情报学报　2023 年 11 月　第 42 卷　第 11 期
Journal of the China Society for Scientific and Technical Information, Nov. 2023, 42(11): 1369-1381

DOI: 10.3772/j.issn.1000-0135.2023.11.010

基于可解释图神经网络模型的社交媒体谣言识别研究

汪子航，言鹏韦，蒋卓人

（浙江大学公共管理学院信息资源管理系，杭州　310058）

摘　要　随着社交媒体数据规模的增长与数据形式的复杂化，社交媒体谣言识别研究面临新的挑战。一方面，谣言传播网络中的复杂结构特征难以被充分挖掘；另一方面，亟须探索基于深度神经网络的谣言识别模型的可解释性。本文设计和实现了具备可解释性的图神经网络模型应用于谣言识别任务。具体而言，本文在运用残差图卷积神经网络模型进行谣言识别的基础上，进一步训练基于掩码学习的图神经网络解释器，不仅将谣言传播网络结构特征纳入识别模型，而且从传播网络结构和传播节点属性两个视角对图神经网络模型自动生成解释。本文基于新浪微博（中文）和推特（英文）来源的两个网络谣言数据集进行实验，并从全局与个案两个层面进行解释性分析。研究结果显示，本文提出的图神经网络模型可以有效利用谣言传播网络结构特征，在谣言识别任务中的表现超过了一系列对照组模型。结合图神经网络解释器生成的解释可以发现，在较大规模的谣言传播树中，长传播链条是谣言的关键网络拓扑结构；在规模较小的谣言传播树中，文本特征是关键的节点属性。

关键词　社交网络；谣言识别；图神经网络；模型可解释性

Interpretable Graph Neural Network for Social Media Rumor Detection

Wang Zihang, Yan Pengwei and Jiang Zhuoren

(*Department of Information Resources Management, School of Public Affairs, Zhejiang University, Hangzhou*　310058)

Abstract: The increasing scale of social media data and data heterogeneity pose new challenges. In contrast, complex structural features in rumor propagation networks are difficult to explore; however, the interpretability of a deep neural network-based rumor detection model must be further investigated. In this study, we design and implement an interpretable graph neural network model for a rumor detection task. Specifically, we train a graph neural network interpreter based on mask learning using a residual graph neural network model. This framework incorporates the structural features of rumor propagation and provides an automatic interpretation of the graph neural network model, considering both network structure and node features. The experiments are conducted on two online rumor datasets sourced from Sina Weibo (Chinese) and Twitter (English). Interpretive analyses are performed at both the global and case levels. The experimental results show that the proposed graph neural network model effectively exploits the communication structure features and outperforms a series of baseline models in the rumor detection task. Using the trained graph neural network interpreter, we discovered

收稿日期：2022-08-29；**修回日期**：2023-05-10

基金项目：国家自然科学基金青年科学基金项目“基于图神经网络的政策舆情演化机制研究”（72104212）；浙江省自然科学基金项目“基于深度表示学习的政策舆情动态画像研究”（LY22G030002）；中央高校基本科研业务费专项资金资助项目。

作者简介：汪子航，男，2000年生，硕士研究生，主要研究方向为计算社会科学；言鹏韦，男，1996年生，博士研究生，主要研究方向为社交网络、图深度学习、可解释学习；蒋卓人，通信作者，男，1986年生，博士，百人计划研究员，博士生导师，主要研究方向为社交媒体挖掘、自然语言处理、计算社会科学等，E-mail：jiangzhuoren@zju.edu.cn。

that long propagation chains are the key network topology for rumors in larger-scale rumor propagation trees and text features are the key node attributes in smaller-scale rumor propagation trees.

Keywords: social networks; rumor detection; graph neural networks; model interpretability

0 引 言

当前，社交媒体已经逐渐取代电视、报纸等传统媒体，成为人们日常生活中接收信息的重要途径[1]。然而，社交媒体的开放属性也为谣言的产生和传播创造了条件。谣言，即“信息流传中的尚未得到证实的陈述”[2]，这些未经验证的信息可能会引起社会公众的怀疑或焦虑[3]，推动虚假信息传播或导致舆论问题。此外，社交媒体数据规模日益增长、数据形式日趋复杂，谣言信息在社交媒体中以信息传播树的形式层层扩散并大量传播。因此，高效、准确地识别网络谣言并针对谣言传播特征进行治理是亟待解决的研究问题。

然而，目前社交媒体谣言识别的研究仍然面临两大挑战。其一，谣言传播网络中的复杂结构特征难以被充分挖掘。目前自动化谣言识别重点关注的是传播内容、传播用户等特征[4-9]，而对于传播结构的特征关注较少。其二，随着深度神经网络为代表的复杂模型在谣言识别任务上的应用，模型可解释性的需求也日益高涨。由于深度神经网络等复杂模型的推理过程通常是一个“黑箱”过程，在社会实践中使用黑盒模型可能引发算法歧视等社会问题。相关法律法规（如欧盟的《通用数据保护条例》）对机器学习算法的可解释性也提出了明确要求。因此，社交媒体谣言识别算法急需可解释性分析，以提升其可信度。

针对上述挑战，本文设计并实现了一个可解释图神经网络谣言识别模型。该模型利用图神经网络对谣言数据进行识别，同时从网络结构和节点特征两个角度对模型决策进行解释分析。具体而言，首先，本文基于社交媒体的用户信息和消息传播结构等构建谣言数据集，将用户信息，如用户粉丝数等作为传播节点特征；将用户间的交互行为，如转发等作为传播结构中的边，构建包含多维节点特征的信息传播树。其次，利用残差图卷积神经网络模型，同时对信息传播树中的节点信息与结构信息进行学习，以提高谣言识别的准确率。最后，通过训练图神经网络解释器，利用基于网络结构掩码的学习与基于节点特征掩码的学习识别模型决策中对模型预测重要的网络结构和节点特征，并生成模型解释。

本文使用源自新浪微博和推特的谣言数据集进行模型实验验证与可解释分析。模型验证结果表明：①本文采用的残差图卷积神经网络模型在两个数据集的所有评价指标上均超过了对照组模型，证明了该模型的有效性和泛化性；②本文采用的模型仅使用谣言传播的结构特征也可以对谣言进行准确识别，验证了谣言传播结构特征的重要性和图神经网络模型的鲁棒性。结合图神经网络模型解释器生成的解释可以发现：①非谣言数据的传播结构较为扁平，而谣言数据的传播结构相对纵深较长，在传播路径上往往产生多级转发或评论；②对于复杂的传播网络，图神经网络模型可以充分学习用户互动中的结构信息，取得良好的预测结果；对于简单的传播网络，其传播结构特征较弱，需要结合更丰富的节点特征进行谣言识别。

本文的主要贡献：第一，本文从模型可解释性出发，提出了可解释图神经网络谣言识别模型，不仅能够对谣言进行精准识别，而且能够结合图神经网络解释器对谣言传播机理进行分析。本文既增强了深度学习模型的可信度，也在谣言识别领域对信息行为和信息价值的理论进行了全新的探索。第二，与已有研究侧重于传播内容等特征进行谣言识别相比，本文将谣言传播结构特征纳入了谣言识别模型，以提高识别准确率。同时，利用图神经网络模型对谣言传播结构进行可解释分析，更好地总结了谣言传播行为的规律和模式。第三，在中英文两个数据集上，本文综合全局和案例分析得出谣言的拓扑结构特征和节点特征，并基于发现提出谣言治理的策略建议，对于谣言鉴别和舆情治理具有积极的实践意义。

1 相关研究

1.1 社交媒体谣言识别研究

谣言的本质属性在于未经验证[2-4]，本文将社交媒体谣言定义为在社交媒体平台上广泛传播的、未经验证的信息陈述。传统的谣言识别研究通常将其转化为一个基于特征挖掘的分类问题。谣言的传播过程形成了从信息源到不同受众逐级传播

的信息传播树，其中包含了谣言信息本身的内容特征、传播路径中涉及的用户特征以及由传播中用户间互动产生的序列特征与结构特征。早期的谣言识别主要采用传统机器学习算法，关注对文本特征与时序特征的挖掘[4-6,10-12]，重点在于通过特征挖掘提高谣言识别模型的准确率。Afroz 等[10]利用词汇、句法和内容特定的特征，在众多数据集上的谣言检测任务中取得了较好的效果。Ma 等[12]使用动态时间序列的结构从信息传播的过程中抽取时间属性，以观察谣言在传播过程中随时间的变化。此外，部分研究通过统计的方式对信息传播特征进行提取，初步验证了传播特征对谣言识别的重要性[6,11]。

近年来，深度学习被广泛应用于谣言识别中。与传统的机器学习方法相比，深度学习方法可以依靠复杂的模型结构对文本、图像等信息进行特征挖掘与提取。这一研究方法的转变，使问题从对输入特征的建模转变成设计一种有效解决谣言识别任务的网络结构。例如，Yu 等[8]与 Wang 等[9]分别利用 CNN（convolutional neural network）与 LSTM（long short-term memory）加强了对文本信息提取的设计。

上述谣言识别方法关注的重点为谣言传播内容、传播用户和传播时序的特征，但对于谣言传播网络的结构特征，已有研究难以直接对传播结构进行建模与分析，而图神经网络的发展为此提供了新的研究可能[13]。图卷积网络（graph convolution network，GCN）是一种专门针对图数据进行空间特征提取的卷积神经网络[14]，包括 GraphSAGE（graph sample and aggregate）[15]、GAT（graph attention network）[16]、PGC（partition graph convolution）[17]等。近年来，研究者开始探索图神经网络在谣言识别中的运用。例如，Bian 等[18]将会话表示为一个有向树的结构，使用图卷积网络，分别自上而下和自下而上对会话树进行编码，用于学习谣言传播结构的特征；Bai 等[19]基于原始信息和回复构建图数据，建立了具有节点比例分配机制的集成图卷积网络；王昕岩等[20]采用边权重来描述事件之间联系的紧密程度，提出了一种基于加权图卷积神经网络模型，对新浪微博中的谣言进行检测。

本文旨在结合谣言内容、用户属性以及谣言传播网络结构等多维度的特征对谣言识别进行可解释性分析，利用图神经网络方法进行研究。

1.2　机器学习可解释性

近年来，机器学习技术在谣言识别场景的应用提高了谣言识别任务的准确率，但模型决策过程缺乏可解释性，限制了其在诸多场景的应用，因此，越来越多的研究者开始探索机器学习的可解释性。

一方面，对于非图结构的神经网络，目前常用的一种解释方法是基于代理模型的方法，如 LIME（local interpretable model-agnostic explanations）[21]模型用一个简单的线性模型作为代理，对一个复杂的黑箱模型的局部区域进行近似。另一方面，许多研究使用基于梯度的反向传播来计算输入的重要性，如类激活映射（class activation mapping，CAM）等[22-26]，以及应用反事实推理的方法 LEWIS[27]等，这些方法的关键思想是将梯度作为输入重要性的近似值。

社交网络是一种图结构。近年来，对图神经网络的解释工作的研究日益增加。Pope 等[28]将最终的节点嵌入映射到输入空间，利用原始 GNN（graph neural network）的网络参数和分类器的输出计算输入的重要性，将基于梯度的方法扩展到 GNN 中，用于解释图分类模型。Huang 等[29]将 LIME 扩展到图模型中，采用非线性代理模型的重要特征来近似对 GNN 的解释。Vu 等[30]通过随机扰动的方法获取局部的数据集，借助可解释的贝叶斯网络来拟合该数据集，可以同时用于解释节点分类和图分类任务。Ying 等[31]和 Luo 等[32]分别提出了基于掩码学习的 GNNExplainer 与 PGExplainer，对于给定一个训练好的 GNN 模型和一个预测结果，掩码学习可以用于识别一个小的子图结构和一个节点特征的子集，使整个输入图中 GNN 预测的互信息最大化，可以同时识别对预测重要的网络结构和节点特征。

2　可解释图神经网络谣言识别框架

本文提出的可解释图神经网络谣言识别框架包含两个重要组成部分：谣言传播树网络结构与节点特征预处理模块（2.1 节）和可解释图神经网络谣言识别模型（2.2 节）。其中，网络结构与节点特征预处理模块主要负责网络谣言原始数据的预处理工作，可解释图神经网络谣言识别模型由残差图卷积神经网络模型（2.2.1 节）和基于掩码学习的图神经网络解释器（2.2.2 节）两大模块组成。该框架的输

入是谣言传播树原始数据，通过预处理后，可解释图神经网络谣言识别模型可通过残差图卷积神经网络模型对信息是否为谣言进行预测，而图神经网络解释器通过训练可以分别生成基于网络结构的解释和基于节点特征的解释。总体而言，本文提出的识别框架如图 1 所示。

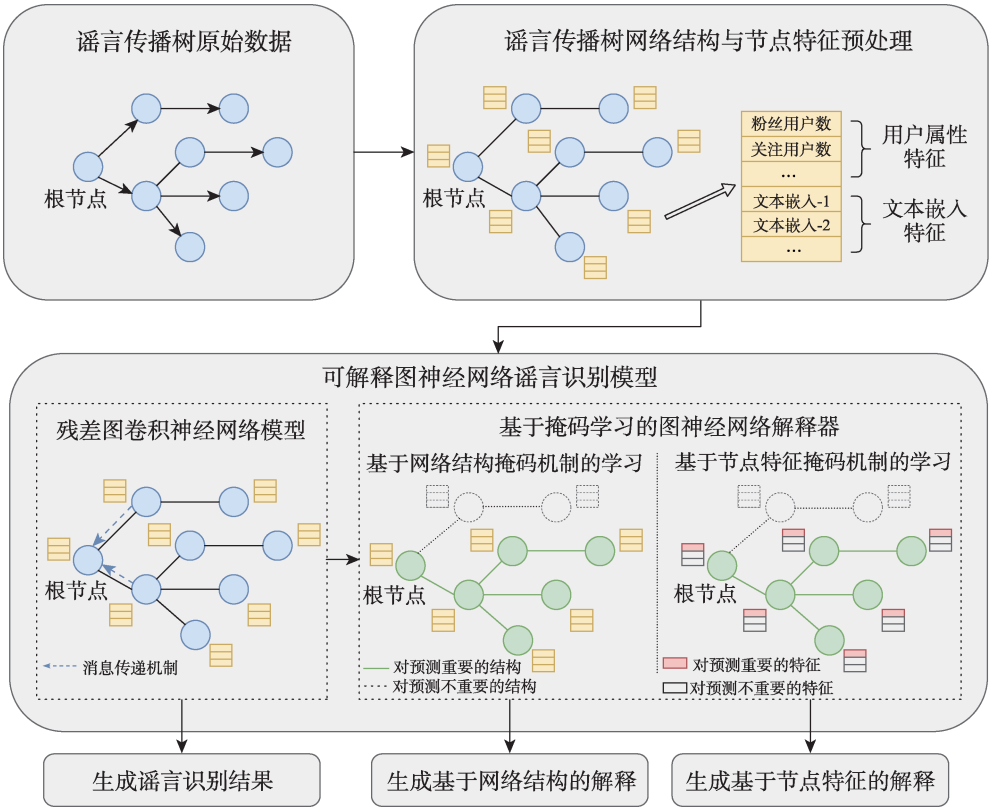


图 1 可解释图神经网络谣言识别框架

2.1 谣言传播树网络结构与节点特征预处理

在一个信息的传播树中，以原始信息为根节点，每一次评论/转发作为一个节点，使用粉丝用户数等用户属性特征和基于信息文本内容的文本嵌入特征等作为节点特征，根据信息的转发/评论关系，建立由被转发/评论者指向转发/评论者的边，构建谣言传播树网络结构数据。

对于数据集中的第 i 个事件，用 y_i 表示其对应的标签，$Y=\{y_1,y_2,y_3,\cdots,y_n\}\in R^{n\times 2}$ 表示数据集标签的集合。根据信息的转发关系，构建信息的传播结构 $G=\langle V,E\rangle$。其中，$V=\{v_1,v_2,v_3,\cdots,v_n\}$ 表示图中所有的节点，$E=\{e_1,e_2,e_3,\cdots,e_n\}$ 表示图中所有的边。$A\in R^{n_i\times n_i}$ 代表 c_i 的邻接矩阵，表示节点间的连接情况，若节点 i 和节点 j 之间有连接，则 $A_{ij}=1$，否则，$A_{ij}=0$。$X\in R^{n_i\times m}$ 代表数据集的特征矩阵，其中 m 为特征的维度。

2.2 可解释图神经网络谣言识别模型

2.2.1 残差图卷积神经网络模型

本文设计了一个基于残差图卷积神经网络的谣言识别模型。该模型的输入为以事件为单位的信息传播树，输出为该事件的预测标签。其核心思想是使用残差图卷积网络抽取节点的邻居节点的信息来更新每一层隐藏层的信息，从而整合谣言的扩散过程中的关键结构信息，获取社交网络结构的深层表征。第 k 个图卷积层的隐藏特征矩阵 H_k 的计算过程为

$$H_k=M(A,H_{k-1},W_{k-1}) \tag{1}$$

其中，W_{k-1} 表示可训练参数。第 k 层的隐藏特征被聚合后再通过信息传播函数 M 得到下一层的隐藏特征，当 $k=1$ 时，隐藏特征即节点原始特征。基于此，本文为每一层添加了残差连接，作为残差图卷积网络（ResGCN）[33]模型。该方法已经被验证能够有效地避免梯度消失，可以在深层次上获得更

好的聚合能力，即

$$H_k = R(H_{k-1}, W_{k-1}) = \rho(\hat{A}H_{k-1}W_{k-1}) + H_{k-1} \quad (2)$$

残差图卷积神经网络模型结构如图 2 所示。为了避免过拟合，在 Readout 层使用 Dropout 策略。

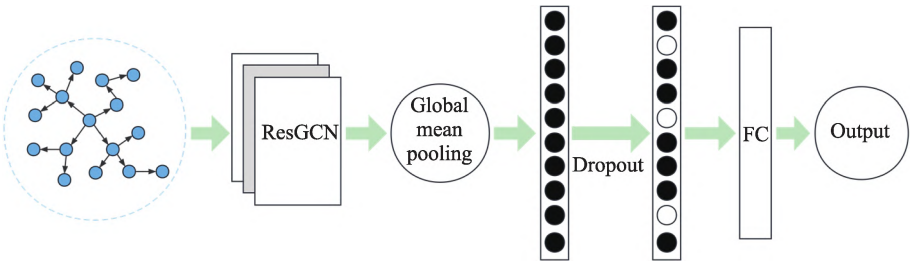


图 2　残差图卷积神经网络模型结构

将图数据输入图卷积网络，可得到输出的隐藏特征矩阵 $H = \{h_1, h_2, h_3, \cdots, h_n\}$，使用全局平均池化操作来聚合该表征的信息。通过全连接层和 Softmax 层计算事件 i 的预测标签 $\hat{y}$，即

$$\hat{y} = \text{Softmax}(\text{FC}(\text{MEAN}(H))) \quad (3)$$

其中，$\hat{y}$ 是一个二维向量，表示预测为谣言和非谣言的概率。通过最小化预测标签 $\hat{y}$ 和真实标签 y 的交叉熵来训练模型参数。

2.2.2　基于掩码学习的图神经网络解释器

本文采用基于掩码学习的图神经网络解释器[31]对残差图卷积神经网络的预测进行解释，通过改变输入模型的信息、对比模型结果，来监测被改变的信息的重要性，从而确定边和特征对决策的重要程度，如图 3 所示。具体来说，给定一个训练好的图神经网络模型和一个预测结果，该图神经网络解释器可以通过学习边和节点特征的掩码来预测解释结果，从而识别一个小的子图结构 $G_S \subseteq G$ 和一个节点特征的子集 $X_S \subseteq X$，通过使整个输入的原始图的预测以及通过掩码获得的新的 GNN 预测的互信息 MI 最大化来优化掩码，即

$$\max \text{MI}(Y, (G_S, X_S)) = H(Y) - H(Y|G = G_S, X = X_S) \quad (4)$$

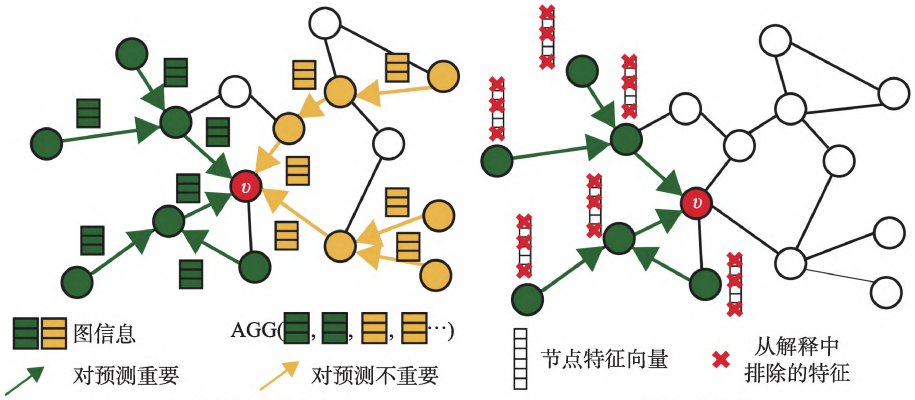


图 3　图神经网络解释器工作原理[31]

3　实验数据

3.1　数据来源与预处理

新浪微博和推特（Twitter）分别是中文领域和世界范围内两大重要的社交媒体平台，具有用户量多、曝光度大、信息传播快的特点，容易成为谣言传播的重灾区。为了验证本文提出的可解释性图神经网络模型的有效性和鲁棒性，选择来自上述两个平台的公开数据集进行实验。Chinese_Rumor_Dataset 数据集（https://github.com/thunlp/Chinese_Rumor_Dataset）是依据新浪微博社区管理中心公示内容中的谣言微博进行收集的，本文使用的是该数据集的第二部分，即 CED_Dataset 数据集[34]，该数据集包含事件的原始微博及其相应的转发/评论微博。本文通过爬虫进一步获取数据集中相关用户的粉丝数、博文发布数等用户属性信息。PHEME 数据集[35]对“弗格森骚乱”“《查理周刊》枪击事件”等 9 个突

发新闻事件相关推文数据进行收集，具体包含原始推文及其回复，以及相关用户的属性信息。

按照2.1节谣言传播树网络结构与节点特征预处理模块，根据交互关系对上述两各数据集构建传播网络数据。每个传播网络对应一个二分类的标签，表示谣言或非谣言。剔除数据缺失的事件，使用Python的PyTorch Geometric库构建图结构数据集，其中CED_Dataset包含3300个图结构，PHEME包含6425个图结构。两数据集的标签分布情况如表1所示。

表1 实验数据集标签分布情况

数据集	谣言数量	非谣言数量	谣言占比（%）	总数
CED_Dataset数据集	1451	1849	44.0	3300
PHEME数据集	2402	4023	37.3	6425

3.2 特征选择

Wellman[36]认为，人们之间的互动和交流形成了社会关系，使得人们能够学习他人的意见。因此，人们传播消息很大程度上受到其社交关系的影响。本文选择粉丝用户数、关注用户数、博文发布数作为用户特征，用于衡量一名用户的社会关系的复杂程度和强弱程度。

在本文所研究的数据集中，CED_Dataset数据集中传播树节点更多，即谣言事件本身的转发或评论等交互信息更丰富，传播结构较为复杂；PHEME数据集传播结构相对较为简单。图4展示了两个数据集的传播树的节点数量分布情况。

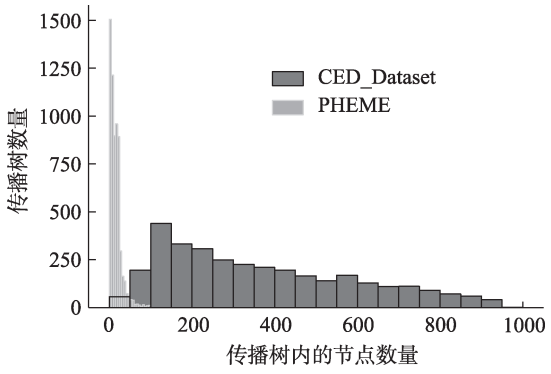


图4 数据集传播树节点数量分布直方图

从理论角度来看，根据Bai等[19]的研究结果，当传播树的结构较为简单时，谣言和非谣言之间的图的全局结构特征可能难以区分。因此，对于简单的传播树，文本特征对于谣言检测更为重要。从实际数据角度来看，在微博的信息传播网络结构中，大量用户仅仅转发而不做评论，不存在文本内容。

基于上述两点原因，本文对CED_Dataset数据集和PHEME数据集采用不同的特征选择策略。在CED_Dataset数据集中，使用粉丝用户数、关注用户数和博文发布数3个用户属性信息作为节点特征。PHEME数据集中的传播树结构更为简单，只依靠用户信息难以对是否为谣言进行判断。因此，PHEME数据集中节点特征除了用户的关注用户数、粉丝用户数和博文发布数特征以外，还包含了每个节点所包含的文本信息。对于节点的文本信息，本文使用Sentence-BERT（bidirectional encoder representations from transformers）[37]获取其向量表示，将文本转换为768维向量，并将其作为节点特征的一部分。两个数据集的具体使用特征信息如表2所示。

表2 CED_Dataset数据集和PHEME数据集特征选择

特征	CED_Dataset数据集	PHEME数据集
粉丝用户数	√	√
关注用户数	√	√
博文发布数	√	√
文本向量	×	√
特征维度	3	771

4 结果与分析

基于预处理的数据集，本文从3个方面来全面验证可解释图神经网络谣言识别模型：①网络谣言识别的效果（4.2节）；②基于传播结构的可解释性分析（4.3节）；③基于节点特征的可解释性分析（4.4节）。

4.1 实验设置

本文的实验在Windows 10操作系统和Python 3.7环境下进行。采用Pytorch Geometric框架构建谣言识别模型，并进行模型训练。模型选择的优化器为Adam，损失函数为交叉熵，分类层激活函数为Sigmoid。超参数设置如表3所示。

本节使用训练后的图神经网络模型对数据集进行预测。对于预测的结果，使用准确率（accuracy）、精确率（precision）、召回率（recall）和F1值（F1 score）来衡量其准确性。

本文设置了5组实验作为对照组，分别检验不

表3　超参数设置

参数	参数值
训练集占比	50%
Epoch	200
Dropout	0.5
学习率	0.001
Batch_size	32
Hidden_channels	64

同情况下的实验结果以及本文模型的性能。各组实验具体如下。

（1）支持向量机[38]：支持向量机是一种二分类算法，其目标是得到最好的超平面，对数据进行二元分类。经过扩展，支持向量机也能应用于回归问题。

（2）随机森林[39]：随机森林是Bagging集成算法的一种，多颗随机采样变量的弱分类器组成了随机森林。随机森林的所有基评估器均是决策树，模型的最终分类结果由子树进行投票得到。

（3）XGBoost[40]：XGBoost是一种集成算法，通过优化结构化损失函数来生成弱学习器，纠正前面所有弱学习器的残差，最终多个学习器共同用于预测。

（4）朴素贝叶斯[41]：朴素贝叶斯方法是一种基于贝叶斯定理和特征条件独立假设的分类方法。

（5）GCN[14]：GCN是基于不使用残差连接的朴素图操作的图卷积网络模型。

4.2　网络谣言识别实验结果

在实验中，使用十折交叉验证方法，实验结果如表4所示。比较表4中不同模型的表现可以看出，本文设计的基于残差图卷积神经网络的谣言识别模型在两个数据集上的所有评价指标均优于其他对照组模型，这充分证明了该模型的有效性和泛化性。在CED_Dataset数据集上，GCN模型超过了上述机器学习算法的性能；在PHEME数据集上，GCN模型的性能有所欠缺。而残差图卷积神经网络模型则分别在两个数据集上大幅度地超过了上述所有算法模型的准确性。此外，残差图卷积神经网络模型在仅使用少量特征的情况下仍然能够得到较高的准确性，从而验证了该模型的鲁棒性。

4.3　基于传播结构的可解释性分析

图神经网络解释器通过学习传播结构中边的掩码并为边生成权重，表示其对预测结果的影响程度。如上文所述，CED_Dataset数据集中谣言信息扩散范围较广，传播树结构丰富；PHEME数据集中谣言信息的扩散范围较小，传播结构相对简单。因此，本节主要基于CED_Dataset数据集解释结构特征在谣言识别中作用，下文分别从案例和全局层面对模型生成的解释进行阐述。

表4　谣言识别结果

模型	数据集	准确率	精确率	召回率	F1值
SVM（support vector machine）	CED_Dataset	0.735	0.657	0.836	0.735
	PHEME	0.637	0.586	0.010	0.169
XGBoost	CED_Dataset	0.770	0.737	0.745	0.740
	PHEME	0.872	0.845	0.804	0.824
随机森林	CED_Dataset	0.778	0.745	0.753	0.748
	PHEME	0.853	0.851	0.735	0.789
朴素贝叶斯	CED_Dataset	0.567	0.035	0.712	0.018
	PHEME	0.776	0.671	0.783	0.723
GCN	CED_Dataset	0.798	0.793	0.730	0.760
	PHEME	0.741	0.659	0.634	0.646
残差图卷积神经网络	CED_Dataset	**0.853**	**0.828**	**0.839**	**0.834**
	PHEME	**0.901**	**0.894**	**0.832**	**0.862**

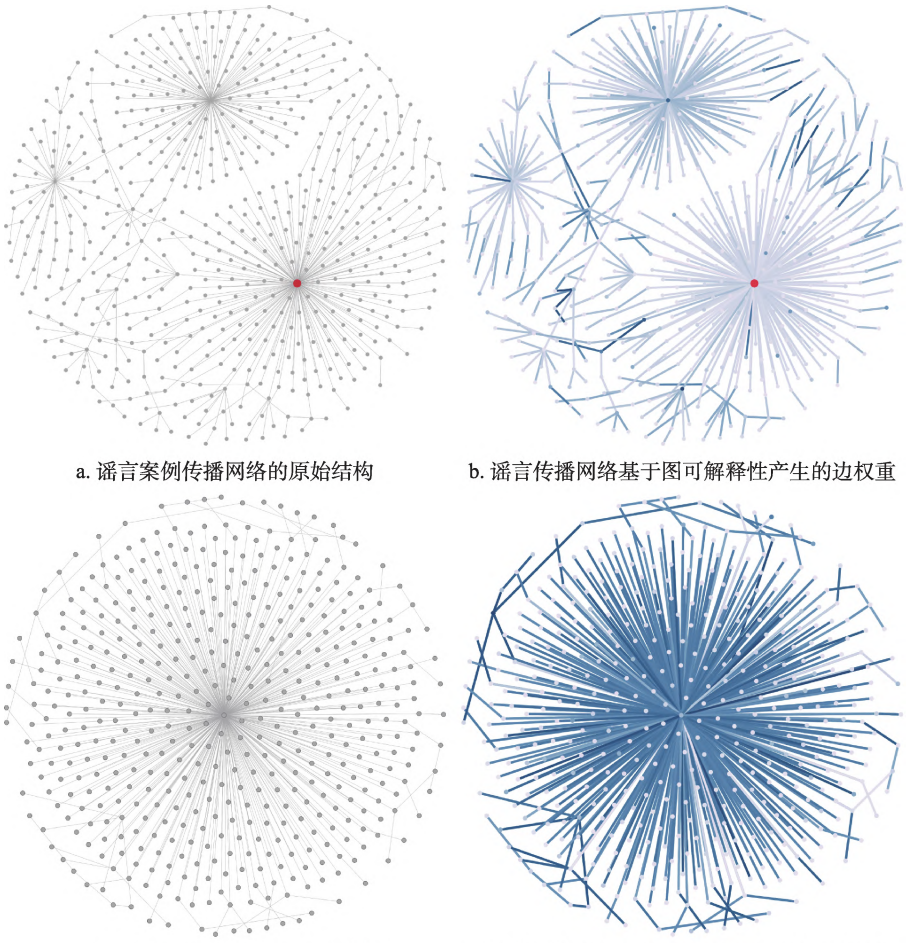
注：粗体表示最优结果。残差图卷积神经网络是本文使用的方法。

4.3.1　基于案例的传播结构可解释性分析

抽取CED_Dataset数据集中节点较多、结构相对复杂的数据进行案例分析，选取其中一则关于“火车盒饭价格”的谣言数据与关于“下架日系产品”的非谣言数据进行对比。图5展示了这两个样本的原始网络结构以及标记了边权重的网络结构。图5b和图5d用颜色的深浅来表示边权重的大小，连边颜色越深，表示其权重越大，代表其在谣言识别模型的决策过程中越重要。

由原始传播网络结构（图5a和图5c）可以看出，谣言案例的传播树呈现双中心的结构，两个中心的一层传播节点数目分别占整体节点总数的32.4%与22.4%。在非谣言的信息传播树中，非谣言案例呈现以根节点为中心的单中心结构，其一层传播节点数占整体节点数的86.6%，信息传播的核心节点相对单一。此外，结合表5可知，案例中的谣言网络传播链平均长度是非谣言的2倍，说明其传播深度比非谣言案例更深。

由图5b和图5d可以看出，在谣言案例中，两个中心节点的一层传播并没有被赋予较高的权重，相反地，模型更多地关注长传播链条的末段传播部分；在非谣言的传播结构的解释中，整体边的权重未呈现显著差异。



a. 谣言案例传播网络的原始结构　b. 谣言传播网络基于图可解释性产生的边权重

c. 非谣言案例传播网络的原始结构　d. 非谣言传播网络基于图可解释性产生的边权重

图 5　CED_Dataset 典型案例——传播结构

表 5　CED_Dataset 案例——传播链平均长度

案例	传播链平均长度	传播链最大长度
谣言	2.31	7
非谣言	1.16	6

结合实际传播过程中的具体讨论内容与用户进行分析发现，在初始传播阶段，即一级转发/评论中，用户通常是基于自身经历对事件进行客观讨论；随着传播继续进行，类似“欺骗消费者[抓狂]”的讨论内容增加，讨论情绪倾向明显增强。结合对案例中不同传播层级的用户属性进行分析。表 6 展示了一至三级转发/评论层次下的节点的用户特征的平均情况，表明随着转发层次的深入，用户的粉丝用户数、关注用户数、博文发布数逐渐减少，尤其是粉丝数显著下降。

上述分析说明，随着转发层次的深入，转发/评论者更多为在网络中影响力较小的普通用户，这些用户对信息的甄别和筛选能力相对较弱[42]，面对不确定性的谣言事件，容易引发不同立场并产生递进式的讨论，使得谣言事件具有相对更深的传播结构。因此，较深的传播结构反映了该微博的不确定性与争议性。而模型通过关注这些传播链较长的传播结构，有助于对谣言进行判别。

表 6　CED_Dataset 谣言案例——不同转发层次节点特征均值

传播层级	粉丝用户数	关注用户数	博文发布数
一级转发/评论	32210.76	7522.48	12183.40
二级转发/评论	6348.67	950.42	8921.49
三级转发/评论	4535.26	809.60	8043.26

4.3.2　基于全局的传播结构可解释性分析

为进一步探索普遍性的规律，本文结合社交网络的属性，分别对谣言和非谣言数据进行全局统计分析。对谣言与非谣言传播树的节点数量、一级转发/评论占比、近似传播深度（以图数据中最大传播长度/节点总数来近似事件传播的深度）的平均

值进行统计与可视化，如图 6 所示。

图 6a 展示了一级转发/评论在传播网络中的占比情况，非谣言数据的一级转发/评论占比呈现左偏分布，说明针对非谣言事件的讨论主要直接发生在与原始信息之间，而谣言事件中直接转发/评论占比分布相对均衡。在整体上，非谣言事件的直接转发/评论占比大于谣言事件。图 6b 对比了谣言与非谣言传播树中信息传播深度的分布情况，相比于非谣言事件，谣言事件的传播层次更深。

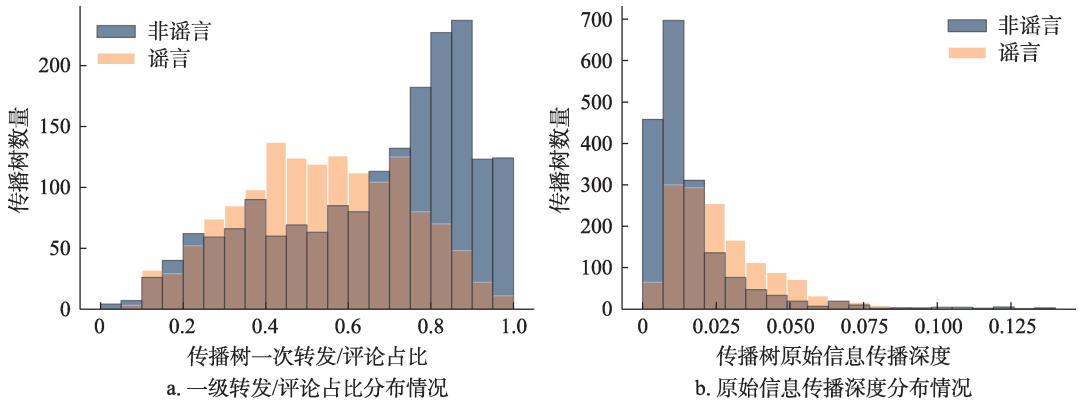


图 6　CED_Dataset 数据集数据分布直方图(彩图请见 https://qbxb.istic.ac.cn/)

总体而言，谣言事件更容易引发用户间的深层次传播。上述全局统计情况验证了图神经网络解释器在具体案例中生成的解释结论。Prasad[43]认为，谣言是一种群体反应，具有情绪化的倾向。这在熟人之间的网络中更加突出。刘于思等[44]的研究表明，熟人关系居多的社交网络提升了人们之间的依赖度和信赖度，人们更倾向于彼此分享信息。因此，当谣言事件出现在用户的社交网络中时，用户往往倾向于通过转发以分享情绪，在累次转发的过程中，随着转发层次的递增，情绪化现象变得越发严重，从而进一步加剧了用户对谣言的转发。

4.4　基于节点特征的可解释性分析

如 2.2 节所述，图神经网络解释器不仅可以通过边权重的学习生成网络结构的解释，也可以通过学习节点特征的掩码得出节点特征的权重。某一节点特征的权重越高，说明在这个网络中，该节点特征的改变对预测结果产生的影响越大。相较于 CED_Dataset，PHEME 数据集拥有更丰富的节点特征，节点特征除了包含用户的关注用户数、粉丝用户数和博文发布数外，还纳入了每个节点所包含的文本特征信息。因此，本节主要基于 PHEME 数据集，通过节点特征的重要程度来解释节点特征在谣言识别中作用。

4.4.1　基于案例的节点特征可解释性分析

选取 PHEME 数据集中“悉尼人质”事件下的谣言数据进行分析。“悉尼人质”事件是指 2014 年 12 月 15 日澳大利亚悉尼市中心一家咖啡馆发生的人质劫持事件。下文结合一则事件中关于“炸弹布置”的谣言数据与一则关于“警方行动”的非谣言数据的节点特征进行可解释分析。

图 7 以谣言传播树的形式对比了谣言案例和非谣言案例的部分节点特征：①节点颜色深浅表示节点的博文发布数，颜色越深表示数量越多；②节点大小表示节点粉丝用户数，节点越大表示数量越多；③节点的文本嵌入通常包含情感立场等语义信息，因此结合文本的立场进行案例分析；④用边的颜色表示子节点对其父节点的立场，可以划分为认同、不认同、中立或无关三类，分别标记为蓝色、红色与黑色。在图 7a 中，谣言案例的根节点和 12 号节点的博文发布数显著大于其他节点，并且 12 号节点对根节点持不认同的态度；而在图 7b 中，仅有根节点的博文发布数显著大于其他节点。博文发布数代表了用户的在社交平台上的活跃度。从案例数据上看，在谣言案例的传播中出现了高活跃度节点的对立；在非谣言案例中，整个传播结构由单个观点输出者主导。

通过图神经网络解释器可以得出案例中不同特征对预测重要性的权重，其中，文本嵌入特征的权重由 768 维文本嵌入权重的和得到，作为整体文本嵌入特征的重要性表示。对全部 4 个特征的重要性进行 0-1 标准化，得出谣言和非谣言案例中各个特征的相对重要程度，可以发现文本嵌入特征相对重要程度超过 99%，说明模型对案例是否为谣言的预测几乎全部依赖于文本信息。

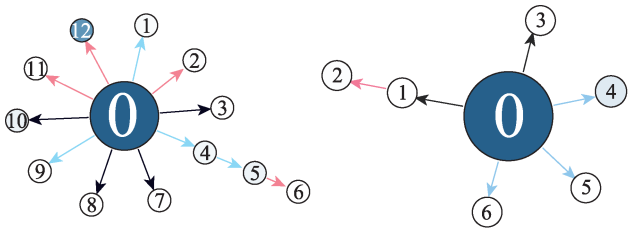


图 7 PHEME 案例传播树结构
（彩图请见 https://qbxb.istic.ac.cn）

图 8 展示了谣言案例中的文本向量的情况，本文使用 *t*-SNE（*t*-distributed stochastic neighbor embedding）[45]对其进行降维，将 768 维向量转化为 32 维向量，并采用热图分组展示图中 13 个节点的文本向量。文本向量由上到下包括谣言原文、持认同立场的文本、中立或无关立场的文本以及持不认同立场的文本。如图 8 所示，不同内容、立场的文本在部分维度上存在显著差异，图中使用方框标记了这些维度，说明文本向量包含了丰富的语义信息，对谣言识别有重要意义。

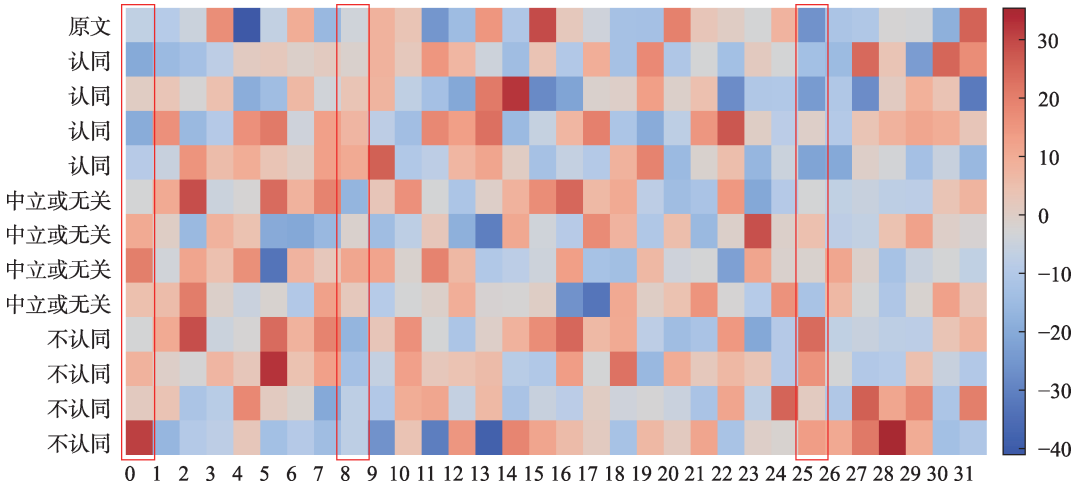


图 8 PHEME 谣言案例——文本向量热力图

4.4.2 基于全局的节点特征可解释性分析

为了进一步探索普遍性的规律，参照上述案例分析对特征权重的处理对 PHEME 整体数据集进行统计，得出谣言和非谣言数据中各个特征的相对重要程度，具体如表 7 所示。就 PHEME 数据集而言，谣言识别模型在进行预测时很大程度上依赖于传播中的文本信息。其原因可能是，对于简单的传播网络，由于互动者较少，谣言和非谣言之间的网络结构特征相似，文本特征对于谣言的识别更为重要。

表 7 PHEME 数据集标准化重要性权重

类别	特征	标准化重要性权重
谣言	粉丝用户数	0.0001
	关注用户数	0.0007
	博文发布数	0.0005
	文本嵌入特征	0.9987
非谣言	粉丝用户数	0.0001
	关注用户数	0.0002
	博文发布数	0.0005
	文本嵌入特征	0.9992

利用 KL 散度（Kullback-Leibler divergence）[46]验证文本特征在谣言与非谣言上分布的差异性。经计算，谣言的文本特征各维度重要性分布与非谣言的文本特征各维度重要性的 KL 散度为 0.342，说明谣言和非谣言的文本嵌入特征值在重要性分布上存在显著差异。

此外，对谣言识别中的用户属性特征进行统计分析并做可视化处理，图 9a~图 9c 所示为谣言数据，图 9d~图 9f 所示为非谣言数据；图中对比了 PHEME 数据集中谣言与非谣言传播树中用户平均粉丝用户数、平均关注用户数、平均博文发布数的分布情况。可以看出，谣言数据中的平均用户粉丝用户数、平均关注用户数、平均博文发布数均少于非谣言数据，这说明谣言数据中涉及的用户在社交平台上的影响力与活跃度均较低。另外，谣言与非谣言数据在用户属性特征上的分布相似，削弱了用户属性特征在谣言识别模型预测中的重要性。

4.5 小 结

本节通过图神经网络解释器的学习分别生成基

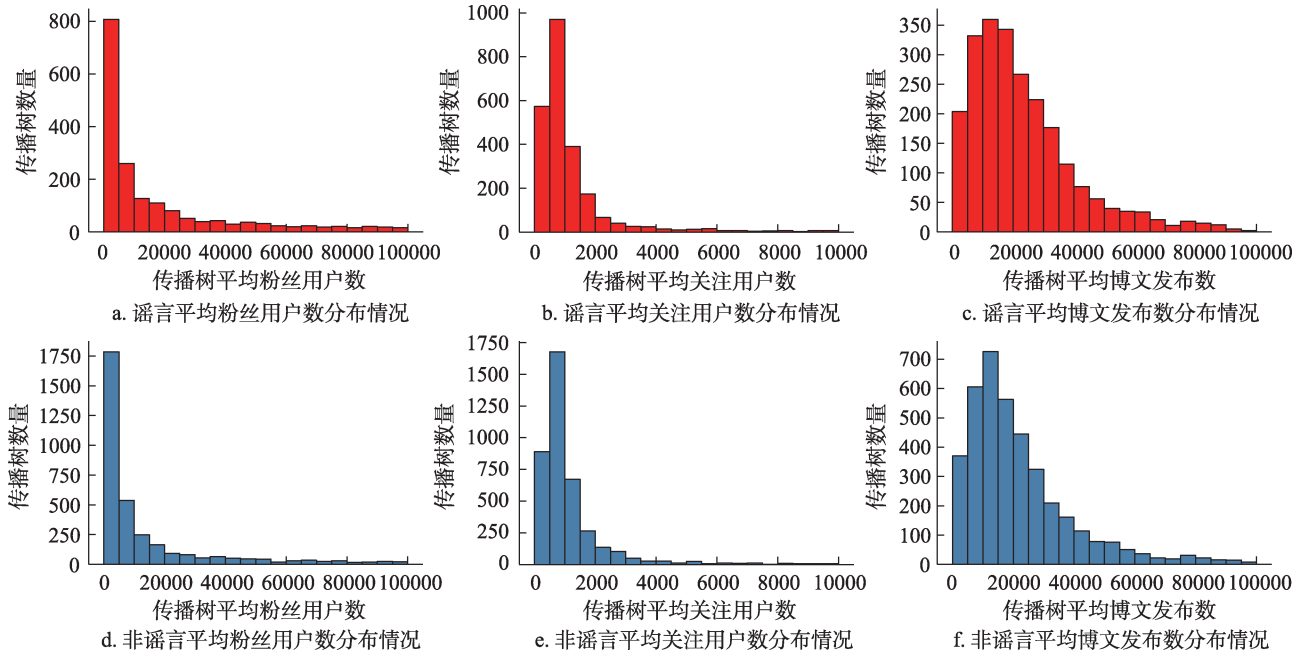


图 9　PHEME 数据集平均用户数据分布直方图

于节点特征的解释和基于传播结构的解释，结合案例分析与全局统计分析发现：①从传播结构来看，一方面，非谣言的传播结构相对较为扁平，往往由信息源直接传播到网络中的绝大多数用户，而谣言传播纵深往往相对较长，经由较多次转发引发广泛传播与讨论；另一方面，非谣言数据传播核心单一，而谣言数据传播中且容易出现双中心甚至多中心的结构，引发观点的对峙与辩论。②从传播的用户及内容来看，谣言文本内容本身是进行谣言识别的根本依据，尤其是在谣言传播初期等传播网络较小的情形下。此外，谣言传播涉及的用户通常在社交媒体上影响力较低、活跃度较低。

5　总结与展望

针对当前谣言识别研究中传播网络结构信息应用不足与可解释性不足的问题，本文提出可解释图神经网络谣言识别框架。一方面，可通过残差图卷积网络谣言识别模型充分挖掘谣言传播网络中的结构信息，提高谣言识别准确率；另一方面，可通过图神经网络解释器的学习分别生成基于节点特征的解释和基于传播结构的解释，提升模型的可解释性与可信度。

根据研究发现，谣言在传播结构上具有路径长、中心多的特点，而涉及谣言传播的用户通常活跃度与影响力较低。从谣言治理角度来看，在谣言传播初期，需要更关注谣言文本，通过文本特征对谣言进行有效快速识别；在谣言传播后期，需要关注引发深度讨论与传播的事件，长而深的传播链条对谣言识别起着关键作用。对于社交媒体用户，在使用社交媒体时，也可以根据本文总结出的谣言传播的结构特点及内容特征增强对谣言的辨别能力。

本文具有一定的理论意义和实践价值。

理论意义上，在信息行为的视角下，本文将谣言识别任务结合图神经网络进行可解释分析，可以有效挖掘谣言传播行为发生的机理，从行为特征以及传播结构等方面提取模式和规律，从而更好地解释和预测人们在不同情境下的谣言传播行为，为验证现有理论提供实证结果。在信息价值的视角下，本文在社交媒体数据的基础上自动生成谣言识别的传播网络结构和传播节点属性的双重解释，是决策中信息质量和价值评估理论框架的一种新探索。

实践价值上，在谣言鉴别方面，图神经网络模型可以从海量社交媒体数据中筛选出潜在的谣言信息，并对其进行分析和评估。这有助于平台管理方了解谣言的传播路径、传播者及其背后的潜在动机，从而更好地洞察谣言事件的本质。在舆情治理方面，社交媒体上的谣言传播对社会舆情产生深远影响，管理部门可以利用图神经网络模型对谣言进行实时检测和可解释性分析。有助于政府部门及时了解社会舆情的变化趋势、民意的走向，并为政府决策提供依据。

未来可考虑传播网络的动态变化，进一步提升谣言识别模型的性能。信息的传播结构是一个随时间变化的过程，对此动态网络进行研究，有助于更深入地理解谣言传播机制。

参 考 文 献

[1] Bondielli A, Marcelloni F. A survey on fake news and rumour detection techniques[J]. Information Sciences, 2019, 497: 38-55.

[2] DiFonzo N, Bordia P. Rumor, gossip and urban legends[J]. Diogenes, 2007, 54(1): 19-35.

[3] Zubiaga A, Liakata M, Procter R, et al. Towards detecting rumours in social media[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 35-41.

[4] 贺刚, 吕学强, 李卓, 等. 微博谣言识别研究[J]. 图书情报工作, 2013, 57(23): 114-120.

[5] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter[C]// Proceedings of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 675-684.

[6] Kwon S, Cha M, Jung K, et al. Prominent features of rumor propagation in online social media[C]// Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Piscataway: IEEE, 2013: 1103-1108.

[7] 曾子明, 王婧. 基于LDA和随机森林的微博谣言识别研究——以2016年雾霾谣言为例[J]. 情报学报, 2019, 38(1): 89-96.

[8] Yu F, Liu Q A, Wu S, et al. A convolutional approach for misinformation identification[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 3901-3907.

[9] Wang W Y. "Liar, liar pants on fire": a new benchmark dataset for fake news detection[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 422-426.

[10] Afroz S, Brennan M, Greenstadt R. Detecting hoaxes, frauds, and deception in writing style online[C]// Proceedings of the 2012 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2012: 461-475.

[11] Liu X M, Nourbakhsh A, Li Q Z, et al. Real-time rumor debunking on Twitter[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1867-1870.

[12] Ma J, Gao W, Wei Z Y, et al. Detect rumors using time series of social context information on microblogging websites[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1751-1754.

[13] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[C]// Proceedings of the 2nd International Conference on Learning Representations. ICLR, 2014: 1-14.

[14] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]// Proceedings of the 5th International Conference on Learning Representations. ICLR, 2017: 1-14.

[15] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 1025-1035.

[16] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[C]// Proceedings of the 6th International Conference on Learning Representations. ICLR, 2018: 1-12.

[17] Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 7444-7452.

[18] Bian T A, Xiao X, Xu T Y, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 549-556.

[19] Bai N, Meng F R, Rui X B, et al. Rumour detection based on graph convolutional neural net[J]. IEEE Access, 2021, 9: 21686-21693.

[20] 王昕岩, 宋玉蓉, 宋波. 一种加权图卷积神经网络的新浪微博谣言检测方法[J]. 小型微型计算机系统, 2021, 42(8): 1780-1786.

[21] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1135-1144.

[22] Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2921-2929.

[23] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C]// Proceedings of the 16th IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 618-626.

[24] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks[C]// Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2018: 839-847.

[25] Zhang J M, Bargal S A, Lin Z, et al. Top-down neural attention by excitation backprop[J]. International Journal of Computer Vi-

sion, 2018, 126(10): 1084-1102.

[26] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS One, 2015, 10(7): e0130140.

[27] Galhotra S, Pradhan R, Salimi B. Explaining black-box algorithms using probabilistic contrastive counterfactuals[C]// Proceedings of the 2021 International Conference on Management of Data. New York: ACM Press, 2021: 577-590.

[28] Pope P E, Kolouri S, Rostami M, et al. Explainability methods for graph convolutional neural networks[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10764-10773.

[29] Huang Q, Yamada M, Tian Y, et al. GraphLIME: local interpretable model explanations for graph neural networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 6968-6972.

[30] Vu M N, Thai M T. PGM-Explainer: probabilistic graphical model explanations for graph neural networks[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2020: 12225-12235.

[31] Ying R, Bourgeois D, You J X, et al. GNNExplainer: generating explanations for graph neural networks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019: 9244-9255.

[32] Luo D S, Cheng W, Xu D K, et al. Parameterized explainer for graph neural network[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2020: 19620-19631.

[33] Li G H, Müller M, Thabet A, et al. DeepGCNs: can GCNs go as deep as CNNs?[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 9266-9275.

[34] Song C H, Tu C C, Yang C, et al. CED: credible early detection of social media rumors[OL]. (2018-11-10). https://arxiv.org/pdf/1811.04175.pdf.

[35] Kochkina E, Liakata M, Zubiaga A. PHEME dataset for rumour detection and veracity classification[DS/OL]. (2018-06-10). https://doi.org/10.6084/m9.figshare.6392078.v1.

[36] Wellman B. Culture of the Internet[M]. Marva: Lawrence Erlbaum Associates Publishers, 1997: 179-205.

[37] Ni J M, Abrego G H, Constant N, et al. Sentence-T5: scalable sentence encoders from pre-trained text-to-text models[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 1864-1874.

[38] Vapnik V N, Chervonenkis A. A note on one class of perceptrons [J]. Automation and Remote Control, 1964, 25(12): 821-837.

[39] Breiman L. Random forests[J]. Machine Language, 2001, 45(1): 5-32.

[40] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785-794.

[41] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification[C]// Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. Palo Alto: AAAI Press, 1998: 41-48.

[42] Shu K, Wang S H, Liu H. Understanding user profiles on social media for fake news detection[C]// Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval. Piscataway: IEEE, 2018: 430-435.

[43] Prasad J. The psychology of rumour: a study relating to the great Indian earthquake of 1934[J]. British Journal of Psychology General Section, 1935, 26(1): 1-15.

[44] 刘于思, 徐煜. 在线社会网络中的谣言与辟谣信息传播效果: 探讨网络结构因素与社会心理过程的影响[J]. 新闻与传播研究, 2016, 23(11): 51-69, 127.

[45] van der Maaten L, Hinton G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

[46] Kullback S, Leibler R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

（责任编辑　王克平）