为训练集/验证集/测试集。Twitter15 和 Twitter16 数据集的谣言传播图数量较少, 因此这两个数据集按照 8: 2 的比例划分为训练集/测试集并进行五折交叉验证(与之前的研究工作^[20,24,25]实验设置一致)。划分后的每个子集都具有相同的类别分布, 并且每个子集的类别分布都是平衡的。为进行半监督实验, 将训练集按照不同的比例进一步划分为一个有标注样本集合和一个无标注样本集合(无标注集的标签信息被删除)。考虑到不同数据集的规模大小, 对于 Weibo 和 DRWeibo 数据集, 将有标注样本集的大小 N_l 设置为 $\{5, 10, 20, 40, 100, 200, 500\}$; 对于 Twitter15 数据集, 将有标注样本集的大小设置为 $\{10, 20, 40, 80, 100\}$; 对于 Twitter16 数据集, 将有标注样本集的大小设置为 $\{10, 20, 40, 80\}$ 。每个训练集剩余的样本将被用作无标注样本集, 其中所选取的 N_l 最大值都已确保有标注样本的数量小于无标注样本的数量。有标注样本按照随机概率抽取, 为了减少随机采样带来的偏置, 对于每种数据规模 N_l , 随机对有标注样本做五次不同的随机采样, 获得五种不同的有标注集/无标注集的组合, 并在最终实验汇报五次结果的平均值。

在节点初始特征的选择上, 对于 Twitter15、16 数据集, 采用 5000 维的 BOW (Bag-of-words) 词频特征来获取谣言文本的表示作为每个帖子节点的特征表示; 对于 Weibo 和 DRWeibo 数据集, 用 200 维的 word2vec^[35] 特征获取谣言文本的表示。GNN 的层数设置为 2 层, 所有使用的 GNN 都采用平均池化的方式来获取全图的特征表示。在隐层节点特征维度的选择上, Twitter15 和 16 数据集设置为 64, Weibo 和 DRWeibo 数据集设置为 128。全局阈值中伪标注样本选择比例的增长步幅 γ 在 Weibo 和 DRWeibo 数据集设置为 0.1, 在 Twitter15 和 16 数据集设置为 0.2。局部阈值中调节选取比例的指数超参数 α 在 Weibo, DRWeibo, Twitter15 和 16 四个数据集上分别设置为 0.5, 0.25, 1.0, 1.0。调节分布对齐模块强度的指数超参数 β 在四个数据集一致设置为 0.5。无监督损失项和公平性正则损失项的权重 λ_u 和 λ_f 分别设置为 0.8 和 0.4。在 Twitter15、16、Weibo 和 DRWeibo 四个数据集上训练的 mini-batch 大小分别设置为 256, 128, 32, 32。模型通过 Adam 优化器^[62] 和反向传播算法来更新参数, 学习率和权重衰减系数分别设置为 0.0005 和 0.0001。最优超参数的组合通过 Twitter15、16 数据集的第一折 (fold-0 set) 和 Weibo 与 DRWeibo 数据集的验证集来确定。所有模型在 PyTorch²环境实现, 并且所有基线模型方法都被重新复现。

²<https://pytorch.org/>

4.3.4 实验结果与分析

所提 RDST 方法及基线模型在 Weibo、DRWeibo、Twitter15、Twitter16 四个数据集上的半监督实验结果如表4-2，表4-3，表4-4和表4-5所示。与现有多数半监督学习相关研究^[49,53-55] 的实验设置类似，本节汇报了不同有标注数据规模下的分类精度 (Accuracy, Acc.)，表中 “# Label” 表示使用的有标注样本数量，即 N_l ，表中 “All” 表示完全有监督训练（训练集中所有样本都为有标注样本）。另外，表中最优性能已做加粗显示。

根据表4-2，表4-3，表4-4和表4-5的实验结果有如下发现：

表 4-2 Weibo 数据集的半监督实验结果

Table 4-2 Semi-supervised rumor detection results on Weibo dataset

# Label	准确率 (Acc.)							
	5	10	20	40	100	200	500	All
GIN	0.691	0.708	0.746	0.791	0.874	0.907	0.940	0.945
BiGCN	0.631	0.653	0.762	0.832	0.879	0.909	0.945	0.948
ClaHi-GAT	0.665	0.705	0.773	0.823	0.887	0.914	0.947	0.953
GIN-GCLP	0.766	0.773	0.831	0.856	0.912	0.927	0.946	0.962
GIN-GAEP	0.759	0.823	0.854	0.861	0.900	0.924	0.957	0.966
CL	0.687	0.712	0.764	0.836	0.896	0.927	0.949	-
RDST-GCLP	0.860	0.868	0.895	0.917	0.937	0.945	0.960	-
RDST-GAEP	0.824	0.853	0.878	0.909	0.930	0.951	0.963	-

表 4-3 DRWeibo 数据集的半监督实验结果

Table 4-3 Semi-supervised rumor detection results on DRWeibo dataset

# Label	准确率 (Acc.)							
	5	10	20	40	100	200	500	All
GIN	0.557	0.586	0.667	0.713	0.739	0.775	0.819	0.868
BiGCN	0.553	0.609	0.650	0.701	0.754	0.797	0.836	0.885
ClaHi-GAT	0.568	0.614	0.654	0.705	0.751	0.780	0.841	0.888
GIN-GCLP	0.562	0.607	0.721	0.749	0.769	0.798	0.834	0.876
GIN-GAEP	0.559	0.592	0.685	0.726	0.747	0.792	0.839	0.882
CL	0.553	0.591	0.714	0.740	0.777	0.792	0.827	-
RDST-GCLP	0.653	0.689	0.746	0.771	0.801	0.825	0.852	-
RDST-GAEP	0.626	0.65	0.708	0.751	0.787	0.838	0.866	-

表 4-4 Twitter15 数据集的半监督实验结果

Table 4-4 Semi-supervised rumor detection results on Twitter15 dataset

# Label	准确率 (Acc.)					
	10	20	40	80	100	All
GIN	0.445	0.524	0.625	0.737	0.757	0.848
BiGCN	0.495	0.577	0.678	0.753	0.791	0.880
ClaHi-GAT	0.507	0.585	0.676	0.760	0.785	0.875
GIN-GCLP	0.464	0.556	0.663	0.756	0.788	0.865
GIN-GAEP	0.501	0.588	0.699	0.778	0.809	0.887
CL	0.456	0.549	0.665	0.766	0.808	-
RDST-GCLP	0.569	0.653	0.740	0.793	0.827	-
RDST-GAEP	0.603	0.665	0.771	0.825	0.842	-

表 4-5 Twitter16 数据集的半监督实验结果

Table 4-5 Semi-supervised rumor detection results on Twitter16 dataset

# Label	准确率 (Acc.)				
	10	20	40	80	All
GIN	0.510	0.594	0.688	0.779	0.868
BiGCN	0.561	0.656	0.737	0.816	0.879
ClaHi-GAT	0.565	0.675	0.765	0.820	0.887
GIN-GCLP	0.549	0.632	0.720	0.803	0.883
GIN-GAEP	0.566	0.641	0.737	0.825	0.892
CL	0.537	0.625	0.769	0.812	-
RDST-GCLP	0.635	0.753	0.805	0.839	-
RDST-GAEP	0.670	0.774	0.819	0.857	-

(1) 尽管现有根据谣言传播特点精心设计的架构复杂的谣言检测模型 (BiGCN、ClaHi-GAT) 在有标注样本充足的条件下表现出了良好的性能, 但当有标注数据匮乏时, 它们表现出了明显的过拟合倾向。例如, 当每个类别只有 5 个或 10 个有标注样本可用时, 这些架构复杂的谣言检测模型甚至比结构简单的基线模型表现更差, 此现象在 Weibo 数据集尤为明显, 当每个类别只有 5 个有标注的谣言样本可用时, BiGCN 和 ClaHi-GAT 的检测精度比结构简单的朴素 (vanilla) GIN 模型还要低 6.0% 和 2.6%。

(2) GIN-GCLP 和 GIN-GAEP 利用无标注数据的谣言传播图进行图自监督预训练并利用有标注数据微调, 在所有的有标注数据规模下均超过了 GIN 模型。当与架构更复杂的 BiGCN 和 ClaHi-GAT 模型比较时, GIN-GCLP 和 GIN-GAEP 实

现了与其相当的性能表现。其中，在 Weibo 数据集的表现较为理想，在半监督设置下，GIN-GAEP 在不同数据规模下比 BiGCN 和 ClaHi-GAT 提升了 1.0%~17.0%。同样，在完全监督（利用了全部有标注数据）设置下，GIN-GAEP 的检测精度比 BiGCN 提高了 1.8%，比 ClaHi-GAT 提高了 1.3%。以上结果证明了所提自监督预训练方法在增强图编码器泛化性、缓解模型过拟合上的作用，因此 GIN-GCLP 和 GIN-GAEP 可以被当作合适的自训练初始模型来减少模型早期的偏置。

(3) 自训练方法的代表性工作 CL 得益于其高效的伪标注策略，在多数有标注数据规模下的表现一致地超过了它的基模型 GIN。然而当每个类别仅有 5 个有标注样本可用时，在 Weibo 和 DRWeibo 两个数据集，CL 的表现差于 GIN，如表4-2和表4-3所示。此现象印证了本章之前的假设，即由极少有标注数据训练的模型不可避免地含有偏置，随自训练进行，误差逐渐累积，导致模型性能的退化。

(4) 本章所提消偏自训练谣言检测方法 RDST-GCLP 和 RDST-GAEP 在全部四个数据集的所有有标注数据规模下均比所有基线模型的表现有大幅提升。特别地，当有标注数据极其稀少时，如每类只有 5 个有标注样本可用时，RDST-GCLP 在 Weibo 和 DRWeibo 数据集上的检测精度比 CL 分别高出了 17.3% 和 10.0%。同样，当每类只有 10 个有标注样本可用时，RDST-GAEP 在 Twitter15 和 Twitter16 数据集上的检测精度比 CL 分别高出了 14.7% 和 13.3%。此外，在最大的有标注数据规模下，RDST-GAEP 的表现与一些基线模型在完全监督设置下的表现相当，甚至超过了部分基线模型在完全监督设置下的表现。以上实验结果证明了本章提出的消偏自训练方法 RDST，在模型自监督预训练初始化和自适应阈值伪标注方法的帮助下，显著地减少了确定性偏置 (confirmation bias)，提高了自训练过程的效率。

4.3.5 消融实验

为验证 RDST 方法各部分模块的有效性，设计如下消融性实验：

对四个数据集各选取三个不同的有标注样本规模，对于 Weibo 和 DRWeibo 数据集，选取 $N_l = \{5, 40, 100\}$ ，对于 Twitter15、16 数据集，选取 $N_l = \{10, 40, 80\}$ 。在消融实验中，比较以下变体模型：(1) -w/o LCST (without local class specific threshold)，表示在伪标注阶段不使用类别特定的局部阈值，仅保留全局阈值；(2) -w/o PT (without pre-training)，表示自训练的初始模型由随机参数初始化，并直接利用有标注数据训练；(3) -w/o DA (without distribution alignment)，表示移除分布对齐模块；(4) -w/o LA (without labeled data allocated averagely)，表示原始的有标注样本被随机分配到新有标注集的每个 mini-batch，而不再是平均分配。

如图4-3所示，整体上看，每个模块都能为整体的自训练方法 RDST 带来性能上的提升。具体地，-w/o PT 在 Weibo 和 DRWeibo 数据集上的准确率有明显的下

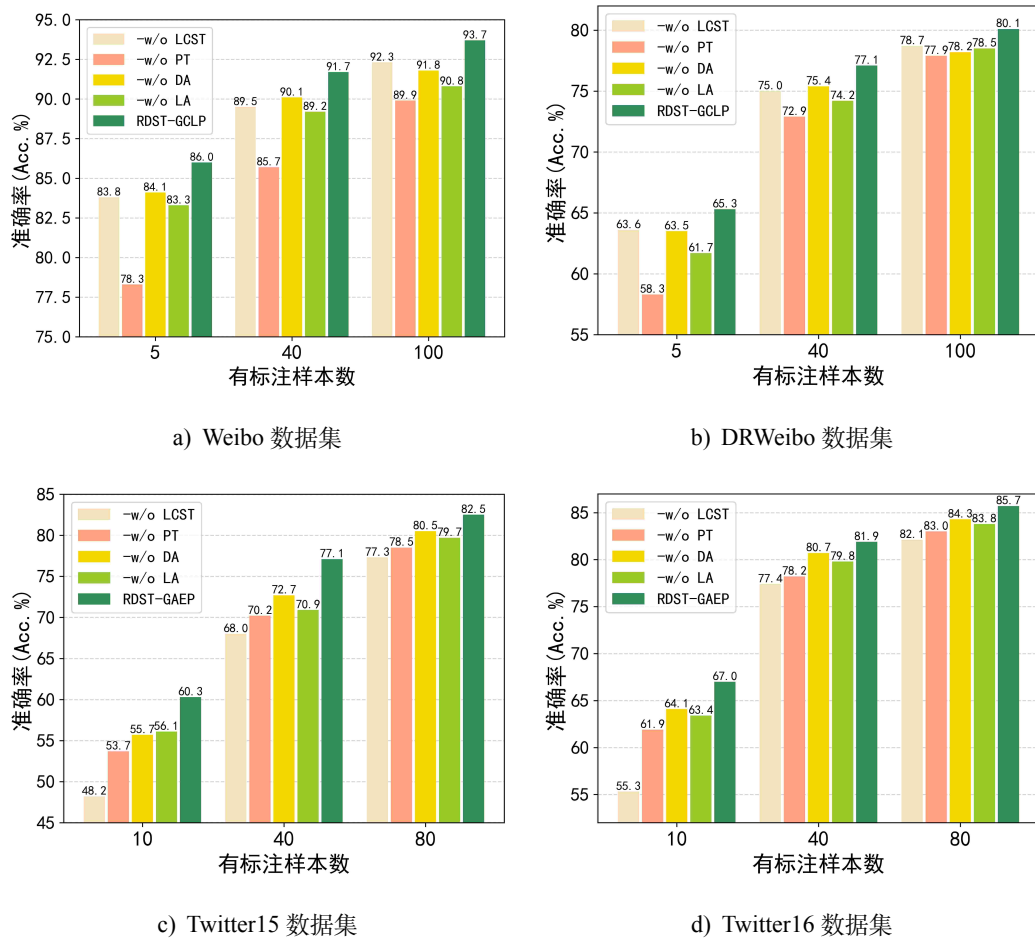


图 4-3 消融实验结果

Figure 4-3 Experimental results of ablation study

降，这表明以图自监督预训练方法来对自训练模型初始化在减小模型误差上具有重要作用。此外，在 Twitter15 和 Twitter16 两个数据集上，-w/o LCST 在所有模型变体中表现最差，一种可能的原因是 Twitter15 和 Twitter16 数据集与 Weibo 和 DRWeibo 数据集相比包含更多的谣言类别，其分类任务更困难，因此在自训练过程中，关注模型对每个类别的学习情况并设定类别特定的局部阈值更加重要。-w/o DA 和 -w/o LA 都产生了精度的下降，验证了它们在校正伪标签和利用伪标注样本上的有效性。

此外，本节设计消融实验单独探究了自定进度全局阈值(self-paced global threshold) 模块的有效性：在自训练的伪标注选择过程中采用一组固定置信度阈值 {0.8, 0.9, 0.95} 与 RDST 中的自定进度全局阈值进行对比。四种阈值的迭代的步骤次数统一设定为自定进度阈值法的迭代次数。表4-6展示了自训练每轮迭代过程中伪标签选择的数量（表中用“# PL”表示）和测试准确率。结果表明，使用一个相对较高的固定阈值会导致自训练的表现不稳定，并且最终的准确率与自定进度的阈值

方法相比也有较大差距。而自定进度的动态阈值方法从易到难地利用伪标注数据，在选取伪标签时兼顾了其质量和数量，准确率随自训练迭代过程稳步提升。

表 4-6 固定阈值和自定进度阈值的性能对比

Table 4-6 Performance comparison among fixed and self-paced thresholds

迭代步骤	伪标注样本选择阈值							
	0.8		0.9		0.95		自定进度	
	# PL	准确率	# PL	准确率	# PL	准确率	# PL	准确率
Iteration-0	-	0.853	-	0.853	-	0.853	-	0.853
Iteration-1	1978	0.863	1464	0.866	1159	0.864	255	0.865
Iteration-2	1835	0.872	1096	0.865	1540	0.848	504	0.864
Iteration-3	2038	0.870	432	0.852	463	0.854	747	0.868
Iteration-4	2212	0.874	707	0.865	1406	0.851	1079	0.872
Iteration-5	1760	0.886	585	0.863	1560	0.844	1317	0.874
Iteration-6	1597	0.882	1342	0.859	572	0.836	1473	0.873
Iteration-7	2265	0.889	826	0.868	1530	0.849	1813	0.883
Iteration-8	2008	0.891	1215	0.854	1594	0.857	2081	0.899
Iteration-9	2194	0.887	1047	0.850	1527	0.835	2355	0.907
Iteration-10	2041	0.887	1378	0.861	1524	0.842	2370	0.910

4.3.6 参数分析

本小节将对 RDST 方法中的一些重要超参数进行详细研究：

(1) 全局阈值的增长步幅 γ ：是控制自训练每轮迭代中伪标注样本选取数量的重要参数。最优的增长步幅 γ 应使自训练快速收敛，同时保证伪标签的准确率，实验选取三种不同的增长步幅 $\{0.05, 0.1, 0.2\}$ ，在全部四个数据集中每类选取 40 个有标注样本进行半监督实验。图4-4展示了在不同增长步幅 γ 下自训练每轮迭代的准确率变化。结果表明当 γ 分别设置为 0.1, 0.1, 0.2, 0.2 时，模型在 Weibo, DRWeibo, Twitter15 和 Twitter16 四个数据集获得最佳表现。减小 γ 并不会带来性能提升，相反会造成更不稳定的性能表现，并且会增加运行的时间，这种现象在 DRWeibo 和 Twitter15 数据集上表现得较为明显。

(2) 类别再平衡的强度 α ：是类别特定局部阈值中控制各类再平衡比例的指数超参数，图4-5展示了采用不同的类别再平衡强度 α 下的准确率变化曲线。当 $\alpha = 0$ 时，所有类别的采样比例都为 1，意味着类别特定局部阈值完全失效，这时模型明显表现出对某些特定类别的偏置，在四个数据集的表现都是最不理想的。当 α 分

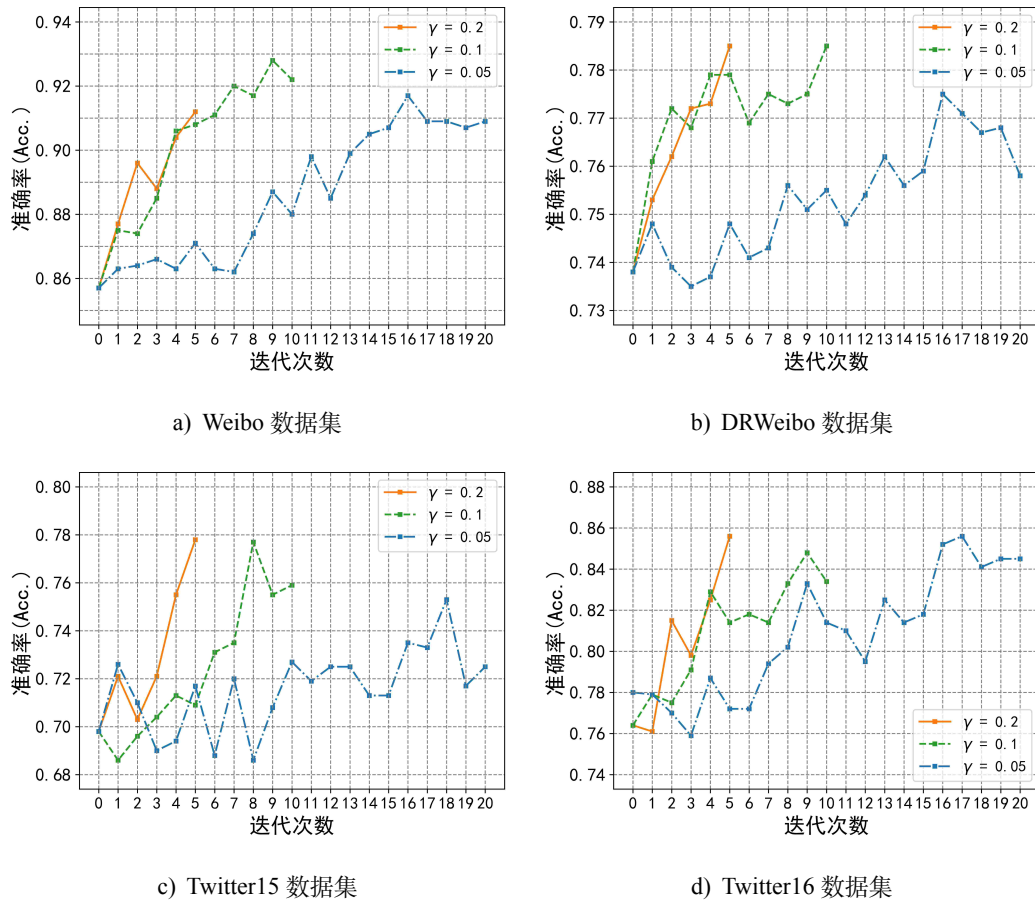


图 4-4 不同增长步幅的影响

Figure 4-4 Effect of various incremental stride

别设置为 0.5, 0.25, 1.00, 1.00 时, 模型在 Weibo, DRWeibo, Twitter15 和 Twitter16 四个数据集获得最佳表现。据此推测 Twitter15 和 Twitter16 两个数据集的最优 α 值较大可能因为这两个数据集包含更多的谣言类别, 类间更容易发生混淆。此外, 过大的 α 会致使类别再平衡的程度过大, 造成相反方向的类别不平衡并使模型性能退化。

(3) 分布对齐的强度 β : 是分布对齐模块中控制分布对齐强度的指数超参数, 图4-6展示了不同分布对齐强度 β 下的准确率变化曲线。当 $\beta = 0$ 时, 分布对齐模块完全失效, 所有数据集上的自训练精度一致达到最低值。这表明在自训练早期, 由少量有标注样本训练而得的具有偏置的模型可能会盲目将大多数无标注样本预测为某一特定类, 产生了大量错误的伪标签。而分布对齐模块通过将模型对无标注样本预测结果的类别分布与有标注样本集类别分布对齐, 校正了之前预测的伪标签, 解决了上述问题。随着 β 增大, 模型自训练的表现逐渐提升。当 β 分别设置为 0.25, 0.50, 0.50, 0.50 时, 模型在 Weibo, DRWeibo, Twitter15 和 Twitter16 四个数据集获得最佳表现。随 β 继续增大, 模型的表现并没有随之提升, 可能的

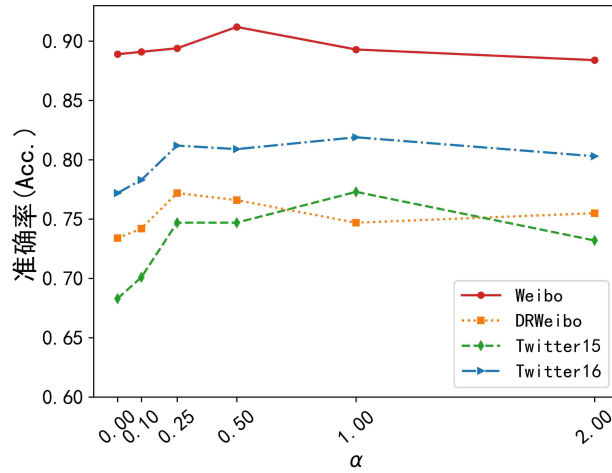


图 4-5 不同类别再平衡强度的影响

Figure 4-5 Effect of the class-rebalancing strength

原因在于：伪标签的分布过度平衡会使模型将更多样本错误地预测为少数类，这降低了少数类别预测的精确率（precision），与所提方法利用更多具有高精确率的少数类样本的思路相矛盾。

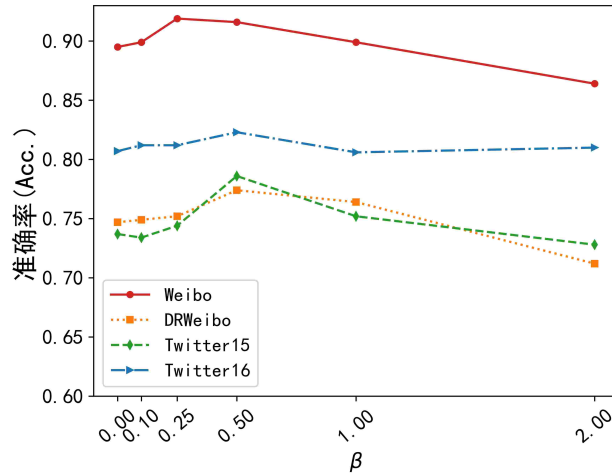


图 4-6 不同分布对齐强度的影响

Figure 4-6 Effect of the distribution alignment strength

(4) 无监督损失项和公平化损失项权重 λ_u 和 λ_f ：在 Weibo 数据集每个类使用 5 个和 40 个有标注样本的设置下进行了实验，表4-7展示了 λ_u 和 λ_f 不同组合下的准确率，最优值已通过加粗显示，下划线表示本章所选取的参数组合。所采用的参数组合 ($\lambda_u = 0.8$, $\lambda_f = 0.4$) 达到了较好表现，过大或过小的 λ_u 都会使模型对伪标签过于信赖或忽视，导致次优的表现。同样地，一个合适的 λ_f 值能够确保模型对每个类别公平地预测。

表 4-7 不同损失项权重的影响
Table 4-7 Effect of trade-off hyper-parameters

# Label	5					40				
λ_u/λ_f	0.1	0.4	0.8	1.0	2.0	0.1	0.4	0.8	1.0	2.0
0.1	0.824	0.795	0.829	0.833	0.857	0.890	0.901	0.891	0.903	0.895
0.4	0.825	0.832	0.861	0.841	0.832	0.890	0.907	0.892	0.902	0.880
0.8	0.812	<u>0.860</u>	0.869	0.840	0.842	0.892	0.916	0.907	0.885	0.901
1.0	0.842	0.833	0.817	0.857	0.863	0.893	0.895	0.902	0.901	0.896
2.0	0.832	0.778	0.814	0.807	0.840	0.889	0.890	0.881	0.892	0.898

4.3.7 特征可视化

为了更直观地反映出所提方法学习到的不同类别的特征表示, 使用 T-SNE 进行了可视化研究。选取 Weibo 和 Twitter15 两个数据集, 并利用以下模型来提取所有测试数据的图特征表示。随后, 利用它们的原始标签, 对获取的图特征表示进行可视化。

对于 Weibo 数据集, 在每个类别只使用 5 个有标注数据的条件下, ClaHi-GAT、GIN-GCLP、CL 和 RDST-GCLP 学到的图表示的对比如图4-7所示。从图中可以发现, 在有标注数据极度匮乏的情况下, ClaHi-GAT 难以学到不同类别的判别性特征, 没有清晰的类别簇形成。提出的 GIN-GCLP 在利用大量无标注数据进行图自监督预训练后, 增强了图编码器的泛化能力, 学到的特征被清晰地分成两个簇。因此, 将 GIN-GCLP 作为提出的自训练方法 RDST-GCLP 的初始模型是合理的。典型的自训练方法 CL 没有设计用于消除偏差的模块, 尽管它获得了更紧凑的特征表示, 但由于自训练的偏置和误差积累, 来自两个类别的特征大量混合交织在了一起。然而, RDST-GCLP, 得益于图自监督预训练和自适应阈值的伪标注策略, 减少了自训练过程中的误差积累, 学到了更紧凑和更易区分的特征表示。

对于 Twitter15 数据集, 每个类别使用 40 个有标注数据, 利用 ClaHi-GAT、GIN-GAEP、CL 和 RDST-GAEP 提取图特征表示, 如图4-8所示。从图中可以发现类似的结果, 即 ClaHi-GAT 未能生成具有区分性的图特征表示。而 GIN-GAEP 通过使用掩码图自编码器进行预训练, 形成了几个类别簇。然而, 每个类别簇的边界仍然不够清晰。对于典型的自训练方法 CL, 尽管它形成了四个明显的类别簇, 但可以观察到模型对 NR 类别存在高度偏置。具体来说, NR 类别特征的簇中包含大量实际属于其他三个类别的样本, 这与第4.2.3节的发现一致。在这种情况下, 模型倾向于将所有类别预测为 NR, 导致 NR 类别的召回率高但精确率低, 相反, 其他三个类别的精确率较高。因此, 在自训练过程中选择伪标注样本时, 遵循从少数类别

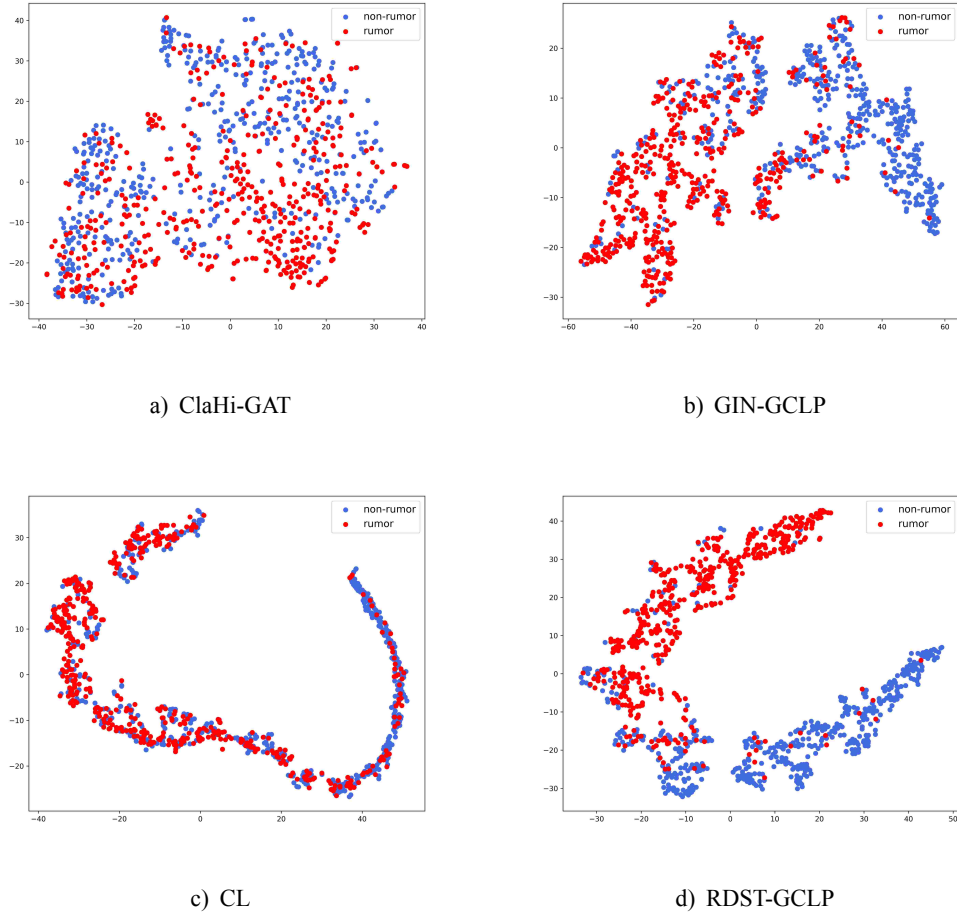


图 4-7 Weibo 数据集特征可视化

Figure 4-7 Visualization of features on Weibo dataset

选择更多样本，从多数类别选择更少样本的原则是合理的。得益于自监督预训练，特别是类别特定的局部阈值，提出的 RDST-GAEP 在一定程度上缓解了模型在自训练学习过程中的偏差，NR 类别簇中属于其他类别样本的数量已经显著减少。

4.4 本章小结

目前现有的谣言检测方法在完全有监督的实验设置下取得了良好表现，而有标注的谣言数据难以大量获得，利用有限的有标注数据和大量无标注数据联合学习的半监督学习方法在谣言检测领域仍有待探索和研究。在此情景下，本章提出了一个自监督预训练辅助的消偏自训练方法 RDST 来进行半监督的谣言检测。为了在自训练的初始阶段减少噪声和偏置，RDST 利用大量谣言无标注数据的传播图结构来进行图自监督预训练，增强了图编码器的泛化性与鲁棒性，用其作为自训练初始模型，减少了自训练初始阶段错误伪标签的生成。为了提高每轮迭代选

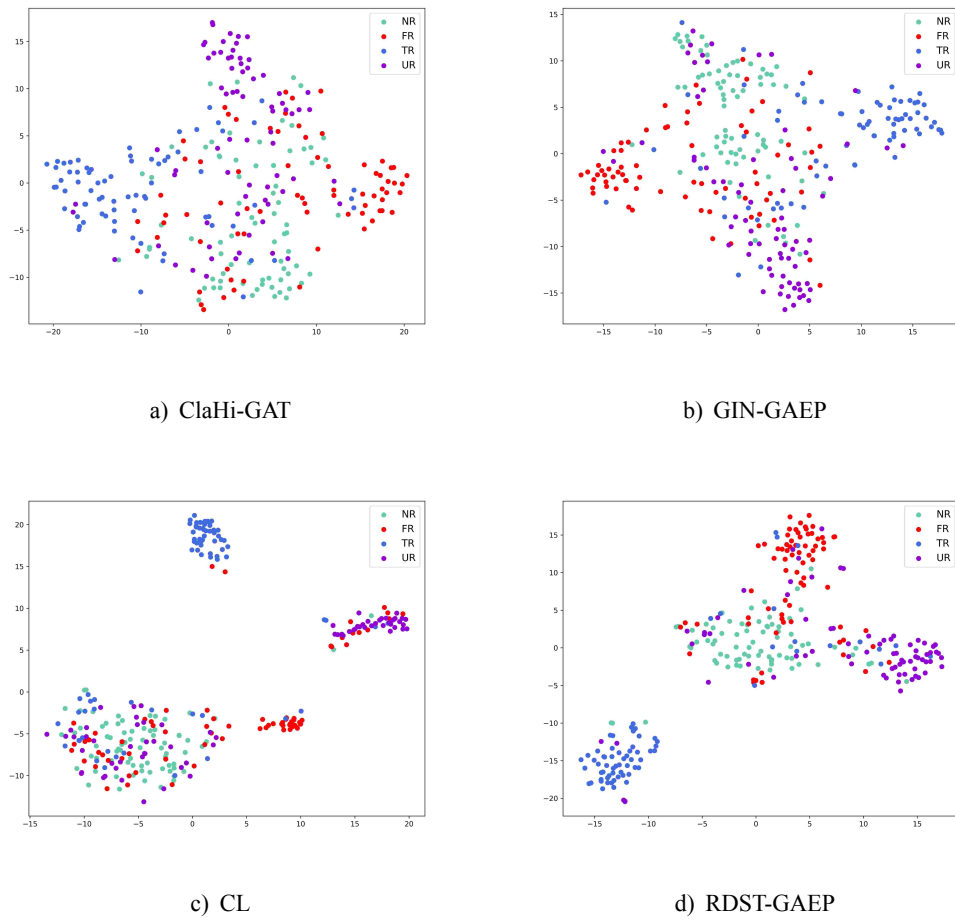


图 4-8 Twitter15 数据集特征可视化

Figure 4-8 Visualization of features on Twitter15 dataset

取的伪标注样本的质量，RDST 设计了一种自适应阈值的伪标注策略，包括自定进度的全局阈值来控制伪标签的整体利用进度和类别特定的局部阈值来关注模型对每个类别的具体学习情况。通过一系列精心设计的消偏技术：分布对齐，模型参数重新初始化，公平性正则化项，RDST 进一步减少了自训练迭代过程中的误差累积，提高了自训练方法的整体性能表现。

在四个谣言公开数据集上进行的实验表明：所提方法在半监督实验设置下的性能表现显著超过了现有架构复杂的谣言检测模型，特别是在有标注数据极其稀少的情况下，RDST 仍具有良好表现。与传统的自训练方法相比，RDST 融合了多个消偏模块，因此在自训练过程中减少了噪声的累积，表现出了良好的性能。

5 总结与展望

5.1 工作总结

本文主要针对谣言数据集规模小，现有谣言检测模型泛化性和鲁棒性差，并倾向于过拟合的问题，将谣言检测建模为图数据的分类问题并利用自监督学习和半监督学习的思路予以解决。本文研究的主要贡献如下：

(1) 提出了一种基于图自监督对比学习的谣言检测方法 RD-GCSL。现有谣言检测模型在有标注数据充足时能有效解决分类问题，然而常用谣言数据集规模较小，各种针对谣言特点精心设计的模型倾向于过拟合。同时，现有模型鲁棒性不足，谣言制造者或传播者在谣言的传播过程中通过某些恶意操作破坏了原有的谣言传播结构，致使现有模型的检测结果出现了错误。针对以上问题，本文结合社交媒体中谣言传播特点，设计了三种图级的数据增强策略来模拟对原图的扰动，在此基础上建立自监督对比学习任务并和有监督分类任务端到端地联合训练，使谣言图编码器能够捕获谣言更趋本质的特征，缓解了过拟合现象，提高了模型的鲁棒性与泛化性能。在来源于主流社交媒体平台的三个谣言公开数据集上进行的实验中，本文提出的 RD-GCSL 方法在使用全部有标注数据和仅使用部分有标注数据的设置下均比基线方法获得了更好的性能表现，验证了所提图自监督对比学习方法在谣言检测问题上的有效性。通过消融实验，探究了图编码器模块、数据增强模块和投影头模块对整体模型的影响，同时验证了 RD-GCSL 方法不依赖于特定的谣言图编码器，能作为一个通用的框架来提高现有谣言检测模型的性能。

(2) 提出了一个自监督预训练辅助的消偏自训练半监督谣言检测方法 RDST。现有各种架构复杂的谣言检测模型在完全监督设置下能取得良好的检测效果，然而有标注的谣言数据匮乏，无标注谣言数据更易获得，而同时利用有限有标注数据和大量无标注数据联合学习的半监督学习方法在谣言检测领域仍有待探索和研究。自训练作为半监督学习中的典型方法，尽管使用起来简单，但是其内在机制存在噪声积累的缺陷。针对以上问题，本文提出了一个消偏的自训练谣言检测方法 RDST。首先，为减少自训练初始阶段噪声的产生，RDST 利用无标注谣言数据的传播图结构对图编码器进行自监督预训练，以提高其泛化性能与鲁棒性，随后用有限的有标注数据微调作为自训练的初始模型。为了进一步提高自训练每轮迭代所选取伪标签的质量，本文进一步设计了一种自适应阈值的伪标注方法，包括一个自定进度的全局阈值来控制伪标注样本的整体使用进度和一个类别特定的局部阈值来关注模型对不同类别的学习情况。随后通过一系列自训练消偏技术包括：

分布对齐、模型参数重新初始化、公平性正则项，进一步减少了自训练过程中误差的累积。在四个公开谣言数据集进行的实验表明，RDST 在半监督实验设置下的性能表现大幅超过了现有架构复杂的谣言检测模型，特别是在有标注数据极其稀少的情景下仍有良好表现。此外，与传统的自训练方法相比，本文所提 RDST 方法融入了多个消偏模块，大幅减少了噪声的积累，获得了更好的性能表现。

5.2 未来展望

本文所提 RD-GCSL 和 RDST 方法在谣言检测任务上表现出了较为良好的效果，但仍存在一些局限性，所提方法在未来仍有一些可以改进或拓展之处：

(1) RD-GCSL 方法采用的是较为常规的对比学习方法，由于此方法的内在机制要求，需要将原图和数据增强生成的两种扰动图都输入给图编码器来提取图表示特征，因此在时间和空间上需要一定开销。近年来还有其它设计思路更新颖的对比学习框架可供使用，比如在正负样本的构建上或是数据增强策略的选择上，如何在提高模型泛化性能的基础上减少训练所需的时间和空间成本是一个值得研究的方向。

(2) RDST 方法采用的是基于自训练的半监督方法，此外，半监督的另一大类方法一致性正则化（consistency regularization）近年来的性能表现已逐渐超过自训练方法的表现。然而对于图模态的数据，如何设计相应的“强数据增强”和“弱数据增强”策略以使得一致性正则化方法有效是在未来值得研究的一个方向。

(3) RDST 方法是在谣言类别平衡的条件下提出的半监督学习方法，在真实情况中，谣言的类别分布往往是不平衡的。因此，在类不平衡的情景下进行半监督的谣言检测更符合真实情景需要，如何设计一种类不平衡的半监督学习方法使其在半监督谣言检测任务上有效是未来值得研究的另一个方向。

参考文献

- [1] KEMP S. DIGITAL 2024: GLOBAL OVERVIEW REPORT [EB/OL]. [2024-01-31]. <https://datareportal.com/reports/digital-2024-global-overview-report>.
- [2] DiFonzo N, Bordia P. Rumor, gossip and urban legends. [J]. Diogenes, 2007, 54 (1): 19–35.
- [3] Ma J, Gao W, Mitra P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks [C]. //Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, 2016: 3818–3824.
- [4] Ma J, Gao W, Wong K. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning [C]. //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017: 708–717.
- [5] Guo H, Cao J, Zhang Y, et al. Rumor Detection with Hierarchical Social Attention Network [C]. //Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, 2018: 943–951.
- [6] Shu K, Sliva A, Wang S, et al. Fake News Detection on Social Media: A Data Mining Perspective [J]. SIGKDD Explor., 2017, 19 (1): 22–36.
- [7] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C]. //Proceedings of the 20th International Conference on World Wide Web, Hyderabad, 2011: 675–684.
- [8] Qazvinian V, Rosengren E, Radev D R, et al. Rumor has it: Identifying Misinformation in Microblogs [C]. //Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, 2011: 1589–1599.
- [9] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo [C]. //Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, Beijing, 2012: 1–7.
- [10] Kwon S, Cha M, Jung K, et al. Prominent Features of Rumor Propagation in Online Social Media [C]. //2013 IEEE 13th International Conference on Data Mining, Dallas, 2013: 1103–1108.
- [11] Ma J, Gao W, Wei Z, et al. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites [C]. //Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, 2015: 1751–1754.
- [12] Liu X, Nourbakhsh A, Li Q, et al. Real-time Rumor Debunking on Twitter [C]. //Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, 2015: 1867–1870.
- [13] Ma J, Gao W, Mitra P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks [C]. //Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, 2016: 3818–3824.
- [14] Yu F, Liu Q, Wu S, et al. A Convolutional Approach for Misinformation Identification [C]. //Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, 2017: 3901–3907.

- [15] Liu Y, Wu Y B. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks [C]. //Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, 2018: 354–361.
- [16] Khoo L M S, Chieu H L, Qian Z, et al. Interpretable Rumor Detection in Microblogs by Attending to User Interactions [C]. //The Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, 2020: 8783–8790.
- [17] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need [C]. //Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, 2017: 5998–6008.
- [18] Zubiaga A, Aker A, Bontcheva K, et al. Detection and Resolution of Rumours in Social Media: A Survey [J]. ACM Comput. Surv., 2018, 51 (2): 32:1–32:36.
- [19] Ma J, Gao W, Wong K. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks [C]. //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, 2018: 1980–1989.
- [20] Bian T, Xiao X, Xu T, et al. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks [C]. //The Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, 2020: 549–556.
- [21] Zhang P, Ran H, Jia C, et al. A lightweight propagation path aggregating network with neural topic model for rumor detection [J]. Neurocomputing, 2021, 458: 468–477.
- [22] 张鹏飞. 轻量级可解释谣言检测方法研究 [D]. 北京: 北京交通大学, 2021.
- [23] 韩雪明, 贾彩燕, 李轩涯, 等. 传播树结构结点及路径双注意力谣言检测模型 [J]. 计算机科学, 2023, 50: 22–31.
- [24] Wei L, Hu D, Zhou W, et al. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection [C]. //Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual Event, 2021: 3845–3854.
- [25] Lin H, Ma J, Cheng M, et al. Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks [C]. //Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual Event, 2021: 10035–10047.
- [26] Ran H, Jia C, Zhang P, et al. MGAT-ESM: Multi-channel graph attention neural network with event-sharing module for rumor detection [J]. Inf. Sci., 2022, 592: 402–416.
- [27] 冉宏艳. 在线社会网络中的谣言检测方法研究 [D]. 北京: 北京交通大学, 2023.
- [28] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database [C]. //2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, 2009: 248–255.
- [29] Zubiaga A, Liakata M, Procter R, et al. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads [J]. PLOS ONE, 2016, 11 (3): 1–29.
- [30] Cui C, Jia C. Propagation Tree Is Not Deep: Adaptive Graph Contrastive Learning Approach for Rumor Detection [C]. //Thirty-Eighth AAAI Conference on Artificial Intelligence, Vancouver, 2024: 73–81.

- [31] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks [C]. //5th International Conference on Learning Representations, Toulon, 2017.
- [32] Velickovic P, Cucurull G, Casanova A, et al. Graph Attention Networks [C]. //6th International Conference on Learning Representations, Vancouver, 2018.
- [33] Xu K, Hu W, Leskovec J, et al. How Powerful are Graph Neural Networks? [C]. //7th International Conference on Learning Representations, New Orleans, 2019.
- [34] Hamilton W L, Ying Z, Leskovec J. Inductive Representation Learning on Large Graphs [C]. //Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, 2017: 1024–1034.
- [35] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [C]. //1st International Conference on Learning Representations, Scottsdale, 2013.
- [36] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]. //Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019: 4171–4186.
- [37] Chen T, Kornblith S, Norouzi M, et al. A Simple Framework for Contrastive Learning of Visual Representations [C]. //Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020: 1597–1607.
- [38] He K, Fan H, Wu Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning [C]. //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020: 9726–9735.
- [39] Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization [C]. //7th International Conference on Learning Representations, New Orleans, 2019.
- [40] Velickovic P, Fedus W, Hamilton W L, et al. Deep Graph Infomax [C]. //7th International Conference on Learning Representations, New Orleans, 2019.
- [41] Sun F, Hoffmann J, Verma V, et al. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization [C]. //8th International Conference on Learning Representations, Addis Ababa, 2020.
- [42] Hassani K, Ahmadi A H K. Contrastive Multi-View Representation Learning on Graphs [C]. //Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020: 4116–4126.
- [43] You Y, Chen T, Sui Y, et al. Graph Contrastive Learning with Augmentations [C]. //Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual, 2020: 5812–5823.
- [44] Hou Z, Liu X, Cen Y, et al. GraphMAE: Self-Supervised Masked Graph Autoencoders [C]. //KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, 2022: 594–604.
- [45] Oord A v d, Li Y, Vinyals O. Representation learning with contrastive predictive coding [J]. arXiv preprint arXiv:1807.03748, 2018.

- [46] Lee D-H, et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks [C]. //Workshop on challenges in representation learning, International Conference on Machine Learning, 2013: 896.
- [47] Iscen A, Tolias G, Avrithis Y, et al. Label Propagation for Deep Semi-Supervised Learning [C]. //IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 5070–5079.
- [48] Arazo E, Ortego D, Albert P, et al. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning [C]. //2020 International Joint Conference on Neural Networks, Glasgow, 2020: 1–8.
- [49] Cascante-Bonilla P, Tan F, Qi Y, et al. Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning [C]. //Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, 2021: 6912–6920.
- [50] Rizve M N, Duarte K, Rawat Y S, et al. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning [C]. //9th International Conference on Learning Representations, Virtual Event, 2021.
- [51] Berthelot D, Carlini N, Goodfellow I J, et al. MixMatch: A Holistic Approach to Semi-Supervised Learning [C]. //Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, Vancouver, 2019: 5050–5060.
- [52] Xie Q, Dai Z, Hovy E H, et al. Unsupervised Data Augmentation for Consistency Training [C]. //Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual, 2020: 6256–6268.
- [53] Sohn K, Berthelot D, Carlini N, et al. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence [C]. //Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual, 2020: 596–608.
- [54] Zhang B, Wang Y, Hou W, et al. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling [C]. //Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, virtual, 2021: 18408–18419.
- [55] Wang Y, Chen H, Heng Q, et al. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning [C]. //The Eleventh International Conference on Learning Representations, Kigali, 2023.
- [56] Wang X, Wu Z, Lian L, et al. Debaised Learning from Naturally Imbalanced Pseudo-Labels [C]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022: 14627–14637.
- [57] Devries T, Taylor G W. Improved Regularization of Convolutional Neural Networks with Cutout [J]. arXiv preprint arXiv:1708.04552, 2017.
- [58] Cubuk E D, Zoph B, Shlens J, et al. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space [C]. //Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual, 2020: 18613–18624.
- [59] Berthelot D, Carlini N, Cubuk E D, et al. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring [J]. arXiv preprint arXiv:1911.09785, 2019.

- [60] Cubuk E D, Zoph B, Mané D, et al. AutoAugment: Learning Augmentation Strategies From Data [C]. //IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 113–123.
- [61] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning [C]. //Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, 2009: 41–48.
- [62] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [C]. //3rd International Conference on Learning Representations, San Diego, 2015.
- [63] He Z, Li C, Zhou F, et al. Rumor Detection on Social Media with Event Augmentations [C]. //44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 2021: 2020–2024.
- [64] Gao Y, Wang X, He X, et al. Rumor detection with self-supervised learning on texts and social graph [J]. *Frontiers Comput. Sci.*, 2023, 17 (4): 174611.
- [65] Sun T, Qian Z, Dong S, et al. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning [C]. //ACM World Wide Web Conference 2022, Virtual Event, 2022: 2789–2797.
- [66] Chen B, Jiang J, Wang X, et al. Debiased Self-Training for Semi-Supervised Learning [C]. //Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, New Orleans, 2022: 32424–32437.
- [67] Beyer L, Zhai X, Oliver A, et al. S4L: Self-Supervised Semi-Supervised Learning [C]. //2019 IEEE/CVF International Conference on Computer Vision, Seoul, 2019: 1476–1485.
- [68] Chen T, Kornblith S, Swersky K, et al. Big Self-Supervised Models are Strong Semi-Supervised Learners [C]. //Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual, 2020: 22243–22255.
- [69] Zeng J, Xie P. Contrastive Self-supervised Learning for Graph Classification [C]. //Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, 2021: 10824–10832.
- [70] Zhang Y, Zhang X, Li J, et al. Semi-Supervised Contrastive Learning With Similarity Co-Calibration [J]. *IEEE Trans. Multim.*, 2023, 25: 1749–1759.
- [71] Wei C, Sohn K, Mellina C, et al. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning [C]. //IEEE Conference on Computer Vision and Pattern Recognition, virtual, 2021: 10857–10866.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

乔禹涵，男，出生于 1997 年 11 月，教育经历如下：

2016.09-2020.06 中国矿业大学（北京） 机电与信息工程学院 信息工程 学士

2020.09-2024.06 北京交通大学 计算机与信息技术学院 计算机科学与技术 硕士

二、发表论文

[1] **乔禹涵**, 贾彩燕, 基于图自监督对比学习的社交媒体谣言检测 [J]. 南京大学学报 (自然科学), 2023,59(05):823-832.

[2] **Yuhan Qiao**, Chaoqun Cui, Yiyang Wang, Caiyan Jia, A Debaised Self-Training Framework with Graph Self-Supervised Pre-training Aided for Semi-Supervised Rumor Detection, *NeuroComputing* (CCF-C Journal) 在审

三、参与科研项目

[1] 2021.1-2021.12, 百度红果基金，面向事件不变特征和传播结构学习的社交平台谣言检测方法研究

[2] 2019.10-2024.12, 国家重点研发计划，“新一代人工智能”重大项目：新一代神经网络模型，批准号：2018AAA0100302

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：乔易涵

签字日期：2024 年 6 月 1 日

学位论文数据集

表 1.1 数据集页

关键词 *		密级 *	中图分类号	UDC	论文资助
谣言检测；图表示；自监督学习；半监督学习；自训练		公开			
学位授予单位名称 *			学位授予单位代码 *	学位类别 *	学位级别 *
北京交通大学			10004	工学	硕士
论文题名 *			并列题名		论文语种 *
基于图表示学习的谣言检测方法研究					中文
作者姓名 *	乔禹涵			学号 *	20120400
培养单位名称 *			培养单位代码 *	培养单位地址	邮编
北京交通大学			10004	北京市海淀区西直门外上园村 3 号	100044
学科专业 *			研究方向 *	学制 *	学位授予年 *
计算机科学与技术			谣言检测	3 年	2024
论文提交日期 *	2024 年 6 月				
导师姓名 *	贾彩燕			职称 *	教授
评阅人	答辩委员会主席 *			答辩委员会成员	
	李晓龙			常冬霞 滕竹 瞿有利 丁春涛	
电子版论文提交格式 文本 (√) 图像 () 视频 () 音频 () 多媒体 () 其他 () 推荐格式：application/msword; application/pdf					
电子版论文出版（发布）者			电子版论文出版（发布）地		权限声明
论文总页数 *	68				
共 33 项，其中带 * 为必填数据，为 21 项。					