计算机科学
*Computer Science*
ISSN 1002-137X,CN 50-1075/TP

# 《计算机科学》网络首发论文

题目：基于双向图注意力网络的潜在热点话题谣言检测
作者：李劭，蒋方婷，杨鑫岩，梁刚
网络首发日期：2024-05-29
引用格式：李劭，蒋方婷，杨鑫岩，梁刚．基于双向图注意力网络的潜在热点话题谣言检测[J/OL]．计算机科学．
https://link.cnki.net/urlid/50.1075.TP.20240528.0940.010

网络首发时间：2024-05-29 10:02:18
网络首发地址：https://link.cnki.net/urlid/50.1075.TP.20240528.0940.010



# 基于双向图注意力网络的潜在热点话题谣言检测

李劭[1] 蒋方婷[1] 杨鑫岩[1] 梁刚[1]

1 四川大学网络空间安全学院 四川成都 610211
（lishao@stu.scu.edu.cn）

**摘 要** 现有社交网络谣言检测方法大多将社交网络中的单个帖子视为检测目标，存在因数据量不足而导致的检测冷启动问题，影响检测性能。其次，现有方法没有对海量社交网络信息中与检测无关的信息进行过滤，导致检测时延较长，性能较差。另外在分析谣言的传播特征时，现有方法大多侧重于谣言传播过程中的静态特征，难以充分利用节点间的动态关系对复杂的传播过程进行表证，导致性能提升存在瓶颈。针对以上问题，本文提出了一种基于潜在热点话题和图注意力神经网络的谣言检测方法，该方法采用神经主题模型和潜在热点话题发现模型进行话题级别的谣言检测以克服冷启动问题，并设计了一个基于双向图注意力神经网络的检测模型 TPC-BiGAT，分析谣言话题传播过程中的动态特征以进行谣言真实性检测。在三个公开数据集上进行的多次实验证明，该方法在准确率上较现有方法取得了 3%-5%的显著提升，验证了本文方法的有效性。

**关键词：**谣言检测；社交网络；潜在热点话题；图神经网络；主题聚类

**中图法分类号** TP181

# Rumor Detection on Potential Hot Topics with Bi-Directional Graph Attention Network

LI Shao[1]， JIANG Fangting[1]， YANG Xinyan[1] and LIANG Gang[1]

1 School of Cyber Science and Engineering，Sichuan University， Chengdu，China

**Abstract** The existing methods for detecting rumors on social media networks typically focus on individual posts as the target of detection, which leads to a cold start problem due to insufficient data, adversely affecting the detection performance. Moreover, these methods do not filter out the vast amount of irrelevant information in social media networks, resulting in longer detection latency and poorer performance.
Additionally, current methods tend to emphasize static features during the spread of rumors when analyzing the characteristics of rumor propagation, making it difficult to fully leverage the dynamic relationships between nodes to model the complex propagation process. To address these issues, this paper proposes a rumor detection method based on potential hot topics and graph attention neural networks. The method employs a neural topic model and a potential hot topic discover model for topic-level rumor detection to overcome the cold start problem. Furthermore, a detection model named TPC-BiGAT is designed to analyze the dynamic features of rumor topic propagation for authenticity detection. Multiple experiments conducted on three public datasets demonstrate that this method achieves a significant improvement of 3%-5% in accuracy compared to existing methods, thereby validating the effectiveness of the proposed approach.

**Keywords** Rumor detection，Social network，Potential hot topic，Graph neural network，Topic cluster

## 1 简介

在线社交网络的增长改变了人们获取和传播信息的方式。We Are Social 和 Hootsuite 合作发布的 DIGITAL 2020[1]显示，超过 38 亿人使用社交媒体。社交网络用户可以通过推特、微博等各种社交网络以极低的成本来传播信息，从而导致了谣言的迅速传播。微博官方发布的辟谣报告[2]显示，2021 年微博官方共处理了 66251 条谣言。这些谣言会误导人们对事件的看法，严重的可能会引发恐慌并影响社会秩序[1]。因此，如何及时、精确的检测出社交网络中传播的谣言信息，控制其传播以降低其后续影响，成为了各大社交网络迫在眉睫的问题。

目前，社交网络谣言检测方法主要分为两大类：手动谣言检测方法和自动谣言检测方法[1]。大多数社交网络使用手动谣言检测方法[2]，将疑似谣言交给经验丰富的检测人员，依赖其知识和经验进行判断。尽管手动检测方法具有较高的准确性，但在检测效率和成本方面存在较大的局限性[1]。因此，自动谣言检测方法成为当前研究的主流方法。自动谣言检测方法根据模型结构的不同可分为机器学习方法[3-6]和深度学习方法[7-13]，通过训练检测模型来自动判断社交网络信息的真实性。

[1] https://datareportal.com/reports/digital-2020-global-digital-overview

2 https://weibo.com/1866405545/LcFuud7ml

基金项目：国际自然科学基金重大项目（No.62162057）
This work was supported by National Natural Science Foundation of China under Grant (No.62162057)
通信作者：梁刚（lianggang@scu.edu.cn）

现有自动谣言检测方法在一定程度上抑制了谣言的传播，但是其依然存在一些局限性。首先，大多数方法局限于检测每个单独帖子的真实性，而在突发事件的早期阶段，社交媒体用户发布的信息通常只包括原始帖子，缺乏相关评论和转发帖子。这导致了谣言检测的冷启动问题，影响谣言检测的性能[1]。一种有效的解决方法是将检测对象从单个帖子转变为包含多个帖子的话题，获取更多信息以进行话题级别的谣言检测。为此，一些研究采用手动方法或传统的聚类算法进行话题级别的谣言检测[14]，然而在处理来自社交网络的大量短文本时，此方法存在特征稀疏的问题。其次，现有方法大多没有对社交网络中的检测无关信息进行过滤。社交网络中包含大量与谣言检测无关的信息，例如个人日常信息和广告。将这些信息纳入谣言检测几乎没有价值并会增加检测时延，因此有必要在检测前过滤掉这些无关信息。此外，为了对谣言的传播过程进行分析，现有方法大多将传播结构建模为图结构并使用 GCN 等模型分析传播特征。然而，现有方法大多只能通过图的空间关系提取静态传播特征，无法根据节点间的关系动态学习传播特征，对谣言传播过程的表征性能较差，检测效果存在瓶颈。

为了解决现有方法的局限性，本文提出了一种基于潜在热点话题和图注意力神经网络的话题级别谣言检测方法，旨在降低冷启动问题影响并提升谣言检测效果。本文的主要贡献如下：

1）本文首次将神经主题模型引入谣言检测任务，将单独帖子聚类成话题来提升用于检测的数据量，并进行话题级别的谣言检测以降低冷启动问题对检测效果的影响；

2）为了过滤社交网络信息中的无关信息，降低其对检测效果的影响，本文提出了一个话题热度模型，利用话题热度表示谣言话题受关注的程度，通过计算话题热度的增长速度筛选大量谣言话题中潜在的热点话题，以其作为谣言检测的目标；

3）本文提出了一个基于图注意力神经网络的检测模型 TPC-BiGAT，将谣言传播表示为 TPC 图并以用户发布的文本信息及传播结构作为输入，进而利用双向图注意力网络动态的学习表征能力更强的谣言传播特征，以提升模型在区分谣言和非谣言信息时的性能。

4）在三个公开数据集上的大量实验证明，与现有的谣言检测基线方法相比，本文提出的方法取得了更好的检测性能，显示出本文方法的优越性。

本文的组织结构如下：第 2 节回顾了谣言检测的相关工作，第 3 节介绍了本文使用的定义和符号，第 4 节描述了本文提出的方法并对每个模块进行详细说明，第 5 节展示了该方法与其他基线方法的实验结果，第 6 节对本文进行总结并对未来研究方向进行讨论。

## 2 相关工作

随着社交网络的普及和发展，学术界对社交网络谣言检测领域的研究日益增多，研究者们主要探索了两大类谣言检测方法，即人工检测方法和自动检测方法。社交平台广泛使用人工谣言检测方法[2]，将怀疑是谣言的信息交给经验丰富的专业人士，由他们根据自身知识和经验判断其真实性。人工谣言检测方法易于使用，准确率也较高。然而，它们也面临一些挑战，如检测延迟显著、严重依赖检测人员的主观看法、无法处理社交网络中海量的信息。因此，当前谣言检测的研究主要集中于自动谣言检测方法。

早期的自动谣言检测研究主要依赖于特征工程提取人工构建的特征，采用传统机器学习算法学习这些特征以捕捉谣言和正常信息之间的区别，从而获得有效的检测模型。Castillo 等人[15]提取了谣言的文本特征，包括总词数和字符数、文本情感特征等。Yang 和 Kwon 等人[16,17]根据先前的研究提出了外部 URL 以及人称代词的数量等特征。Morris 和 Liu 等人[3–6]也设计了传播特征、用户特征和多媒体特征以增加用于谣言检测的特征数量。这些方法在初期研究中得到了广泛应用并取得了一些成功。然而，特征工程需要大量的时间和人力，且提取的特征难以在平台和领域间泛化，导致检测结果不稳定[1]。深度神经网络的出现为谣言检测引入了一种新的研究方向，其自动学习谣言表示的能力不仅不需要进行人工的特征工程，还能够提取更深层次、表征能力更强的特征[18]。因此，越来越多的研究者采用深度神经网络作为谣言检测的主要方法。

传统的深度学习的方法将谣言的文本信息建模为序列数据，通过循环神经网络学习其深层特征。Ma 等人[7]首次使用循环神经网络进行谣言检测，他们将原始谣言和评论信息构造成文本序列，利用循环神经网络从中学习谣言的文本表示，取得了比传统方法更好的结果。Xu 等人[19]将原始谣言信息分为三部分：原始信息、转发信息和用户特征，然后连接所有特征向量形成最终的输入向量作为分类器的输入。Guo 等人[20]的工作将谣言信息表示为层次结构，作为双向 LSTM 模型的输入，其输出则输入分类器以获得最终的检测结果。Ajao 等人[21]采用卷积神经网络处理文本信息并提取文本的深层表示，然后将这些深层表示转化为时间序列由循环神经网络进行分析。随着研究的不断发展，一些学者提出分析谣言信息传播过程并提取其深层特征可以为谣言检测提供关键信息[22]。Ma 等人[8]提出了一种基于核学习的谣言检测方法，该方法将谣言的传播建模为传播树，然后通过比较传播树的相似性来识别谣言信息。在此基础上，Ma 等人[22]又引入了递归神经网络(RvNN)来学习谣言的内容语义和传播过程，可以学习传播过程的深层特征，并在分

析文本的语义特征时捕捉节点之间的关系。这些研究大多将传播过程定义为传播树，结合文本信息和传播特征并使用循环神经网络进行分析。然而，这些方法难以处理复杂的谣言传播过程，对传播结构的分析能力较差。

图神经网络的提出为分析社交网络谣言传播过程带来了新的方法，越来越多的研究开始将谣言传播过程建模为图结构。图结构的节点表示谣言的原始帖子或评论转发帖子，图的边表示谣言传播的过程。Huang 等人[23]首次将图卷积网络应用于谣言检测。他们根据一组帖子的转发和回复关系构建了传播树，而将用户集合表示为一个关系图。采用了 Ma 等人提出的双向递归神经网络架构[22]处理文本信息，并使用图卷积神经网络对用户集合进行处理。最终将两个集合合并以进行谣言检测。Bian 等人[9]首次使用图卷积神经网络分析谣言传播过程，提出了一个 Bi-GCN 模型，从正向和反向两个方向提取传播特征。该模型包含两个独立的图卷积神经网络，学习不同方向上的传播过程并进行融合，获得最终的特征矩阵。

综上所述，现有谣言检测方法虽然在一定程度上抑制了社交网络谣言的泛滥，但是依然存在一些挑战。例如，现有方法主要依赖于经过较长时间传播的信息进行训练，存在检测冷启动问题，导致检测存在滞后性。现有方法大多对社交网络中的所有信息都进行真实性鉴别，没有过滤其中检测价值较低的无关信息，影响检测效率。

## 3 定义说明

为了更好地理解本文方法，以下是一些定义和符号的说明：

1. 原始帖子：指由用户发布的，不是对其他帖子进行回复或转发的帖子，本文使用 $p_i$ 表示集合中第 $i$ 个原始帖子。

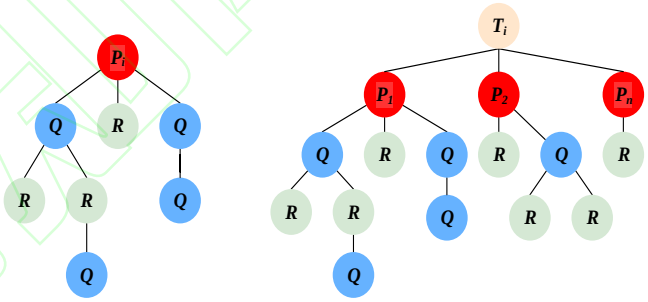
2. 评论帖子：指的是对原始帖子或其他评论帖子进行回复的帖子。在不同的社交网络平台上评论帖子的呈现方式会有差异，以推特为例，评论帖子可以分为两类：回复评论和转发评论。本文将所有类型的评论视为相同类型，使用 $c_t^i$ 表示与原始帖子 $p_i$ 相关的第 $t$ 个评论帖子。

3. 原帖-评论集合：指包含一个原始帖子及其所有评论帖子的集合，第 $i$ 个原帖-评论集合即表示为 $\delta_i = \{p_i, c_1^i, \dots, c_t^i, \dots, c_k^i\}$，其中 $k$ 是该原帖-评论集合中评论帖子的数量。每个原帖-评论集合不仅包含文本信息，还包括发布时间和属性信息，如评论数量和状态等。

4. 话题集合：指包含在同一话题中的所有原帖-评论集合组成的集合，第 $i$ 个话题集合即表示为 $E_i = \{\delta_1, \dots, \delta_m\}$，其中 $m$ 是该话题集合中原帖-评论集合的数量。每个话题集合对应一个标签 $y$，取值为非谣言或者谣言。

5. 原帖-评论图：指基于每个原帖-评论集合构建的无向图结构，其结构如图 1(a) 所示。原帖-评论图中有两种类型的节点：根节点表示的是原始帖子 $p_i$，叶子节点是原帖-评论集合 $\delta_i$ 中所有的评论帖子 $c_t^i$，节点的值是文本的词嵌入向量。图的边表示两个节点之间的关系。如果两个节点之间存在传播关系则二者间存在边，否则就不存在边。

6. 话题-原帖-评论图（TPC 图）：本文使用 TPC 图的形式来表示话题级别谣言的传播过程，该图是基于原贴-评论图构建的。通过将相同话题中的每个原帖-评论图中的每个原始帖子节点 $p_i$ 与其相应的话题节点连接后得到的无向图结构，如图 1(b) 所示。话题-原帖-评论图中的根节点代表话题节点，话题节点的值是该话题关键词的词嵌入向量。话题节点与每个原始帖子节点之间的边表示该原始帖子属于该话题，即该原帖-评论集合属于当前话题集合。其他节点和边的含义与原帖-评论图中的相同。




**(a) 原帖传播图** **(b) 话题传播图**

图 1 原帖-评论图与话题-原帖-评论图

Fig.1 Post-Comment and Topic-Post-Comment graph

7. 话题热度：在潜在热点话题发现模块中，本文引入了话题热度来衡量话题受关注程度的指标，作为选择潜在热点话题和过滤无关信息的标准。对于话题集合 $E_i$，话题热度 $H_{E_i}$ 可以根据其相关属性进行计算，计算过程在 4.2 节中进行解释。

## 4 方法描述

本文提出的方法使用 BERTopic 神经主题模型进行主题聚类，并通过潜在热点话题发现模块发现潜在的热点话题，然后将话题建模为TPC图结构，最终将其输入谣言真实性检测模型进行谣言检测。该方法流程如图 2 所示，主要分为四个部分，分别为话题聚类模块、潜在话题发现模块、TPC 图构建以及谣言真实性检测模块，每个部分的细节将在以下小节中进行说明。

### 4.1 主题聚类模块

在谣言传播的早期阶段，基于单个帖子的谣言检测可能受到由于数据不足而导致的冷启动问题的影响。然而，社交网络用户通常会围绕特定话题发布相关信息。在社交网络中引发广

泛讨论的热点话题通常涉及庞大的用户群体和大量相互交互的信息。因此，由于一个话题比单个帖子包含更多的信息，因此可以进行话题级别的谣言检测，以减轻冷启动问题对检测的影响，然而，社交网络中的帖子并不直接说明它们所讨论的话题，因此需要一种方法来对各话题下的帖子进行分类。常见的方法是通过人工分析用户发布的原始帖子信息，并将属于同一话题的帖子归为一类。但是人工方法存在较大的延迟，无法处理庞大的数据量，无法满足谣言检测的要求。

鉴于这些问题，本文采用了主题模型对文本内容进行聚类以克服手动方法的局限性。考虑到社交网络中的文本信息主要由短文本组成，本文采用了一种名为 BERTopic 的神经主题模型[24]进行主题聚类。BERTopic 模型结合了基于 Transformer 的预训练模型与 HDBSCAN 聚类算法来进行主题聚类。与传统聚类算法和主题模型相比，BERTopic 神经主题模型对文本进行了词嵌入处理，解决了在处理短文本信息时出现的稀疏性问题，更适用于多数为短文本信息的社交网络信息。

BERTopic 模型包括三个主要过程：文本嵌入、主题聚类和主题表示。在本文中，对于多个原帖-评论集合 $\delta_i$，主题聚类模块的目标是根据每个原帖-评论集合 $\delta_i$ 中的原始帖子 $p_i$ 的文本进行聚类，将属于相同话题的原帖-评论集合聚类到一个话题中得到多个话题集合$E_i$。首先，该模块会将原始数据中所有的帖子进行分类，选择其中所有的原始帖子输入 BERTopic 模型中。然后，BERTopic 模型会使用 BERT 等预训练模型对帖子的文本内容进行词嵌入，将文本内容通过向量进行表示。进而，BERTopic 会通过 HDBSCAN 聚类算法对文本向量进行聚类，得到多个话题与各话题的表示。最终，该模块将原始帖子与其相关的评论帖子合并后得到的原帖-评论集合 $\delta_i$按照聚类结果进行合并，最终得到多个话题集合 $E_i$ 作为该模块的输出。

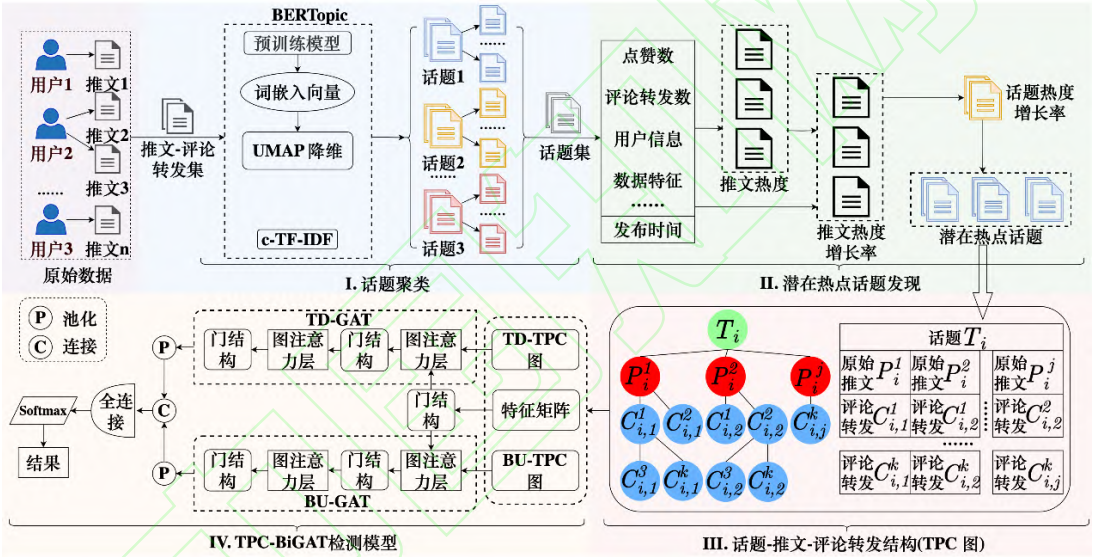



图 2 本文模型框架图

Fig.2 The framework of the proposed model

## 4.2 潜在热点话题发现模块

社交网络中存在着海量的信息，如果对所有信息都进行谣言检测会严重影响检测效率，因此需要对其中检测价值较低的信息进行过滤。因此，本文设计了潜在热点话题发现模块，关注于从所有话题中筛选出在经过一段时间的传播后更有可能发展成为引起广泛关注和讨论的潜在热点话题，并将这些潜在热点话题作为谣言检测的目标，达到过滤无关信息的目的。

潜在热点话题发现模块基于话题热度实现，其结构如图 3 所示。对于话题聚类后的某个话题，该模块会根据其中每个帖子的相关信息分别计算其热度基值和权值，进而得到每个帖子的热度，通过聚合该话题内所有帖子的热度来计算该话题的热度。然后，选择一定时间段并计算该话题热度在该时间段的增长速率。最后，通过比较所有话题的热度增长速率来发现潜在的热点话题。其工作流程如下：

首先，潜在热点话题发现模块会遍历所有话题集合，对于其中的某一个话题集合 $E_i$，其中会存在多个原帖-评论集合 $\delta_i$，其中包括原始帖子和评论帖子。帖子热度的计算与多种因素相关，一般来说，帖子的评论和点赞数可以直观的显示出当前帖子受到的关注程度。除此之外，帖子的自身特征也会影响其传播速度和受关注程度。因此，潜在热点话题发现模块在计算帖子热度的时候不仅考虑了由评论和点赞数计算得到的热度基值，还考虑到根据帖子自身特征得到的热度权值。本文中使用多媒体信息、外部链接、用户认证信息以及用户粉丝数量等特征来计算帖子的热度权值，具体如下：

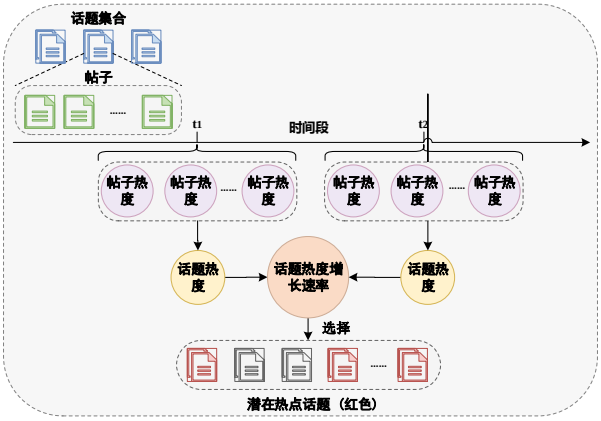



图 3 潜在热点话题发现模块结构图

Fig.3 The structure of Potential Hot Topic Discover module

1. 多媒体信息。帖子中的多媒体信息会吸引更多用户的关注，可以作为计算热度权值的指标。多媒体信息特征 $S_{media}$ 等于帖子中多媒体信息的数量，若不存在则 $S_{media}$ 为 0；

2. 外部链接。帖子中的外部链接会作为帖子的来源或补充信息，外部链接更多的帖子往往会被更多的用户传播。外部链接特征 $S_{url}$ 等于帖子中外部链接的数量，若不存在则 $S_{url}$ 为 0；

3. 用户认证信息。已认证用户发布的信息往往会受到更多用户的关注，并且较未认证用户发布的信息更容易被其他用户信服和传播。如果信息的发布用户是已认证用户则用户认证特征 $S_{au}$ 为 1，反之则 $S_{au}$ 为 0；

4. 用户粉丝数量。用户粉丝数量代表能够直接收到该用户发布信息的其他用户数量，发布者粉丝数量越多，信息就越可能被更多的人传播。用户粉丝数量特征 $S_{fo}$ 的值等于帖子发布者的粉丝数量。

本文选取所有特征的几何平均数来计算热度权值以减少极端值造成的影响，计算过程如公式(1)所示。

$$H_t = \sqrt[4]{S_{media} * S_{url} * S_{au} * S_{fo}} \tag{1}$$

热度基值 $H_b$ 通过计算帖子的评论数 $N_{co}$ 和点赞数 $N_{st}$ 的几何平均数得到，如公式(2)所示。

$$H_b = \sqrt{N_{co} * N_{st}} \tag{2}$$

将公式(2)和(3)得到的帖子热度基值与权值相乘即可得到该帖子的热度值。重复此过程可以得到原帖-评论集合 $\delta_i$ 中所有原始帖子和评论帖子的热度，进而根据公式(3)得到原帖-评论集合 $\delta_i$ 的热度 $H_{\delta_i}$，最终通过公式(4)计算得到话题集合 $E_i$ 的热度 $H_{E_i}$。

$$H_{\delta_i} = H_{p_i} + \sum_1^k c_i^t \tag{3}$$

$$H_{E_i} = \sum_1^m H_{\delta_i} \tag{4}$$

一般来说，话题热度的增长速度与其是否会成为热点话题有着较大的关系，一个话题如果显示出热度快速增长的趋势，那么它在传播过程中有可能吸引更多的关注，从而其成为热点话题的概率就更高。因此，话题热度的增长速率可以作为选择潜在热点话题的标准。话题热度增长速率的计算首先需要确定一个时间段的开始时间 $t_1$ 以及结束时间 $t_2$，分别计算当前话题集合在两个时刻的话题热度 $H_{t_1}$ 和 $H_{t_2}$，最终根据公式(5)得到当前话题集合在 $t_1$ 到 $t_2$ 时间段内的热度增长速率 $R$。

$$R = \left(H_{t_2} - H_{t_1}\right) \div (t_2 - t_1) \tag{5}$$

最后，对于所有话题集合 $E_i$，在相同的时间段内计算其热度增长速度 $R_i$，再将 $R_i$ 与预先设置的阈值进行比较。若 $R_i$ 大于阈值，则表示话题 $E_i$ 在一段时间的发展后成为热点话题的可能性较其他话题更高，即将 $E_i$ 视为一个潜在热点话题，反之则将 $E_i$ 视为无关信息进行过滤。最终，该模块将所有被视为潜在热点话题的话题集合进行汇总，作为潜在热点话题发现模块的输出，这些话题即为话题级别谣言检测的目标，作为谣言真实性检测模块的输入。

### 4.3 谣言真实性检测模块

本文设计了一个基于传播结构的谣言真实性检测模块，其核心思想是使用图注意力神经网络模型挖掘谣言传播过程中的深层特征以提升区分谣言和非谣言信息的效果。该模块首先将谣言的传播过程建模为话题-原帖-评论图（TPC 图）的结构，进而根据该结构中的无向图构造两个方向相反的有向图，将两个有向图作为双向图注意力神经网络模型的输入，获取得到不同方向上的谣言传播特征，将两个特征通过连接的方式进行融合，最终将融合特征输入分类器对谣言真实性进行检测。

#### 4.3.1 TPC 图构建

当前基于传播结构的谣言检测工作一般都将谣言的传播过程建模为图结构，但是其大多都针对于某个单独帖子的传播过程。而本方法对原始数据进行了主题聚类以进行话题级别的谣言检测，需要将同一话题下的谣言信息的传播结构进行聚合，因此，本文将数据集中的每个话题集合建模为 TPC 图的形式。TPC 图可以从节点和边两方面进行描述，其包含三类节点：

1. 话题节点：每一个 TPC 图中包含且只包含一个话题节点，其作为图的根节点表示其他节点属于该主题，节点特征使用主题聚类模型中得到的主题关键词嵌入向量进行表示；

2. 原帖节点：表示所有属于当前主题的原始帖子，每个原帖节点都是话题节点的子节点，且其互相之间不存在边，节点特征使用原始帖子的文本词向量进行表示；

3. 评论节点：表示从属于某个原帖-评论集合，包含于某个原始帖子传播过程中的评论信息。节点特征与原帖节点相同，使用评论帖子的文本词向量进行表示。

TPC 图为无向图结构，不同类型节点之间的边表示不同的

关系，可以分为两种：话题节点与原帖节点之间的边表示从属关系，代表该原始帖子属于当前话题；而原帖节点与评论节点、评论节点和评论节点之间的边表示谣言的传播过程，节点间的边代表两个节点在传播过程中存在关联。

该过程使用邻接矩阵和特征矩阵来将某个特定的话题集合构建为一个 TPC 图。根据输入的帖子总数为 $n$ 的话题集合 $E_i$，基于其中的从属关系和谣言传播过程，可以得到其对应的邻接矩阵 $A \in R^{n\times n}$ 与包含所有节点特征的特征矩阵 $X$，通过 $A$ 和 $X$ 即可得到神经网络的输入。

4.3.2 生成输入图

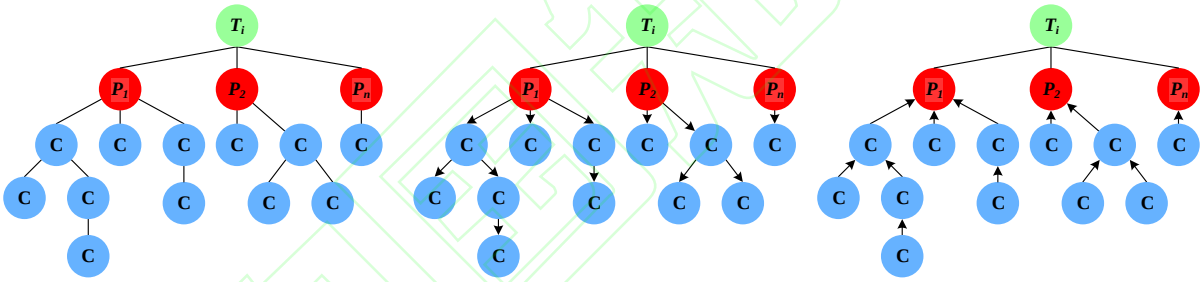
为了从多个方向提取谣言传播过程中的深层特征，本文参考了 Bian 等人[9]提出的双层图神经网络结构，将 4.3.1 节中得到的 TPC 无向图转化为两个方向相反的有向图 TD-TPC 图和 BU-TPC 图。图 4(a) 为初始的无向 TPC 图，图 4(b)为表示传播结构中深度传播过程的自顶向下的有向图，称为 TD-TPC 图，图 4(c) 为表示传播结构中广度传播过程的自底向上的有向图，称为 BU-TPC 图。与 Bian 等人不同的是，TPC 图中包含话题节点，其与原帖节点之间的关系不属于谣言传播过程，而是表示原始帖子与话题之间的从属关系，因此话题节点与原帖节点间的边是双向的，从而将从属关系与传播结构进行区分。

根据 TPC 图的邻接矩阵 $A \in R^{n\times n}$，可以得到只包含单方向的 TPC 图的邻接矩阵 $A' \in R^{n\times n}$（假设其只包含从上节点到下节点的边）。除此之外，该过程使用了矩阵 $P \in R^{n\times n}$，其为一个第一行与第一列的值都为 1 的矩阵，可以表示话题节点与原帖节点之间的双向关系。根据公式(6)(7)可以计算 TD-TPC 图与 BU-TPC 图邻接矩阵 $A^{TD}$ 与 $A^{BU}$。两个有向图都使用与 TPC 图相同的特征矩阵 $X$ 作为各自的特征矩阵。

$$A^{TD} = A' \cup P \tag{6}$$

$$A^{BU} = A'^{T} \cup P \tag{7}$$

另外，本文在图的生成过程中使用了 DropEdge 方法[25]，其为一种增加输入数据随机性，减少图神经网络中存在的过拟合问题的方法，其关键是删除图中一定比例的边。具体来说，在谣言真实性检测模型的每一次训练过程中，DropEdge 会先通过修改邻接矩阵的方式随机的从 $A'$ 中删除一定比例的边，再通过删除边后的 $A'$ 计算新的 $A^{TD}$ 与 $A^{BU}$，最终经过删边处理后的 $A^{TD}$ 与 $A^{BU}$ 将作为此次训练过程的输入。




(a) UD-TPC 图　(b) TD-TPC 图　(c) BU-TPC 图

图 4 三种不同的图结构

Fig.4 Three different graph structures

4.3.3 谣言真实性检测

图注意力神经网络提出的目的是增强图中其他节点对当前节点的表征效果。在图卷积神经网络中，每个节点的邻居节点的地位都相同，无法根据邻居节点重要程度调整其对当前节点的影响，限制了其挖掘空间特征的能力。图注意力神经网络将注意力机制引入图神经网络中，使用多头注意力权重的邻居节点特征聚合函数替代图卷积神经网络模型中的特征标准化函数。模型可以根据邻居节点的特征计算其注意力系数，从而为每个邻居节点分配不同的权重，根据其权重大小不同程度的影响当前节点的表征，从有效的提升了模型效果。

本文中设计了一个基于双向图注意力神经网络的谣言真实性检测模型，该模型以 4.3.2 节中生成的邻接矩阵 $A^{TD}$，$A^{BU}$ 和特征矩阵 $X$ 作为输入，由两个不同方向的图注意力神经网络组成。两个邻接矩阵分别输入到不同方向的图注意力神经网络中。其中一个图注意力神经网络称为 TD-GAT，其以 $A^{TD}$ 为输入；另一个称为 BU-GAT，其以 $A^{BU}$ 为输入。将两个 GAT 的输出结果融合进行最终分类。另外，为了学习到更准确的节点表示，本文借鉴了 Lin 等人[26]提出的方法，设计了一个门模块使用根特征来增强其他特征的表示以提高谣言检测的性能。模型具体结构如图 5 所示。

图注意力网络是通过堆叠图注意力层来实现的，图注意力网络的每一层输入是一组向量，第 $l$ 层的输入表示为 $H^l = \left(h^l_{x_0}, h^l_{x_1}, \ldots, h^l_{x_n}\right)^T$，其中 $n$ 是节点的数量。一般来说，上层图注意力层的输出是下一层图注意力层的输入。然而，门模块改变了这个过程。具体过程是首先计算门向量 $g^l_{x_0\to x_i}$，它可以表示在第 $l$ 层图注意力层中 $x_i$ 受到根特征 $x_0$ 增强的程度，如公式(8)所示。

$$g^l_{x_0\to x_i} = sigmoid\left(W^l_g h^l_{x_i} + U^l_g h^l_{x_0}\right) \tag{8}$$

其中，$W_g^l$ 与 $U_g^l$ 为门模块中可训练的权重参数。根据门向量可以计算出原始特征在经过门模块的根特征增强后的输出，如公式(9)(10)所示，其中 $\odot$ 表示矩阵的哈密顿乘积。由门模块增强的输出作为下一层图注意力层的输入。

$$\widetilde{h_{x_0}^l} = h_{x_0}^l \tag{9}$$

$$\widetilde{h_{x_i}^l} = g_{x_0 \to x_i}^l \odot h_{x_i}^l + \left(1 - g_{x_0 \to x_i}^l\right) \odot h_{x_0}^l \tag{10}$$

图注意力层中的每个节点间共享自注意力机制，其中对于第 $l$ 层中的一个节点对 $(x_i, x_j)$ 可以通过公式(11)计算两个节点特征向量之间的相关度 $e_{ij}^l$。

$$e_{ij}^l = a\left(W^l[\widetilde{h_{x_i}^l} \| h_{x_i}^l], W^l[\widetilde{h_{x_j}^l} \| h_{x_j}^l]\right) \tag{11}$$

其中，$W^l$ 是一个可训练的线性变换权重矩阵，$h$ 是节点的特征，$a(\cdot)$是计算两个特征向量相关度的函数，一般使用向量内积。通过节点间的相关度即可计算 $(x_i, x_j)$ 之间的注意力系数$a_{ij}^l$，如公式(12)所示。

$$a_{ij}^l = \frac{\exp(e_{ij}^l)}{\sum_{k \in N_i} \exp(e_{ij}^l)} \tag{12}$$

其中$N_i$ 包含 $x_i$ 的一阶邻接节点和 $x_i$。得到归一化的注意力系数后可以对邻接节点特征进行线性组合，经过激活函数来更新节点特征作为第 $l$ 层图注意力层的输出，如公式(13)所示。

$$h_{x_i}^{l+1} = \sigma\left(\sum_{j \in N_i} a_{ij}^l W^l \widetilde{h_{x_j}^l}\right) \tag{13}$$

其中 $\sigma$ 为激活函数。在公式(13)的基础上，图注意力神经网络引入了多头注意力机制，通过多个注意力机制计算不同的注意力系数以稳定自注意力的学习过程，将得到的 $K$ 个特征进行连接得到输出特征，如公式(14)所示。

$$h_{x_i}^{l+1} = \|_{k=1}^{K} \sigma\left(\sum_{j \in N_i} a_{ij}^{l,k} W_k^l \widetilde{h_{x_j}^l}\right) \tag{14}$$

根据图注意力层的层数，通过多次迭代公式(14)来计算输出特征。在图注意网络的最后一层，多头注意机制的计算结果不再通过连接计算，而是以平均的形式计算。最后对输出特征 $H^L$进行平均池化，得到最终的输出 $S$，如公式(15)所示。

$$S = MeanPooling(H^L) \tag{15}$$

通过以上计算过程可以分别得到 TD-GAT 与 BU-GAT 的输出结果 $S^{TD}$ 与 $S^{BU}$。然后将两个输出进行连接，将连接后的特征输入多个全连接层，最终输入 softmax 层进行分类，得到谣言真实性的检测结果 $\hat{y}$，如公式(16)所示。

$$\hat{y} = softmax\left(FC(S^{TD} \| S^{BU})\right) \tag{16}$$

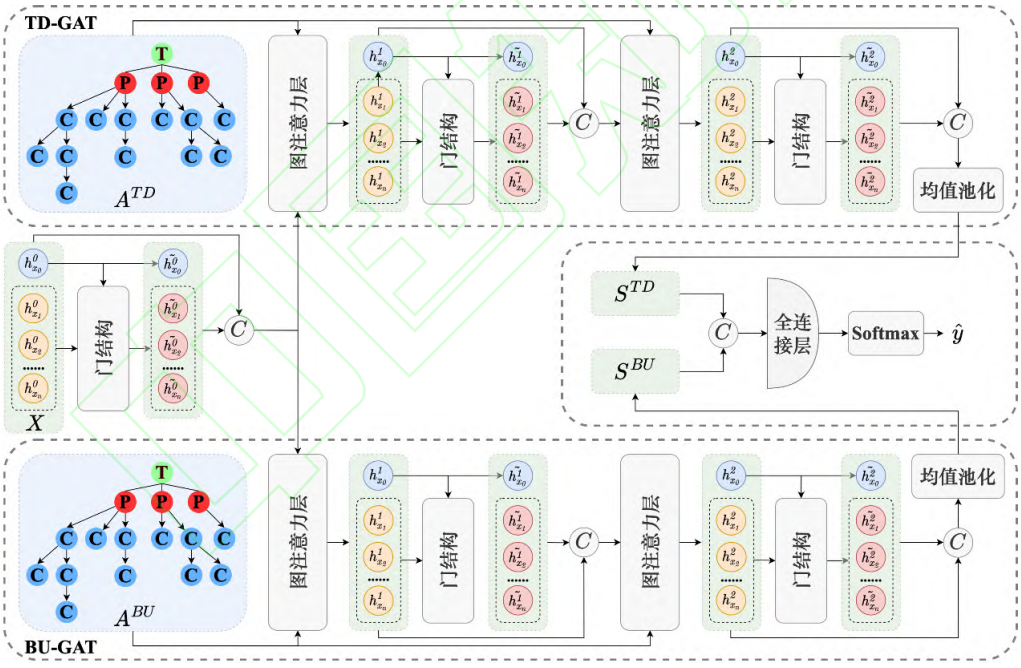



图 5 双向图注意力神经网络模型结构图

Fig.5 Bi-GAT neural network model structure diagram

## 5 实验结果与分析

为了验证本文提出方法的性能，我们在多个公开的谣言检测数据集上进行了实验，将本文模型与其他基线模型的效果进行了比较。然后，我们对模型中的每个模块进行了消融实验，分析了每个模块对整体模型效果的影响。

### 5.1 数据集

现有的谣言检测数据集根据数据的不同组织结构可以分为帖子级别和话题级别两大类。为了从多个方面进行评估，我们分别选择了两类数据集进行实验，详细情况如表 1 所示。

帖子级别：本文使用了 Twitter15[8], Twitter16[8]与 Weibo[7] 三个数据集。Twitter15 和 Twitter16 中每个事件标签根据 Snopes 等事实核查网站中发布文章的真实性标签进行标注，其标注结果分为四类：非谣言、虚假谣言、真实谣言和未验证的谣言。

话题级别：本文使用了 BEARD[27]，一个较新的话题级别谣言数据集。该数据集包含 Snopes 和 Twitter 中约 330 万条相关帖子，其标注结果分为虚假谣言和真实谣言两类。

表 1 数据集统计情况

Table 1 Statistics of the data

| 统计数量 \ 数据集 | 帖子级别 | | 话题级别 |
|---|---|---|---|
| | Twitter15 | Twitter16 | BEARD |
| 话题集合数量 | - | - | 1198 |
| 原帖-评论集合数量 | 1490 | 818 | 17299 |
| 平均帖子数/主题 | - | - | 14 |
| 帖子数量 | 331612 | 204820 | 3302732 |
| 真实谣言数量 | 374 | 203 | 531 |
| 虚假谣言数量 | 370 | 205 | 667 |
| 未证实谣言数量 | 374 | 203 | - |
| 非谣言数量 | 372 | 205 | - |

### 5.2 基线方法

为了展示本文提出方法的优越性，我们将其与其他基线方法进行了比较，这些基线方法可以分为三组：机器学习方法、传统深度学习方法和图神经网络方法。

机器学习方法包括：1. DTC[15]：使用决策树分类器通过给定推文集合中提取的各种手工特征来评估其可信度。2. SVM-TS[28]：将时间序列划分为固定间隔以捕捉时间特征并使用 SVM 分类器检测谣言。

传统深度学习方法包括：3. RNN+CNN[21]：一种使用卷积神经网络和长短时记忆神经网络进行谣言检测的混合模型。4. SVM-TK[8]：通过 SVM 分类器核学习评估谣言传播树结构之间的相似性来捕捉谣言传播结构的差异。5. RvNN[22]：一种带有 GRU 单元的用于学习谣言表示的树状递归神经网络，根据不同方向的传播结构构建了两个传播树。6. RL-ERT[29]：一种基于强化学习的谣言跟踪模型，通过权重调整策略聚合多个组件，并利用特定的社交特征提高谣言检测的性能。

图神经网络方法包括：7. PGNN[30]：一种基于门控图神经网络的传播图神经网络，它可以通过在相邻节点之间传播信息将文本和结构特征嵌入到高层次表示中。8. Bi-GCN[9]：使用两个不同方向的图卷积神经网络从谣言的自顶向下和自底向上传播方向提取特征，并融合两个方向的特征进行谣言检测。9. ClaHi-GAT[26]：将帖子内容和图信息表示为潜在空间，通过在帖子级别上的声明感知注意力和在事件级别上的基于推理的注意力来捕捉谣言特征。10. Dynamic-GCN[31]：一个用于捕捉谣言的影响特征，并基于模式匹配算法学习其传播上的动态结构以实现早期谣言检测的模型。

### 5.3 参数设置

对于 DTC 和基于 SVM 的方法，我们使用 scikit-learn 库进行实现，并将数据集按照 9:1 的比例划分为训练集和测试集。对于其他基于神经网络的方法，我们使用 PyTorch 进行实现，并将数据集按照 8:1:1 的比例划分为训练集、测试集和验证集。所有方法的学习率初始化为 5e-3，并使用 Adam 优化器自动调整学习率。为了缓解梯度爆炸和梯度消失的问题，模型的梯度衰减设置为 1e-3。为了防止过拟合，所有方法都采用了 EarlyStopping[32]。对于基于图神经网络的基线方法，我们使用了预训练的 RoBERTa 模型来得到图中的节点特征。图神经网络结构由两层图神经网络层组成，多头注意力机制头部数量设置为 4。DropEdge 值设置为 0.2，隐藏层大小设置为 128。一些基线方法的特殊参数设置与其各自的论文保持一致。关于结果评估，我们选择了特定类别的准确率和 F1 分数作为评估指标。所有比较的方法都使用了各自论文中报告的默认优化设置。为了确保实验时对所有基线方法进行公平比较，所有基线方法在所有数据集上都进行了五折交叉验证以获得稳健的结果。

### 5.4 谣言检测性能

表 2 展示了本文模型在在 Twitter 数据集与其他基线方法的性能比较，每一列的最佳结果用粗体表示。由于 Twitter 数据集中的帖子数量较少，难以进行话题级别谣言检测。为了更好的展示本文模型的优越性，我们在 BEARD 数据集上分别进行了帖子级别和话题级别的谣言检测，其结果如表 3 所示。

表 3 BEARD 数据集上的谣言检测结果

Table 3 Rumor detection results on BEARD dataset

| 基线方法 \ 数据集 指标 | BEARD (帖子级别) | | | BEARD (话题级别) | | |
|---|---|---|---|---|---|---|
| | Acc. | TR-F1 | FR-F1 | Acc. | TR-F1 | FR-F1 |
| DTC | 0.429 | 0.532 | 0.374 | | —— | |
| SVM-TS | 0.474 | 0.436 | 0.510 | | —— | |
| RNN+CNN | 0.501 | 0.463 | 0.523 | | —— | |
| RL-ERT | 0.544 | 0.481 | 0.602 | | —— | |
| SVM-TK | 0.667 | 0.727 | 0.571 | | —— | |
| RvNN | 0.726 | 0.710 | 0.723 | 0.562 | 0.550 | 0.574 |
| PGNN | 0.764 | 0.757 | 0.784 | 0.791 | 0.773 | 0.801 |
| Bi-GCN | 0.817 | 0.791 | 0.825 | 0.859 | 0.832 | 0.870 |
| Dynamic-GCN | 0.829 | 0.814 | 0.835 | 0.838 | 0.807 | 0.856 |
| ClaHi-GAT | 0.853 | 0.842 | 0.861 | 0.882 | 0.877 | 0.886 |
| TPC-BiGAT | **0.877** | **0.872** | **0.883** | **0.915** | **0.897** | **0.928** |

实验结果显示，基于传统特征工程的检测方法在所有数据集上的性能表现都较差，这是因为手工提取的特征很难体现出谣言信息与其他信息的区别。其中，引入传播结构特征的 SVM-TS 方法的效果要优于其他基于传统特征的检测方法。

第二组基于传统神经网络的基线方法摒弃了繁杂的特征工程，使用神经网络来挖掘深层特征。其中，RNN+CNN 和 RL-ERT 方法仅关注谣言文本层次上的特征，未考虑到谣言信息的传播特征，并未取得较好的效果。而 SVM-TK 与 RvNN 方法关注谣言的传播过程，使用了不同方法来挖掘谣言在传播过程中的深层特征，其结果较其他方法有了较大的提升。以上实验效

果的对比显示出了传播结构特征在谣言检测中的巨大作用。

第三组基于图神经网络的基线方法较第二组方法而言提升了对谣言传播结构的表示和提取深层传播特征的能力，其效果优于第二组方法。其中，Bi-GCN 与 ClaHi-GAT 方法的效果在总体上优于 PGNN 方法。然而，本文提出的 TPC-BiGAT 模型对检测网络结构进行了改进，不仅使用了双向图注意力神经网络分析不同方向上的传播特征，还将门结构加入网络之中以提升原始信息对谣言表示的影响。实验结果显示，该方法在各个数据集上都取得了优越的性能，较其他基线方法有了 3%-5% 的提升，显示出 TPC-BiGAT 模型的优越性。另外，通过对比不同模型进行帖子级别与话题级别两种不同层次的谣言检测效果可以发现，除了方法本身不允许进行话题级别谣言检测的方法以外，只有 RvNN 方法进行话题级别谣言检测时的效果低于进行帖子谣言检测时的效果。这是由于其自身的模型结构较为简单，难以提取话题中大量信息的深层特征所致。除此之外，其他基线方法在进行话题级别谣言检测时的效果都要优于进行帖子级别谣言检测时的效果，证明了本文设计的话题级别谣言方法的优良性能。并且，本文提出的 TPC-BiGAT 模型不仅较其他基线方法取得了更好的话题级别检测效果，也优于其进行帖子级别谣言检测时的效果。

以上结果都表明，本文提出的 TPC-BiGAT 方法通过构造 TPC 图进行话题级别的谣言检测，并使用处理传播结构效果更好的双向图注意力神经网络，在公开谣言检测数据集上取得了较其他基线检测算法更为优秀的谣言检测性能。

表 2 推特数据集上的谣言检测结果

Table 2 Rumor detection results on Twitter datasets

| 数据集 / 指标 / 基线方法 | Twitter15 | | | | | Twitter16 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | TR-F1 | FR-F1 | UR-F1 | NR-F1 | Acc. | TR-F1 | FR-F1 | UR-F1 | NR-F1 |
| DTC | 0.454 | 0.733 | 0.355 | 0.317 | 0.415 | 0.473 | 0.190 | 0.080 | 0.482 | 0.254 |
| SVM-TS | 0.544 | 0.404 | 0.472 | 0.483 | 0.796 | 0.574 | 0.571 | 0.420 | 0.526 | 0.755 |
| RNN+CNN | 0.477 | 0.300 | 0.507 | 0.640 | 0.359 | 0.564 | 0.394 | 0.543 | 0.674 | 0.591 |
| RL-ERT | 0.624 | 0.682 | 0.510 | 0.644 | 0.672 | 0.647 | 0.722 | 0656 | 0.571 | 0.603 |
| SVM-TK | 0.750 | 0.765 | 0.698 | 0.733 | 0.804 | 0.732 | 0.836 | 0.709 | 0.686 | 0.740 |
| RvNN | 0.723 | 0.821 | 0.758 | 0.654 | 0.682 | 0.737 | 0.835 | 0.743 | 0.708 | 0.662 |
| PGNN | 0.765 | 0.792 | 0.743 | 0.819 | 0.711 | 0.776 | 0.806 | 0.721 | 0.785 | 0.751 |
| Bi-GCN | 0.830 | 0.877 | 0.776 | 0.825 | 0.790 | 0.838 | 0.823 | 0.822 | 0.864 | 0.786 |
| Dynamic-GCN | 0.836 | 0.897 | 0.812 | 0.793 | 0.769 | 0.820 | 0.853 | 0.773 | 0.797 | 0.834 |
| ClaHi-GAT | 0.857 | 0.870 | 0.831 | 0.837 | 0.849 | 0.854 | 0.865 | 0.902 | 0.818 | 0.824 |
| TPC-BiGAT | **0885** | **0.906** | **0.842** | **0.868** | **0.898** | **0.901** | **0.886** | **0.914** | **0.869** | **0.913** |

表 4 TPC-BiGAT 模型的消融实验结果

Table 4 Ablation study results on TPC-BiGAT model

| 数据集 / 指标 / 消融模型 | Twitter15 Acc. | Twitter16 Acc. | BEARD Acc. |
|---|---|---|---|
| TPC-BiGAT | 0.885 | 0.901 | 0.872 |
| BiGAT w/o gating | 0.859 | 0.868 | 0.855 |
| BiGAT w/o root | 0.822 | 0.849 | 0.842 |
| GAT/UD | 0.777 | 0.819 | 0.789 |
| GAT/BU | 0.762 | 0.826 | 0.764 |
| GAT/TD | 0.747 | 0.806 | 0.773 |
| GCN | 0.739 | 0.778 | 0.751 |

### 5.5 消融实验

我们通过舍弃一些 TPC-BiGAT 模型的关键组件来对这些组件在模型中起到的作用进行消融实验。首先，为了研究神经网络结构对检测效果的影响，我们将神经网络模型替换为其他结构来进行效果比较。这些结构包括: 1. TPC-BiGAT w/o gating: 模型结构与 TPC-BiGAT 相同，但是舍弃了门结构而是使用 Bian 等人[9]提出的根源节点增强方式进行处理；2. TPC-BiGAT w/o root: 舍弃了所有根源节点增强模块；3. GAT/UD: TPC-BiGAT 使用双向图作为模型输入，该结构使用无向图作为模型输入；4. GAT/BU: 与 3 相同，使用自底向上方向的图作为模型输入；5. GAT/TD: 与 3 相同，使用自顶向下方向的图作为 GAT 网络的输入；6. GCN: 普通图卷积网络。

根据表 4 可以得出结论。首先，不论输入的是有向图还是无向图，图注意力神经网络的效果总要优于图卷积神经网络的效果，这显示出图注意力神经网络所使用的注意力机制在分析图结构时的优良效果。其次，除了 Twitter16 数据集，使用无向图作为输入的图注意力神经网络的效果要优于使用自顶向下和自底向上两个方向上的有向图作为输入的效果。而使用双向图注意力神经网络的 TPC-BiGAT 模型可以将无向图分解为两个有向图，通过不同的网络结构进行分析，其效果较使用无向图作为输入的图注意力神经网络的效果有了较为明显的提升，显示出双向图神经网络在分析双向信息上的效果要优于使用单层

的图神经网络对合并两个方向后的无向图进行分析。另外，在图神经网络中对根源节点的信息进行增强可以有效地提升谣言检测的效果，其证明了原始信息在谣言检测中的重要作用。并且 TPC-BiGAT 模型的效果比舍弃了门结构的模型的效果有所提升，表明模型中的门结构可以更好地学习原始信息对其他信息的影响，显示出门结构对检测效果的重要作用。

## 6 总结与展望

本文提出了一种基于潜在热点话题和图注意力神经网络的社交网络谣言检测模型。该模型能够过滤社交网络中的无关信息，发现其中的潜在热点话题，通过双向图注意力网络从两个不同的方向对这些话题的传播过程进行分析，最终进行话题级别的谣言检测以减少冷启动问题对检测效果的影响。三个公开数据集上的实验结果显示了本文模型的优良性能。

然而，当前的工作也存在一些局限性，在未来的工作中我们将继续围绕这些局限性进行研究。首先，进行话题级别的谣言检测需要大量数据的支持，而目前常用的公开数据集很难进行话题级别的谣言检测，在未来工作中可以构建新的、数据丰富的谣言检测数据集；其次，本文模型使用传播结构中每个节点的特征只考虑了谣言的文本信息，然而谣言还包含图片、视频等模态的信息，在未来工作中可以将本文模型与其他模态的信息融合以构建多模态检测模型。此外，未来工作中可以考虑优化图结构和神经网络架构，引入异构图、动态图和知识图等思想，以增强模型对谣言的表征效果。

## 参考文献

[1] CAO J, GUO J B, LI X, et al. Automatic rumor detection on microblogs: A survey [J]. arXiv preprint arXiv:1807.03505, 2018

[2] GUPTA A, KUMARAGURU P, CASTILLO C, et al. Tweetcred: Real-time credibility assessment of content on twitter [C]. International conference on social informatics, Springer, 2014: 228-243

[3] MORRIS M R, COUNTS S, ROSEWAY A, et al. Tweeting is believing? Understanding microblog credibility perceptions [C]. Proceedings of the ACM 2012 conference on computer supported cooperative work, 2012: 441-450

[4] SUN S Y, LIU H Y, HE J, et al. Detecting event rumors on sina weibo automatically [C]. Asia-Pacific web conference, Springer, 2013: 120-131

[5] LIU X M, NOURBAKHSH A, LI Q, et al. Realtime rumor debunking on twitter [C]. Proceedings of the 24th ACM international on conference on information and knowledge management, 2015: 1867-1870

[6] JIN Z W, CAO J, ZHANG Y D, et al. Novel visual and statistical image features for microblogs news verification [J]. IEEE transactions on multimedia, 2016, 19(3): 598-608

[7] 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, 2016: 3818-3824

[8] MA J, GAO W, WONG F 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 (Volume 1: Long Papers), 2017: 708-717

[9] BIAN T, XIAO X, XU T Y, et al. Rumor detection on social media with bi-directional graph convolutional networks [C]. Proceedings of the AAAI conference on artificial intelligence: volume 34, 2020: 549-556

[10] HUANG Q, ZHOU C, WU J, et al. Deep spatial–temporal structure learning for rumor detection on Twitter[J]. Neural Computing and Applications, 2023, 35(18): 12995-13005

[11] LIN H Z, YI P Y, MA J, et al. Zero-shot rumor detection with propagation structure via prompt learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence: Volume 37, 2023: 5213-5221

[12] LIU J W, XIE J Y, ZHANG F R, et al. Knowledge-Enhanced Hierarchical Information Correlation Learning for Multi-Modal Rumor Detection[J]. arXiv preprint arXiv:2306.15946, 2023

[13] PI DE C, WU Z Y, CAO JUAN. Early Rumor Detection Method Based on Knowledge Graph Representation Learning[J]. ACTA ELECTONICA SINICA, 2023, 51(2): 385

[14] TAN L, MA Z H, CAO J, et al. Rumor detection based on topic classification and multi-scale feature fusion [J]. Journal of Physics: Conference Series: volume 1601, 2020: 032032

[15] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter [C]. Proceedings of the 20th international conference on World wide web, 2011: 675-684

[16] YANG F, LIU Y, YU X H, et al. Automatic detection of rumor on sina weibo [C]. Proceedings of the ACM SIGKDD workshop on mining data semantics, 2012: 1-7

[17] KOWN S, CHA M Y, JUNG K, et al. Prominent features of rumor propagation in online social media [C]. IEEE 13th international conference on data mining, IEEE, 2013: 1103-1108

[18] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumours in social media: A survey [J]. ACM Computing Surveys (CSUR), 2018, 51(2): 1-36

[19] WU Z Y, PI D C, CHEN J F, et al. Rumor detection based on propagation graph neural network with attention mechanism [J]. Expert systems with applications, 2020, 158: 113595

[20] GUO H, CAO J, ZHANG Y Z, et al. Rumor detection with hierarchical social attention network [C]. Proceedings of the 27th ACM international conference on information and knowledge management, 2018: 943-951

[21] AJAO O, BHOWMIK D, ZARGARI S. Fake news identification on twitter with hybrid cnn and rnn models [C]. Proceedings of the 9th international conference on social media and society, 2018: 226-230

[22] MA J, GAO W, WONG F K. Rumor detection on twitter with tree-structured recursive neural networks [C]. Association for Computational Linguistics, 2018

[23] HUANG Q, ZHOU C, WU J, et al. Deep structure learning for rumor detection on twitter [C]. 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019: 1-8

[24] GROOTENDORST M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [J]. arXiv, 2022

[25] RONG Y, HUANG W B, XU T Y, et al. Dropedge: Towards deep graph convolutional networks on node classification [J]. arXiv preprint arXiv:1907.10903, 2019

[26] LIN H Z, MA J CHEN M F, et al. Rumor detection on twitter with claim-guided hierarchical graph attention networks [J]. arXiv preprint arXiv:2110.04522, 2021

[27] ZENG F Z, GAO W. Early Rumor Detection Using Neural Hawkes Process with a New Benchmark Dataset [C]. Proceedings

of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, 2022: 4105-4117

[28] MA J, GAO W, WEI Z Y, et al. Detect rumors using time series of social context information on microblogging websites [C]. The 24th ACM international on conference on information and knowledge management, 2015: 1751-1754

[29] LI G H, DONG M, MING L F, et al. Deep reinforcement learning based ensemble model for rumor tracking [J]. Information Systems, 2022

[30] WU Z Y, PI D C, CHEN J F, et al. Rumor detection based on propagation graph neural network with attention mechanism [J]. Expert systems with applications, 2020, 158: 113595

[31] THOTA R N, SUN X Y, DAI J. Early Rumor Detection in Social Media Based on Graph Convolutional Networks [C]. 2023 International Conference on Computing, Networking and Communications (ICNC), IEEE, 2023: 516-522

[32] YAO Y ROSCSCO L, CAPOMMETTO A. On early stopping in gradient descent learning [J]. Constructive Approximation, 2007, 26(2): 289-315

李劭，1999年生，硕士研究生，主要研究方向为社交网络谣言检测与自然语言处理。

蒋方婷， 1998 年生，硕士研究生，主要研究方向为社交网络谣言检测，自然语言处理与多模态融合。

杨鑫岩，1999年生，硕士研究生，主要研究方向为多媒体虚假信息检测与取证与计算机视觉。

梁刚，1976年生，博士，副教授，硕士生导师，主要研究方向为网络安全，网络舆情分析与预测和人工智能安全。

LI Shao，born in 1999, postgraduates. His main research interests include rumor detection and social network.

Liang Gang, born in 1976. PhD, associate professor, Master supervisor. His main research interests include network security, online public opinion analysis and prediction, and AI security.