A Rumor Detection Method Based on Deep Learning*

JIANG Minmin[1*], *BAN Hao*[2], *ZHAO Li*[2]

(1.*School of Network and Communication, Nanjing Vocational College of Information Technology, Nanjing Jiangsu* 210023, *China*;
2.*School of Information Science and Engineering, Southeast University, Nanjing Jiangsu* 210096, *China*)

Abstract: In order to better learn the characteristic changes in the process of network rumor propagation, a network rumor detection method based on multi hop multimodal fusion is proposed. Faster RCNN is used to extract visual features, GRU is used to extract word features, and BERT is used to extract sentence features. After extracting basic features of words and sentences, RGCN is used to realize information transmission between different nodes in the graph. After extracting multimodal features, multi hop attention mechanism is used to detect rumors. This method can solve complex problems such as negation, ambiguity and long-distance dependency, and can capture remote dependency in a shorter path. By comparing with other rumor detection methods, the effectiveness of this method in the field of rumor detection and even early rumor detection is verified.

Key words: rumor detection; deep learning; multi hop network; attention mechanism; multimodality

EEACC: 6140 **doi**: 10.3969/j.issn.1005-9490.2022.06.027

一种基于深度学习的谣言检测方法*

姜敏敏[1*], 班 浩[2], 赵 力[2]

(1.南京信息职业技术学院网络与通信学院, 江苏 南京 210023; 2.东南大学信息科学与工程学院, 江苏 南京 210096)

摘 要: 为了更好地学习网络谣言传播过程中的特征变化, 提出了一种基于多跳的多模态融合的网络谣言检测方法。该方法采用 faster RCNN 提取视觉特征, 通过 GRU 提取词特征, 通过 BERT 提取句子特征, 在提取词句基本特征后, 利用 RGCN 实现图中不同节点间的信息传递。提取多模态特征后利用多跳注意力机制实现谣言检测。该方法可以较好解决诸如否定、歧义和长距离依赖等复杂问题, 可以在更短路径上捕获远程依赖。通过与其他谣言检测方法的对比实验, 验证了该方法在谣言检测, 甚至早期谣言检测领域应用的有效性。

关键词: 谣言检测; 深度学习; 多跳网络; 注意力机制; 多模态

中图分类号: TP391 **文献标识码**: A **文章编号**: 1005-9490(2022)06-1429-05

随着互联网的飞速发展, 网络上的各种消息爆发式增长, 其中充斥着各种各样的虚假信息。这些虚假信息会混淆视听, 轻则影响大众的判判, 重则影响社会的稳定。因此如何识别网络谣言, 且及早阻止谣言的传播就显得异常重要了。

随着人工智能的飞速发展, 利用人工智能自动判别网络谣言具有重要的实用价值。在人工智能发展的早期, 通常通过人工构造网络谣言鉴别分类器, 通过神经网络进行检测。例如, Xia 等[1]采用半自动化方法在时空上做聚类, 从中提取出密度、直径较大的簇进行人工判断, 评估 Twitter 平台上的特定主题信息是否可靠。这类方法存在人工判别误差、数据库覆盖不全面等问题。

当前随着深度学习的迅猛发展, 机器学习能力有了重大发展。深度学习在机器视觉、语音识别、自然语言处理、统计建模等领域快速发展, 在越来越多的领域扮演着重要角色。

在自然语言处理领域, 深度学习使用分布式向量逐级表示词、短语、句子和篇章等语言单元, 使用神经网络对不同大小的语言单元向量进行合并, 形成语言单元的组合。除了语言单元外, 不同种类的模态如语言、图像等都可以通过类似的合并方式表示在相同的空间中, 从而实现各种复杂的任务。与传统的谣言检测方法相比, 深度学习可以自主学习

项目来源: 江苏省“青蓝工程”优秀青年骨干教师培养项目
收稿日期: 2021-10-09 **修改日期**: 2021-11-05

到更深层次的谣言特征,而不像传统方法一样依赖人工和经验,因此利用深度学习进行网络谣言的检测是行之有效的。

在谣言检测领域,深度学习也有了越来越多的应用研究成果。一些文献对谣言的各种特征进行了分析与建模。在文献[2]中,作者探讨了社交媒体谣言的时间、结构和语言特征,并将这些特征结合起来,帮助更准确地识别谣言。在文献[3]中,作者基于发布者信息、文本信息、转发评论者信息三种特征,构建递归神经网络(Recursive Neural Networks,RNN)实现网络谣言的自动检测。在文献[4]中,作者通过深度多层感知机(Multi-layer Perceptron,MLP)结合神经网络特征、统计特征和人工设计的谣言特征,进行新闻立场检测。

一些文献从信息传播途径入手进行谣言检测。在文献[5]中,作者提出了一种基于异构用户表示的信息传播模型,以观察谣言和可信消息传播模式的区别并加以区分。在文献[6]中,作者将用户的行为作为隐藏线索来发现可能的谣言传播,并得出了基于群体行为的谣言检测优于基于微博固有特征的谣言检测的结论。在文献[7]中,作者开发了一个两层网络来模拟流行病传播和信息扩散之间的相互作用,得到了知识对谣言的渗透强度起着至关重要的作用。在文献[8]中,作者使用循环神经网络,提出了一种基于核的传播树方法来模拟微博文章的扩散,通过分析不同类型谣言传播树结构之间的相似性来获取区分不同类型谣言的高阶模式。在文献[9]中,作者提出了一种基于图核的混合支持向量机(Support Vector Machine,SVM)分类器,能够捕获高阶消息传播模式以及语义特征,用于自动检测新浪微博上的虚假谣言。

还有一些文献提出了谣言的预测方法。在文献[10]中,作者提出了一种基于卷积神经网络(Convolutional Neural Networks,CNN)的卷积误报识别方法,可以灵活地提取分散在输入序列中的关键特征,形成重要特征之间的高层交互,有效地识别错误信息,用于谣言的预测。在文献[11]中,作者提出了一种基于内容特征以及伪反馈的算法来预测未来谣言。

另有一些文献从算法结构上进行了研究。在文献[12]中,作者提出了一种分布式计算方法来计算社交网络中每个用户网络中心度的值,以此来衡量网络节点在网络中重要性。在文献[13]中,基于用户的评分、评论和社交数据,作者提出了一种混合推荐模型来提高推荐精度。在文献[14]中,作者提出了一种基于主题分类和多尺度融合的谣言检测方法,从不同尺度的不同子数据集中提取特征,综合考虑整体、主题间、主题内的相关性和差异性,进行特征融合后再进行判断。

以上这些文献在不同的角度对谣言检测进行了探讨,得到了很好的检测效果。但是,现有检测方法没有很好考虑谣言信息传播过程中出现特征的变迁,为了解决这个问题,本文提出了一种基于多跳的多模态融合的网络谣言检测方法。

文献[15]提出层次注意力机制并应用于文本情感分析,实验表明该模型能够从文本数据中学习到较为丰富的情感信息。多层注意力机制是对情感极性做出判断的流行方法,而多跳推理是在不完全知识图谱上解析问答的有效方法。在前人的研究基础上,本文采用基于分层注意力机制的网络谣言检测模型作为鉴别器,通过多跳推理获取网络谣言的深层抽象特征,从而提高模型的识别准确率和鲁棒性。

1 基于多跳推理的谣言检测方法

网络谣言信息是一种时序数据,目前基于注意力机制的循环神经网络模型[16]可以捕获谣言特征随时间的变化,提升谣言检测的准确率。但是由于网络谣言信息在传播的过程中也会出现文本语义、关键特征的变化,为了提升算法的泛化能力,需要根据谣言传播过程中的多篇文本关联分析,从“多跳”的角度出发构建新的算法。图 1 所示为用于网络谣言检测的多跳网络模型。

多模态谣言信息通过更快速的区域生成卷积神经网络特征(faster Region with CNN Features,faster RCNN)提取视觉特征,通过门控循环单元(Gate Recurrent Unit,GRU)提取词特征,且通过来自变压器的双向编码器表示(Bidirectional Encoder Representations from Transformers,BERT)提取句子特征。采用前期融合的方法,将这些特征向量输入一个全连

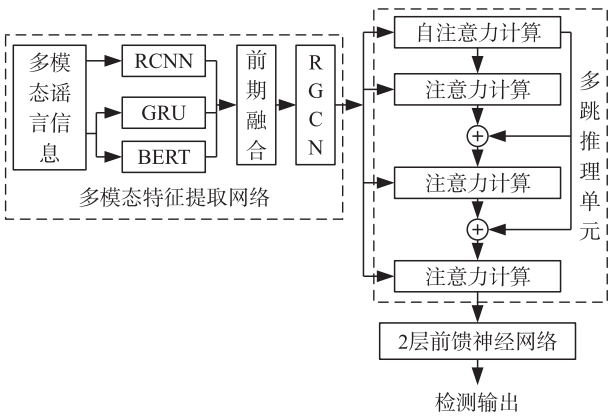

图 1 多跳检测模型

接层,得到每个模态的自注意力权重,进行加权求和得到前期融合特征。

提取融合特征后,利用关系图卷积网络(Relational Graph Convolutional Network, RGCN)[17]实现图中不同节点间的信息传递。为了处理高相关数据特征,使用在第 l 层 RGCN 中给定节点和领域索引的隐状态并关联到邻居,其下一层的隐状态更新为:

$$\boldsymbol{h}_i^{l+1}=\sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^l\boldsymbol{h}_j^l+\boldsymbol{W}_0^l\boldsymbol{h}_i^l\right) \tag{1}$$

式中:上标 l 表示神经网络第 l 层;下标 i 表示第 i 个节点;下标 j 表示第 j 个邻域索引;下标 r 表示第 r 个关系;R 为正则方向和逆方向关系的集合;N_i^r 表示关系 r 下节点 i 的邻域索引;$c_{i,r}$ 为第 i 个节点、第 r 个关系的标准化因子;$\boldsymbol{W}_r^l$ 为神经网络第 l 层特殊关系权重矩阵;$\boldsymbol{W}_0^l$ 为神经网络第 l 层普通权重矩阵;$\boldsymbol{h}_i^l$ 是神经网络第 l 层节点 v_i 的隐藏状态。

在得到节点的隐藏状态 $\boldsymbol{h}_i^{l+1}$ 后,使用注意力机制来平衡不同状态之间的关系以及对整体谣言倾向的贡献,其公式如下所示:

$$\begin{cases}\boldsymbol{u}_i=\tanh(\boldsymbol{W}_r^l\boldsymbol{h}_i^{l+1}+\boldsymbol{b}_r)\\ \boldsymbol{\alpha}_i=\dfrac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_s)}{\sum_i\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_s)}\\ \boldsymbol{v}_i=\sum_i\boldsymbol{\alpha}_i\boldsymbol{h}_i^{l+1}\end{cases} \tag{2}$$

式中:$\boldsymbol{b}_r$ 为上一时间步的记忆内容,$\boldsymbol{\alpha}_i$ 为每个时间步隐藏向量,$\boldsymbol{u}_s$ 表示用于计算注意力权重的查询向量,$\boldsymbol{v}_i$ 表示谣言的特征向量。

由多个注意力计算层叠加的深度模型可以较好解决诸如否定、歧义和长距离依赖等复杂问题,可以更有效应对谣言检测的应用。在本文的多跳推理单元中,下一跳注意力计算时参考上层的历史,从而将前一层的表示输出向更抽象的级别转换。即通过式(2)重新为每个节点计算注意力权重得到第 t 跳的谣言特征向量 $\boldsymbol{v}^t$,再使用额外的 GRU 网络应用到每一跳的谣言特征向量得到最终的谣言特征输出。通过足够多的跳数,输入元素借助注意力的递归计算充分交互,可以在更短路径上捕获远程依赖。

2　实验结果

本文采用 Ma 等人在文献[18]中公布的用于网络谣言检测研究的数据集。该数据集中包含有 Twitter 和新浪微博的数据,是谣言检测领域的经典数据集。在实验中,由隐藏维数为 512 的 GRU 提取维度为 100 的词特征,由 BERT 模型提取维度为 768 的句子特征。采用预先训练的 faster RCNN 来提取 2 048 维的图片特征,每张图片提取 100 个对象,然后用 512 维度的全连接层来嵌入图片特征。多头注意力的数量设置为 4 个,每个注意头的尺寸为 64。多跳推理的跳数设置为 4。模型的批处理大小(batch_size)设置为 32,训练批次大小设置为 64,学习率为 0.001。模型参数随机初始化,最大迭代轮数设为 500。为了和其他文献进行对比,选择了几种检测方法,分别简记为:CNN[10]、PFR[11]、MSFF[14]。本文方法则简记为:HRGCN。

基于新浪微博和 Twitter 数据集,本文方法与其他几种检测方法在谣言检测任务上得到的实验结果如表 1 所示,采用正确率(Accuracy)、准确率(Precision)、召回率(Recall)、F1 等四个评价指标作为实验的评价标准。

从表 1 可以看出,CNN 模型检测效果最差,本文方法检测效果最佳,这是因为本文方法采用了多跳方法,对远程依赖捕获程度较高,对谣言的识别率也较好。从实验结果可以看出,本文方法的整体谣言识别率达到了 90%以上,可以很好识别谣言。

为了验证多跳环节的效果,测试了不同跳数条件下新浪微博数据的谣言检测,其结果如表 2 所示。

表 1　谣言检测结果(R:谣言,N:非谣言)

方法	类别	新浪微博				Twitter			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
CNN	R	0.901	0.854	0.852	0.853	0.904	0.818	0.872	0.844
	N		0.827	0.834	0.844		0.845	0.842	0.843
PFR	R	0.922	0.905	0.849	0.876	0.913	0.820	0.883	0.850
	N		0.856	0.927	0.890		0.868	0.827	0.847
MSFF	R	0.937	0.905	0.913	0.909	0.928	0.823	0.878	0.850
	N		0.909	0.879	0.894		0.897	0.870	0.883
HRGCN	R	0.942	0.958	0.951	0.954	0.941	0.902	0.890	0.896
	N		0.967	0.959	0.963		0.915	0.909	0.912

表 2 不同跳数的谣言检测结果(R:谣言,N:非谣言)

跳数	类别	新浪微博			
		Accuracy	Precision	Recall	F1
1 跳	R	0.881	0.890	0.885	0.887
	N		0.892	0.890	0.891
2 跳	R	0.909	0.916	0.911	0.913
	N		0.926	0.921	0.923
3 跳	R	0.932	0.937	0.934	0.935
	N		0.948	0.942	0.945
4 跳	R	0.942	0.958	0.951	0.954
	N		0.967	0.959	0.963
5 跳	R	0.942	0.959	0.952	0.955
	N		0.969	0.960	0.964

从表 2 可以看出,跳数越多谣言检测的效果越好。但是从 5 跳开始检测效果提升很小,因此选择 4 跳比较适当。

为了进一步确认本文方法在早期谣言检测中的性能,本文选择了事件相关消息发出后的 6 个时间节点,即 1 h、3 h、6 h、12 h、24 h、36 h,截止时间设置为 36 h,仅以当前时间节点范围内的样本作为数据输入,以此来评估本文方法在早期谣言检测中的效果。采用 4 跳结构,最终得到的实验结果如图 2 和图 3 所示。

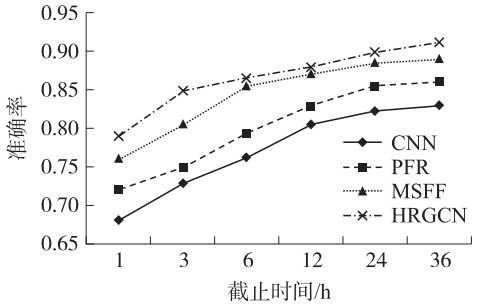

图 2 早期谣言检测结果(Twitter)

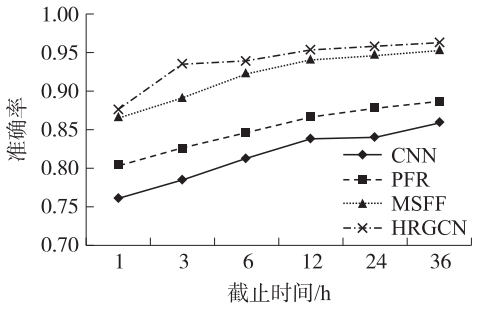

图 3 早期谣言检测结果(新浪微博)

从图 2 和图 3 可以看出,在谣言传播的早期,由于缺乏数据,不同模型的准确率都较低,但随着数据的增多,模型的识别准确率也呈现快速上升的趋势。相对于其他方法,本文方法引入了多跳推理结构,能够在谣言初期更快地学习到深层次特征,能提升谣言初期检测的准确度。以上实验数据说明本文方法在早期谣言检测任务上是有效的。

3 结论

针对网络谣言检测问题,提出了一种结合多跳、注意力机制的深度学习网络。在谣言特征提取模块中采用 faster RCNN、GRU、BERT、RGCN 等网络提取谣言的多模态特征。在多跳推理单元中,利用多跳注意力机制来深度融合谣言的多模态特征。通过对比实验,验证了本文方法具有较好的谣言检测率。

参考文献:

[1] Xia X,Yang X H,Wu C,et al. Information Credibility on Twitter in Emergency Situation[C]//Pacific Asia Conference on Intelligence and Security Informatics,Kuala Iumpur,Malaysia,2012:45-59.

[2] Kwon S J,Cha M Y,Jung K M,et al. Prominent Features of Rumor Propagation in Online Social Media[C]//2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 2013: 1103-1108.

[3] Ruchansky N,Seo S,Liu Y. CSI:A Hybrid Deep Model for Fake News Detection[C]//ACM 2017 Conference on Information and Knowledge Management,Singapore,2017:797-806.

[4] Bhatt G,Sharma A,Sharma S,et al. Combining Neural,Statistical and External Features for Fake News Stance Identification[C]// Companion Proceedings of the Web Conference 2018, Lyon, France,2018:1353-1357.

[5] Yang L,Xu S. Detecting Rumors Through Modeling Information Propagation Networks in a Social Media Environment[J]. IEEE Transactions on Computational Social Systems, 2017, 3(2): 46-62.

[6] Liang G,He W,Xu C,et al. Rumor Identification in Microblogging Systems Based on Users Behavior[J]. IEEE Transactions on Computational Social Systems,2015,2,99-108.

[7] Huang H,Chen Y H,Ma Y F. Modeling the Competitive Diffusions of Rumor and Knowledge and the Impacts on Epidemic Spreading [J]. Applied Mathematics and Computation,2021,388:125536.

[8] Ma J,Gao W,Wong K F. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning[C]//55th Annual Meeting of the Association for Computational Linguistics,Vancouver,Canada, 2017:708-717.

[9] Wu K,Yang S,Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]//IEEE 31st International Conference on Data Engineering,Seoul,South Korea,2015:651-662.

[10] Yu F,Liu Q,Wu S,et al. A Convolutional Approach for Misinformation Identification[C]//26th International Joint Conferences on Artificial Intelligence Melbourne,Australia,2017:3901-3907.

[11] Qin Y M,Dominik W,Tang C C. Predicting Future Rumours[J]. Chinese Journal of Electronics,2018,27(3):514-520.

[12] Ranjan K B,Santanu K R,Sanjay M,et al. Distributed Centrality Analysis of Social Network Data Using MapReduce [J].

Algorithms,2019,12(8):161.

[13] Ji Z Y,Pi H Y,Wei W,et al. Recommendation Based on Review Texts and Social Communities:A Hybrid Model[J]. IEEE Access,2019,7:40416-40427.

[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,2020,1601(3):032032.

[15] Yang Z C,Yang D R,Dyer C,et al. Hierarchical Attention Networks for Document Classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,San Diego,CA,USA,2016:1480-1489.

[16] Xu N,Chen G,Mao W. MNRD:A Merged Neural Model for Rumor Detection in Social Media[C]//2018 International Joint Conference on Neural Networks,Rio de Janeiro,Brazil,2018:1067-1073.

[17] Cao Y,Fang M,Tao D C. BAG:Bi-Directional Attention Entity Graph Convolutional Network for Multi-Hop Reasoning Question Answering[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,Minneapolis,MN,USA,20119:357-362.

[18] Morris M R,Counts S,Roseway A,et al. Tweeting is Believing? Understanding Microblog Credibility Perceptions[C]//ACM 2012 Conference on Computer Supported Cooperative Work,Seattle,WA,USA,2012:441-450.

姜敏敏(1983—),女,安徽蚌埠人,硕士,副教授,研究方向为通信与信号处理,人工智能。