doi:10.3772/j.issn.2095-915x.2026.02.002

基于深度学习的谣言检测研究

常芳瑞[1] 王增光[2] 贾萧丽[2] 王子琪[1] 袁学旺[2]

1. 山东省互联网舆情中心 济南 250000；
2. 潍坊市互联网舆情中心 潍坊 261000

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 网络时代下社交媒体使用更加广泛，互联网谣言随之大量涌现，给社会和公众带来诸多不利影响。如何准确、及时地发现谣言，学术界进行了大量研究和探索。传统谣言检测方法提取数据特征较为单一，而深度学习预测方法能够处理更复杂的特征结构，成为谣言检测的研究热点。[方法 / 过程] 文章对基于深度学习模型的谣言检测方法进行综述，详细介绍了深度学习架构循环神经网络、卷积神经网络、图卷积神经网络、注意力机制以及混合神经网络。其次，总结了现有文献中常用的国内外公共数据集。再次，根据前人的模型实验对谣言检测模型的性能和优劣、改进方向进行了分析。最后，探讨了深度学习谣言检测领域未来的研究方向。[局限] 未深入对各类模型的超参数和使用场景分析进行研究，后续需要进一步总结。[结果 / 结论] 对近年来各类深度模型在谣言检测方面的应用进行了系统梳理，帮助研究者快速了解研究范围和重点，为后续选择合适的模型提供一定参考。

关键词：谣言检测；深度学习；谣言数据集；神经网络

中图分类号：G202；G35

Research on Rumors Detection Based on Deep Learning

CHANG Fangrui[1] WANG Zengguang[2] JIA Xiaoli[2] WANG Ziqi[1] YUAN Xuewang[2]

1. Shandong Province Internet Public Opinion Center, Jinan 250000, China;
2. Weifang Internet Public Opinion Center, Weifang 261000, China

Abstract: [Objective/ Significance] In the Internet era, social media is more widely used, and rumors on the Internet have emerged in large quantities, bringing many adverse effects to society and the public. The academic community has conducted a lot of research and exploration on how to accurately and promptly discover rumors. Traditional rumor detection methods extract

作者简介 常芳瑞（1991-），本科，工程师，主要研究方向为网络舆情分析研判；王增光（1995-），通信作者，硕士，助理工程师，主要研究方向为数据挖掘，E-mail：1716231619@qq.com；贾萧丽（1994-），本科，助理工程师，主要研究方向为数据挖掘；王子琪（1998-），本科，助理工程师，主要研究方向为网络舆情分析研判；袁学旺（1988-），硕士，工程师，主要研究方向为网络传播分析。

引用格式 常芳瑞 , 王增光 , 贾萧丽 , 等 . 基于深度学习的谣言检测研究 [J]. 情报工程 , 2026, 12(2): 15-27.

doi:10.3772/j.issn.2095-915x.2026.02.002

data characteristics are relatively single, while deep learning prediction methods can handle more complex feature structures, becoming a hot topic in rumor detection. [Methods/Processes] The article reviews the rumor detection method based on deep learning models, and introduces in detail the deep learning architecture recurrent neural network, convolutional neural network, graph convolutional neural network, attention mechanism and hybrid neural network. Secondly, the commonly used domestic and foreign public data sets in the existing literature are summarized. Again, based on the model experiments of previous generations, the performance, advantages and disadvantages of the rumor detection model and the direction of improvement were analyzed. Finally, the future research directions in the field of deep learning rumors detection are discussed. [Limitations] The hyperparameters and usage scenario analysis of various models have not been further discussed, and further summary is needed in the future. [Results/Conclusions] The article systematically sorted out the applications of various in-depth models in rumors detection in recent years, helping researchers quickly understand the scope and focus of the research, and providing a certain reference for subsequent selection of suitable models.

Keywords: Rumor Detection; Deep Learning; Rumor Data Sets; Neural Networks

引言

信息化时代下，网络谣言传播速度和范围日益扩大，深刻影响公共安全和公众信任。2024 年度微博辟谣数据报告 [1] 显示，微博官方共有效处置不实信息 138608 条，辟除新增谣言及引导争议事件 1817 例。为有效处置网络谣言，对其实现快速准确鉴别，具有较为紧迫的现实意义。

公众普遍认为谣言是“未经官方证实的，让人们相信的一种主张”[2]，本文将谣言定义为一种在网络上传播且未经验证，或已被官方证实为假的信息。Zubiaga 等 [3] 指出，一个完整的谣言分类系统由谣言检测、谣言跟踪、立场分类和谣言验证组成，谣言检测的主要任务是从社交媒体中抽取不同的特征信息（如文本特征、用户特征、传播特征），判定其是否属于谣言。因谣言内容极具迷惑性且传播规律不固定 [4]，加之传播速度极快 [5]，这就要求谣言检测在具备高准确率的同时还要有很强的时效性，研究者们为此做了大量的工作。

目前网络谣言自动检测方法大致可分为两类：基于机器学习的检测方法与基于深度学习的检测方法。机器学习方法主要在人工特征工程基础上，使用决策树 [6-7]、支持向量机 [8-9]、随机森林 [10-11]、XGBoost[12] 等方法进行谣言检测。此类方法依赖于手工构建特征，特征提取类别较为单一，检测效果并不十分理想。深度学习方法能够从大量数据中自动学习判定谣言需要的深层语义特征，可以有效提高检测准确率，已成为目前谣言检测的研究热点。

国内当前对深度学习方法在谣言检测领域的综述类文章较少，为此本文总结了近年来深度学习在谣言检测领域的研究，总结了现有文献中常用的国内外公共数据集，对比与分析了现有深度学习方法，并探讨了未来谣言检测的发展方向。具体贡献如下：

（1）详细介绍了基于深度学习的谣言检测相关方法，为研究者选择合适的模型提供了参考；

（2）总结了现有文献中常用的国内外公共数据集；

（3）根据前人的模型实验对谣言检测模型的性能及优劣进行了系统对比分析；

（4）探讨了未来研究方法的发展趋势，为后续研究提供了新的视角与思路。

1 基于深度学习的谣言检测方法

传统机器学习模型提取特征相对简单，在处理高维数据集和复杂结构（如音频、图像、社交网络、交通网络等）时存在限制。因此，深度学习技术在谣言检测领域中的优越性逐渐显现。目前用于谣言检测的主流深度学习方法包括：循环神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Networks，CNN）、图神经网络（Graph Neural Networks，GNN）、注意力机制（Attention Mechanism，AM）、混合神经网络等，下面分别进行详细介绍。

1.1 循环神经网络

循环神经网络（RNN）能够捕捉序列数据中的长期依赖关系，适用于具有时序特征的谣言检测。单层 RNN 隐层 h 之间的前向传播过程为：

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

其中，x_t 为 t 时刻的输入，W_{xh} 是输入层到隐藏层的权重，W_{hh} 是隐藏层间的权重，b_h 是隐藏层的偏置，σ_h 是激活函数。

当序列过长时，权值矩阵载荷信息过多，容易发生梯度爆炸或梯度消失。为此，研究者对 RNN 进行了改进，提出了长短期记忆网络（Long Short Term Memory，LSTM）[13]。

LSTM 相对于 RNN 的隐层单元不再是传统的神经元节点，而是由一个记忆单元和三个门组成：即输入门 i_t、遗忘门 f_t 和输出门 o_t，可以选择性地控制信息的传输，能够更好地处理和预测较长序列数据。LSTM 实现包括以下步骤：

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \quad (2)$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \quad (3)$$

$$\tilde{C}_t = \tanh\left(W_C\left[h_{t-1}, x_t\right] + b_C\right) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \quad (6)$$

$$h_t = o_t * \tanh\left(C_t\right) \quad (7)$$

式（2）到（7）中，$\sigma(\cdot)$ 是激活函数，W_f、W_i、W_o、W_C 和 b_f、b_i、b_o、b_C 分别为权重矩阵和偏置向量，$\tilde{C}_t$ 为候选状态向量。由于 LSTM 较为复杂，Cho 等 [14] 在其基础上提出一种门控循环单元 GRU（Gated Recurrent Unit），结构简单且效果良好。

在目前的谣言检测中，大多数研究使用 RNN 来提取数据的时序信息。2016 年，Ma 等 [15] 首次选择使用循环神经网络进行谣言检测，通过对文本序列数据进行时序分析得到谣言上下文信息随时间变化的隐式特征。2018 年，Ma 等 [16] 提出一种树结构递归神经网络（Recursive Neural Network，RvNN）进行谣言检测。后 Chen 等 [17] 对循环神经网络方法进行改进，提出了一种基于 RNN 的深度模型，通过学习文本序列的时间隐式特征对谣言进行识别。为解决 RNN 的长序列依赖问题，Wu 等 [18] 使用 LSTM 对信息的传播路径进行表示和分类，解决了恶意传播者将虚假新闻伪装成真实新闻的问题。2019 年，Khattar 等 [19] 使用多个堆叠的双向 LSTM 来学习文本特征，结果显示优于对比模型。2020 年，尹鹏博等 [20] 使用 LSTM 对微博用户的历史行为

doi:10.3772/j.issn.2095-915x.2026.02.002

特征进行分析，实现了谣言早期检测。在 GRU 模型的应用中，Yuan 等 [21] 提出了一种基于注意力机制和 GRU 的网络谣言检测方法，利用双向 GRU 获取输入信息，并结合结构化注意力机制构建谣言检测模型；廖祥文等 [22] 通过带有注意力机制的双向 GRU 来提取微博间的时序信息，提高了检测的准确性。

1.2 卷积神经网络

除时序信息外，社交媒体数据还包括用户特征、传播特征等，仅依靠循环神经网络构建模型性能有限。在此情况下，有部分研究者选择使用卷积神经网络（CNN）来提取隐藏特征。以文本为例，假设文本 S 表示为 $W_{1:n}=w_1\oplus w_2\cdots\oplus w_n$，其中 n 为文本长度，w_i 表示 s 中的第 i 个词所对应的词向量，式中的⊕符号表示整合操作，$W_{i:j}$ 表示文本 s 中第 i 个词到第 j 个词对应的向量的首尾连接，输入层输入的就是文本 s 的相应词向量表示。卷积运算表示如下：

$$c_j=\sigma(m^T w_{j-h+1:j}+b) \tag{8}$$

其中，$w_{j-h+1:j}$ 表示文本 S 中的第 j–h+1 个词到第 j 个词对应的词向量，m^T 为卷积滤波器，b 表示偏置项，$\sigma(\cdot)$ 是激活函数，最终计算得到特征值 c_j。

池化层的作用是对特征值序列进行降采样操作，最后的全连接层对输出数据进行归一化操作，公式如下：

$$soft\max(f_i)=\frac{e^{f_i}}{\sum_{j=1}^{n}e^{f_j}} \tag{9}$$

$$f(w_i)=m\cdot w_i+b \tag{10}$$

$$y_i=soft\max(f(x)_i) \tag{11}$$

其中，$softmax(\cdot)$ 表示分类函数，$f(w)$ 表示全连接层，y_i 表示分类结果为第 i 类的概率。

最早将 CNN 应用于谣言检测的研究可以追溯到 2017 年，Yu 等 [23] 使用卷积神经网络从相关帖文中提取关键特征进行谣言检测。2018 年，Ajao 等 [24] 将 CNN 与 LSTM 相结合，用于检测 Twitter 中的虚假新闻。2019 年，Chen 等 [25] 提出利用卷积神经网络来提取分散在输入序列中的关键特征，实现了早期谣言的有效识别。2020 年，Zhou 等 [26] 使用假新闻文本和视觉内容进行嵌入，通过卷积神经网络提取特征图判断文本与图像描述之间的差异性。2022 年，周丽娜等 [27] 提出一种基于卷积神经网络的食品安全领域谣言检测方法，将实体特征与卷积神经网络提取的深度文本特征拼接融合，效果好于通用的谣言检测方法。

1.3 图卷积神经网络

卷积神经网络的应用证明了谣言检测方面的有效性，但实际工作中大多都是不规则的图数据，图卷积网络 [28]（Graph Convolutional Network，GCN）的特别之处在于能够处理非欧式结构（如社交网络、知识图谱等）的数据，且卷积运算也更符合实际。CNN 中的卷积通过在规则网格上通过滑动窗口对相邻元素进行加权求和，GCN 则是基于图的拓扑结构，通过消息传递机制让每个节点聚合邻居节点的信息来更新自身特征。在提取特征时，GCN 不仅考虑节点自身的特征，还充分利用图的拓扑结构信息，如节点的度、邻居节点的特征分布等，能够更好地捕捉图数据中的复杂关系特征。

构建图结构首先需要相应的邻接矩阵，我

们设图 $G=<E, V, A>$，其中 $E=\{e_{ij}\}$ 表示边的集合，$V=\{v_1, v_2, \cdots, v_n\}$ 表示节点的集合，A 表示邻接矩阵，e_{ij} 表示节点 v_i 和 v_j 之间通过边进行连接。拓扑图最简单的邻接矩阵可表示为：

$$\begin{cases} A_{ij}=1, \ if \ \ eij \in E \\ A_{ij}=0, \ otherwise \end{cases} \quad (12)$$

图卷积层会对图的节点特征和节点的邻接关系进行卷积，生成一个图嵌入表示向量，将该向量输入全连接神经网络，结合激活函数得到分类结果。GCN 层之间的传播过程为：

$$h^{(l)}=\tilde{A}(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}h^{(l-1)}W^{(l-1)}+b) \quad (13)$$

式中：$\tilde{A}=A+I$；A 为图邻接矩阵，是 $\tilde{D}$ 的 $\tilde{A}$ 度矩阵；I 为单位矩阵；$h^{(l-1)} \in R^{N\times C}$ 是每层的特征输入；N 是节点的数量，C 为节点的特征维度；$\sigma(\cdot)$ 是激活函数；$W^{(l-1)} \in R^{N\times F}$ 为可学习参数；F 为输出维度。

为获取谣言传播的高阶全局结构特征，有研究者开始应用图卷积网络模型。2020 年，Bian 等 [29] 提出谣言检测模型 Bi-GCN，通过双向图卷积网络学习消息转发结构，首次将 GCN 应用于谣言检测任务。2021 年，Wei 等 [30] 提出 EBGCN 改进模型，在 Bi-GCN 中添加了边权重推断模块提升谣言检测精度。同年，王昕岩等 [31] 根据评论之间的关联关系，提出一种加权图卷积神经网络 W-GCN，利用 GCN 学习评论的隐层表示，进而实现谣言检测。次年，Wang 等 [32] 提出一种区域增强的深度图卷积网络模型，通过学习区域化的传播模式来增强谣言的传播特征，探索谣言的区域化传播模式。针对 GCN 在谣言检测时不能充分考虑词义信息的问题，2022 年，王友卫等 [33] 提出一种基于图卷积网络的归纳式微博谣言检测方法，通过 GCN 和 GRU 实现节点信息聚合，提高微博谣言检测效率。此外，有学者开始研究基于 GCN 融合多源特征的方式，2023 年，关昌珊等 [34] 提出了一种基于图卷积网络的多特征融合模型 GCNs-BERT，模型同时融合了原文本特征、传播用户特征和传播结构特征，在新冠疫情数据集上泛化能力良好。

1.4 注意力机制

注意力机制（AM）最早受人类视觉启发，即将注意力集中在信息的关键点上，最早被应用于图像处理 [35]，后 Bahdanau 等 [36] 将注意力机制结合神经网络引入到自然语言处理任务。注意力机制能对提取的特征信息寻找其内部联系，并通过加权平均方式输出最终结果，从而提高模型的预测精度。注意力系数计算方式如下：

$$\alpha_i=soft\max(o_i)=\frac{\exp(score(o_i,O))}{\sum_{j=1}^{M}\exp(score(o_j,O))} \quad (14)$$

$$score(o_i,O)=o_i^T W_\alpha O \quad (15)$$

其中，O 代表最后输出，o 为各单元隐藏层输出，W_α 为可学习的参数矩阵，$score(\cdot)$ 为评分函数。

谣言检测中，注意力机制通常与其他模型融合使用。2017 年，Jin 等 [37] 提出了一种基于注意力的模型 att-RNN，该模型使用 RNN 和 VGG 网络（Visual Geometry Group，VGG）从文本和图像中提取特征，用注意力机制来学习文本、图像和传播结构之间的潜在表示，也是首个使用神经网络整合多模态特征进行谣言检测的模型。2019 年，Zhou 等 [38] 和 Liu 等 [39] 将注意力机制与图神经网络相结合提取数据特征，分别提出了 GEAR 模型和 KGAT 模型进行谣言

doi:10.3772/j.issn.2095-915x.2026.02.002

检测。后续研究中，有学者将以注意力机制为模型基础的 Transfommer 模型 [40] 应用到谣言检测中。2020 年，Khoo 等 [41] 提出一种后层注意力模型 PLAN，利用 Transfommer 的多头注意力机制模拟推文之间的长距离交互，进而实现谣言检测；Ma 等 [42] 使用 Transfommer 将信息的位置特征和结构特征相融合，提高了谣言检测准确率。之后的研究中，注意力机制逐渐被广泛使用。2021 年，Wang 等 [43] 提出了一种基于 GRU 和层次感知注意力机制的方法。2022 年，戚力鑫等 [44] 提出了一种基于自注意力机制的自适应注意力机制（Adaptive-SA），利用注意力机制捕捉与词的关键特征相关的视觉特征，实现了词和视觉对象之间的高层次信息交互。2023 年，冀源蕊等 [45] 提出一种基于 Attention 与 Bi-LSTM 的深度学习谣言识别方法，引入 Attention 机制实现谣言二次传播的有效识别。2024 年，凤丽洲等 [46] 提出一种基于图卷积网络和注意力机制的谣言检测方法，深入挖掘评论间的关联特性，经过验证可以有效提升谣言检测性能。

1.5 混合神经网络

近年来，还有大量研究者以传播树、对抗学习、强化学习、迁移学习、知识图谱等为基础，结合深度学习方法，用以提高谣言检测的准确性。2019 年，Ma 等 [47] 受到生成对抗网络（Generative Adversarial Network，GAN）中对抗学习方法的启发，首次使用类似 GAN 的方法来检测谣言。为获取时效性强且质量高的假新闻标注样本，Wang 等 [48] 于 2020 年在谣言检测任务中引入强化学习方法，提出了弱监督假新闻检测框架（WeFEND），该框架可以利用用户的举报信息作为弱监督来扩大假新闻检测的训练数据量。同年，Singhal 等 [49] 提出一种多模态框架 SpotFake+，使用迁移学习从新闻文章和相关图像中获取语义和上下文信息，从而提高假新闻检测的准确性。最近的研究中，王根生等 [50] 基于知识图谱提取背景知识，将它们和文本转化为加权图结构进行实时谣言检测；张鑫昕等 [51] 根据谣言传播路径构建了一种基于传播树的多特征谣言检测模型，准确率与 Bi-GCN 相比在不同数据集上分别提升 4.8%、4.2% 和 3%。

2 数据集统计

数据对深度学习模型训练十分重要，数据集的质量对模型预测性能和泛化性的影响不可忽视。表 1 对谣言检测研究领域内的常用数据集进行总结，并列出其基本统计信息。

表 1 常用数据集信息统计表

统计信息	Weibo	Twitter15	Twitter16	PHEME	FEVER
用户数	2746818	276663	173487	48843	
贴文数	3805656	331612	204820	197852	
事件数	4664	1490	818	6425	185455
真实谣言数	2351	372	205	1067	93367
虚假谣言数	2313	370	205	638	43107
非谣言数	0	374	205	4023	0
未证实谣言数	0	374	203	698	48971

其中，Weibo 数据集由 Ma 等 [15] 爬取新浪微博社区管理中心提供的谣言信息整理获得，在很多深度学习谣言检测模型对比实验中使用广泛，权威性较高；Twitter15 和 Twitter16 数据集同样由 Ma 等 [16] 根据 Twitter 网站上的推文及相关信息进行构建；PHEME 数据集由 Zubiaga 等 [52] 创建，包含了与九个公共事件相关的谣言；FEVER 数据集由 Thorne 等 [53] 整理发布，内容为非结构化文本，与现实社交网络语言相似。

此外，本文还总结了部分谣言检测研究用到的数据集，并根据其内容分为了纯文本数据集、多模态数据集、事实核查数据集。纯文本数据集主要有 CW-CURE 数据集、BioClaims 数据集、FakeHealth 数据集、LIAR 数据集、CoAID 数据集等 [54-58]，多模态数据集主要有 FakeNewsNet 数据集、FauxBuster 数据集、MR2 数据集、MuMiN 数据集、CHECKED 数据集等 [59-63]，数据核查数据集主要有 FEVEROUS 数据集、CHEF 数据集、MultiFC 数据集等 [64-66]。

3 模型对比

本文对深度学习在谣言领域的文献进行了总结与分析，按照不同的深度学习方法进行了分类，如表 2 所示。谣言检测作为分类任务，常用准确率（Accuracy）、精确率（Precision,）、召回率（Recall）和 F1 分值（F1-Score）等指标分析模型效果。

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

$$Pre = \frac{TP}{TP + FP} \quad (17)$$

$$Rec = \frac{TP}{TP + FN} \quad (18)$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \quad (19)$$

式中：*TP* 为正确分类的谣言事件数量，*TN* 为正确分类的非谣言事件数量，*FP* 为错误分类的谣言事件数量，*FN* 为错误分类的非谣言事件数量。由于数据集和各文献评价指标不尽相同，表中统一使用前人实验结果的准确率作为评价指标。

因谣言数据多源且复杂，给谣言检测任务带来了不小的挑战。考虑到循环神经网络能够捕捉上下文信息，Ma 等 [15] 基于谣言会被转发或评论，将帖文创建为连续变长时间序列，使用 RNN 对其建模，最简单的 RNN 模型在 Twitter、Weibo 数据集上的准确率分别达到 0.827 和 0.873，而改进的 LSTM 和 GRU 模型更优于前者，证明了谣言检测的有效性。之后在 RNN 基础上衍生出许多变体模型，如遵循推文非顺序传播结构的树结构递归神经网络 RvNN[16]；通过学习文本序列的时间隐式特征对谣言进行识别的 MVAE[20]，与 RNN 等基线模型相比均有效提高了谣言检测的准确性。考虑到谣言数据中事件之间的联系，研究者引入了卷积概念，如 CAMI[23] 中利用卷积和池化操作，可以从有限的输入序列中灵活地提取关键特征，相比于 RNN 在谣言早期检测中优势明显；SAFE[26] 通过 CNN 提取文本和图像特征，计算跨模态相似性（文本和图像间的相关性），综合利用三者进行谣言识别，实验证明多模态特征和跨模态关系对谣言检测作用显著。图卷积神经网络则进一步将卷积运算扩展到更适合表示谣言传播的图结构中，如 Bi-GCN[29] 从自顶向下和自底向上两个方向构建传播图和散布图，

doi:10.3772/j.issn.2095-915x.2026.02.002

表 2　深度学习谣言检测模型对比

分类	模型	主要方法	模型超参数	数据集	准确率	训练超参数
循环神经网络	tanh-RNN[15]	传统 RNN	词向量维度、隐藏层数量、激活函数等	Twitter	0.827	学习率、批次大小、迭代次数、正则化系数、优化器选择等
				Weibo	0.873	
	LSTM-1[15]	单层 LSTM		Twitter	0.855	
				Weibo	0.896	
	GRU-1[15]	单层 GRU		Twitter	0.864	
				Weibo	0.908	
	GRU-2[15]	在 GRU-1 基础上增加一个隐藏层		Twitter	0.881	
				Weibo	0.910	
	BU-RvNN[16]	自下而上树状结构的 RvNN 模型变体		Twitter15	0.708	
				Twitter16	0.718	
	TD-RvNN[16]	自上而下树状结构的 RvNN 模型变体		Twitter15	0.723	
				Twitter16	0.737	
	MVAE[20]	多个堆叠的双向 LSTM 来学习文本特征		Twitter	0.745	
				Weibo	0.824	
卷积神经网络	CAMI[23]	使用卷积神经网络提取文本特征	词向量维度、卷积层参数、池化层参数、全连接层参数、激活函数等	Twitter	0.777	
				Weibo	0.933	
	LSTM-CNN[24]	CNN 结合 LSTM		Twitter	80.38	
	ARC[25]	CNN 结合注意力残差网络		TREE	0.879	
				DOT	0.807	
	SAFE[26]	使用卷积神经网络提取文本和图像特征		PolitiFact	0.874	
				GossipCop	0.838	
图卷积神经网络	Bi-GCN[29]	使用双向图卷积提取传播结构特征	词向量维度、图结构参数（如节点数量、归一化方式）、图卷积层数、图卷积隐藏层数量、激活函数等	Twitter15	0.886	
				Twitter16	0.880	
				Weibo	0.961	
	EBGCN[30]	使用图卷积提取传播结构特征		Twitter15	89.2	
				Twitter16	91.5	
	RDGCN[32]			Twitter15	85.6	
				Twitter16	87.8	
	W-GCN[31]	赋予图卷积模型中连边的重		Weibo	0.956	
	GCNs·BERT[34]	基于图卷积融合多特征		Weibo	0.956	
				新冠疫情数据	0.925	
注意力机制	att-RNN[37]	注意力机制结合 RNN	注意力头数、掩码和约束相关超参数、最大序列长度等	Twitter	0.682	
				Weibo	0.788	
	PLAN[41]	多头注意力机制提取文本特征		Twitter15	84.5	
				Twitter16	87.4	
	Attention+Bi-LSTM[45]	注意力机制结合 Bi-LSTM		Twitter	94.3	
	TG-GCN[46]	注意力机制结合 GCN		Twitter15	0.860	
				Twitter16	0870	
混合模型	GAN-GRU[47]	融入生成对抗网络	生成器和判别器层数等	PHEME	0.781	
	WeFEND[48]	融入强化学习	价值估计、策略更新、经验回放、策略更新等相关超参数	WeChat	0.824	
	SpotFake+[49]	融入迁移学习	模型复用与冻结、特征提取等相关超参数	PolitiFact	0.846	
				GossipCo	0.856	
	GAGK-RD[50]	融入知识图谱	表示学习、评分函数相关超参数等	Weibo	92.1	
				PHEME	84.0	

挖掘谣言传播和散布特征，效果优于 RNN 和 CNN 方法；EBGCN[30] 针对谣言传播的不确定性，根据图的先验信念自适应地控制信息传递，重新调整边权重，效果好于基础的 GCN 模型。为了克服长序列数据丢失信息的问题，注意力机制被引入到谣言检测任务中，如与图卷积神经网络结合的 TG-GCN[46] 模型，在图卷积网络各节点信息传输过程中使用注意力机制区分源评论和其他评论对谣言检测的影响，实验结果优于 RNN 和 Bi-GCN 模型。随着深度学习技术的不断发展，多模型、多源数据融合 [59-63] 的思想被用于谣言检测并取得了不错的实验效果，这种方式或将成为谣言检测领域未来的研究趋势。

根据不同模型原理，本文分析总结了基于深度学习的谣言检测方法的优势、局限性和改进方向，如表 3 所示。

表 3　基于深度学习的谣言检测模型总结

模型	优势	局限性	改进方向
循环神经网络	可有效捕获数据中的时间依赖	不能有效捕获传播结构特征	引入能提取传播结构特征的方法，如传播树、卷积神经网络等
卷积神经网络	对图像、音频等多模态数据处理能力强	仅适合欧式特征数据，无法捕获数据中的时间依赖	与 RNN 等处理时间序列的模型相结合
图卷积神经网络	擅长处理谣言传播过程中的结构信息	处理动态图结构有一定的局限性	与其他神经网络模型、机器学习等方法融合
注意力机制	可以突出关键信息，与其他模型的结合能力强	一定程度上会忽略数据的全局信息	通过多模态融合、特征增强等方式，提高注意力机制对特征的学习能力
组合模型	将提取时间特征的模型与提取传播特征的模型相结合，预测精度更高	模型结构更加复杂、参数增加，导致计算成本增加、检测时间变长	对模块间交互进行优化、改进训练策略、共享部分参数等，降低模型复杂度

4　总结与展望

基于深度学习的谣言检测方法在深入挖掘信息特征和提升检测精度方面优势明显，但仍然存在以下问题亟待解决。

4.1　实时检测

网络上的信息不断更新和变化，谣言也在不断产生和传播。目前深度学习方法主要是基于以往的传播信息进行谣言检测，导致检测存在滞后性，而针对实时检测的模型和方法研究相对较少。同时，深度学习模型通常需要较大的计算资源和时间来处理数据，如何提升模型对大规模数据的处理能力、实时数据的接收与分析能力，进而实现实时检测是面临的一个重要挑战。

4.2　数据质量与规模

谣言检测需要大量的标注数据来训练和评估模型。同时，谣言的传播环境复杂多变，新的谣言类型不断涌现，这也增加了标注的难度。在数据规模方面，实际数据中谣言样本往往远少于非谣言样本，这种数据不平衡问题会导致模型在训练过程中出现“分类偏倚”，从而降低对谣言的检测准确率。由此来看，优化数据标注方式、均衡数据分布对后续提高深度学习模型的性能至关重要。

4.3　多模态数据融合

社交媒体上的信息不仅包含文本，还包括

doi:10.3772/j.issn.2095-915x.2026.02.002

图像、视频、音频等多种模态，多模态数据能够提供更丰富的线索，但如何有效地融合这些信息是一个挑战。目前针对多模态融合的方法还存在融合策略不优、特征对齐困难等问题，随着多媒体技术的发展，未来谣言检测需要更加注重多模态信息的融合。

4.4 模型泛化

现有的谣言检测模型往往是在特定的数据集和场景下训练的，当应用于不同的数据集或新的场景时，模型的泛化能力可能不足。例如，不同社交媒体平台的语言风格、用户行为习惯存在差异，同一模型在不同平台上的检测效果可能会有很大波动。此外，随着时间的推移，谣言的传播方式和特点也可能发生变化，模型需要具备适应这些变化的能力。如何开发动态的谣言检测模型、实时更新模型参数，以适应新出现的谣言类型和传播特点，是未来的一个重要研究方向。

参 考 文 献

[1] 2024 年度微博辟谣报告 [EB/OL]. (2025-01-20) [2025-06-27]. https://www.piyao.org.cn/20250114/ffc58c29441b4ac680643bca8178d10b/c.html.

[2] ALLPORT G W, POSTMAN L. The Psychology of Rumor[J]. American Sociological Review, 1948, 13(3): 361.

[3] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumours in socia lmedia: a survey[J]. ACM Computing Surveys, 2018, 51(2): 1-36.

[4] 杨延杰，王莉，王宇航 . 融合源信息和门控图神经网络的谣言检测研究 [J]. 计算机研究与发展，2021, 58(7): 1412-1424.

[5] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online[J]. Science, 2018, 359(6380): 1146-1151.

[6] ZHAO Z, RESNICK P, MEI Q. Enquiring minds: Early detection of rumors in social media from enquiry posts[C]// Proceedings of the 24th international Conference on World Wide Web. 2015: 1395-1405.

[7] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]// Proceedings of the 20th International Conferenceon World Wide Web. New York: ACM, 2011: 675-684.

[8] 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. IEEE, 2013: 1103-1108.

[9] WU K, YANG S, ZHU K Q. False rumors detection on Sina Weibo by propagation structures[C]// Proceedings of the IEEE 31th International Conference on Data Engineering, Seoul, Korea, 2015: 651-662.

[10] ANG F, YU X H, LIU Y, et al. Automatic detection of rumor on Sina Weibo[C]// Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. NewYork: ACM, 2012: 1-7.

[11] 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 Conference on Information and Knowledge Management. New York: ACM, 2015: 1751-1754.

[12] GUMAEI A, AL-RAKHAMI M S, HASSAN M M, et al. An effective approach for rumor detection of Arabic tweetsusing eXtreme gradient boosting method[J]. ACM Transactions on Asian and Low-resource Language Information Processing, 2022, 21(1): 1-16.

[13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[14] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]// Proceedings of the 2014 Conference on

Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

[15] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: IJCAI, 2016: 3818-3824.

[16] MA J, GAO W, WONG K F. Rumor detection on Twitter with tree-structured recursive neural networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1980-1989.

[17] CHEN T, LI X, YIN H, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2018: 40-52.

[18] WU L, LIU H. Tracing fake-news footprints: Characterizing social media messages by how they propagate[C]// Proceedings of The Eleventh ACM International Conference on Web Search and Data Mining. 2018: 637-645.

[19] KHATTAR D, GOUD J S, GUPTA M, et al. Mvae: multimodal variational autoencoder for fake news detection[C]// The World Wide Web Conference. 2019: 2915-2921.

[20] 尹鹏博 , 潘伟民 , 彭成 , 等 . 基于用户特征分析的微博谣言早期检测研究 [J]. 情报杂志 , 2020, 39(7): 81-86.

[21] YUAN S D, YANG Q. Network rumor detection using attention mechanism and BiGRU neural network in big data environment[J]. Journal of Circuits, Systems and Computers, 2024, 33(1): 2450009.

[22] 廖祥文 , 黄知 , 杨定达 , 等 . 基于分层注意力网络的社交媒体谣言检测 [J]. 中国科学 : 信息科学 , 2018, 48(11): 1558-1574.

[23] YU F, LIU Q, WU S, WANG L, et al. A convolutional approach for misinformation identification[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 3901-3907.

[24] 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.

[25] CHEN Y X, SUI J, HU L, et al. Attention-residual network with CNN for rumor detection[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management, New York, USA, 2019: 1121-1130.

[26] ZHOU X, WU J, ZAFARANI R: SAFE: similarity-aware multi-modal fake news detection[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2020: 354-367.

[27] 周丽娜 , 谭励 , 曹娟 , 等 . 基于卷积神经网络的食品安全领域谣言检测方法 [J]. 计算机应用与软件 , 2022, 39(3): 45-50.

[28] 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. Banff: ICLR, 2014. arXiv:1312.6203.

[29] BIAN T, XIAO X, XU T, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 549-556.

[30] 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. Online: ACL, 2021: 3845-3854.

[31] 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.

[32] WANG G, TAN L, SONG T, et al. Region-enhanced deep graph convolutional networks for rumor detection[J]. arxiv preprint arxiv: 2206. 07665, 2022.

[33] 王友卫 , 童爽 , 凤丽洲 , 等 . 基于图卷积网络的归纳式微博谣言检测新方法 [J]. 浙江大学学报（工学版）, 2022, 56(5): 956-966.

[34] 关昌珊 , 邴万龙 , 刘雅辉 , 等 . 基于图卷积网络的多特征融合谣言检测方法 [J]. 郑州大学学报（工

doi:10.3772/j.issn.2095-915x.2026.02.002

学版) , 2024, 45(4): 70-78.

[35] XIAO T, XU Y, YANG K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842-850.

[36] BAHDANAU D, CHO K, BENGIO Y.Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-01) [2025-04-26]. https://arxiv.org/abs/1409.0473.

[37] JIN Z, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017: 795-816.

[38] ZHOU J, HAN X, YANG C, et al. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 892-901.

[39] LIU Z, XIONG C, SUN M, et al.Fine-grained fact verification with kernel graph attention network[C]// Proceedings of the 58th annual meeting of the association for computational linguistics, 2020: 7342-7351.

[40] VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 5998-6008.

[41] KHOO L M S, CHIEU H L, QIAN Z, et al. Interpretable rumor detection in microblogs by attending to user interactions[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(5): 8783-8790.

[42] MA J, GAO W. Debunking rumors on twitter with tree transformer[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 5455-5466.

[43] WANG Y, ZHOU W, WEN J, et al. GRU with Level-Aware Attention for Rumor Early Detection in Social Networks[C]// Neural Information Processing: 28th International Conference. 2021: 548-559.

[44] 戚力鑫 , 万书振 , 唐斌 , 等 . 基于注意力机制的多模态融合谣言检测方法 [J]. 计算机工程与应用 , 2022, 58(19): 209-217.

[45] 冀源蕊 , 康海燕 , 方铭浩 . 基于 Attention 与 Bi-LSTM 的谣言识别方法 [J]. 郑州大学学报 (理学版), 2023, 55(4): 16-22.

[46] 凤丽洲 , 刘馥榕 , 王友卫 . 基于图卷积网络和注意力机制的谣言检测方法 [J]. 数据分析与知识发现 , 2024, 8(4): 125-136.

[47] MA J, GAO W, WONG K F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning[C]// The World Wide Web Conference. 2019: 3049-3055.

[48] WANG Y, YANG W, MA F, et al. Weak supervision for fake news detection via reinforcement learning[C]// Proceedings of the AAAI Conference on Aartificial Intelligence. 2020, 34(1): 516-523.

[49] SINGHAL S, KABRA A, SHARMA M, et al. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)[C]// Proceedings of the AAAI conference on artificial intelligence. 2020, 34(10): 13915-13916.

[50] 王根生 , 朱奕 , 李胜 . 一种融合知识图谱的图注意力神经网络谣言实时检测方法 [J]. 数据分析与知识发现 , 2024, 8(6): 95-106.

[51] 张鑫昕 , 潘善亮 , 茅琴娇 . 基于传播树的多特征谣言检测方法 [J]. 电子学报 , 2024, 52(5): 1609-1618.

[52] 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): e0150989.

[53] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. Fever: A large-scale dataset for fact extraction and verification[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. New Orleans: ACL, 2018: 809-819.

[54] GOLLAPALLI S D, DU M, NG S K. Identifying checkworthy cure claims on twitter[C]// Proceedings of the ACM Web Conference 2023. 2023: 4015-4019.

[55] AMELIE WÜHRL, KLINGE R. Claim detection in biomedical Twitter posts [C]// Proceedings of the 20th Workshop on Biomedical Language Processing. 2021: 131-142.

[56] DAI E, SUN Y, WANG S. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository[C]// Proceedings of the International AAAI Conference on Web and Social Media. 2020(14): 853-862.

[57] WANG W. “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. 2017: 422-426.

[58] CUI L, LEE D. COVID-19 healthcare misinformation dataset[DS/OL]. (2020-05-01) [2025-04-20]. https://doi.org/10.48550/arXiv.2006.00885.

[59] SHU K, MAHUDESWARAN D, WANG S, et al. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171-188.

[60] ZHANG D Y, SHANG L, GENG B, et al. Fauxbuster: A content-free fauxtography detector using social media comments[C]// 2018 IEEE International Conference on Big Data IEEE. 2018: 891-900.

[61] HU X, GUO Z, CHEN J, et al. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media[C]// Proceedings of the 46th international ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 2901-2912.

[62] NIELSEN D S, MCCONVILLE R. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset[C]// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 3141-3153.

[63] YANG C, ZHOU X, ZAFARANI R. CHECKED: Chinese COVID-19 fake news dataset[J]. Social Network Analysis and Mining, 2021, 11(1): 58.

[64] ALY R, GUO Z, SCHLICHTKRULL M, et al. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task[C]// Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) at EMNLP. 2021: 1-13.

[65] Hu X, Guo Z, Wu G Y, et al. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 3362-3376.

[66] AUGENSTEIN I, LIOMA C, WANG D, et al. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4685-4697.

（责任编辑：兰天）