社会网络谣言检测综述

高玉君[1],梁 刚[1],蒋方婷[1],许 春[2],杨 进[1],陈俊任[1],王 浩[3]
(1. 四川大学网络空间安全学院,四川成都 610065; 2. 四川大学信息管理中心,四川成都 610065;
3. 成都信息工程大学软件工程学院,四川成都 610225)

摘 要: 当前社会网络已取代传统媒体成为信息交流的重要平台,社会网络中的信息具有传播速度快,范围广,即时性强等优点.然而,由于发布信息时缺乏有效的监管手段,导致社会网络平台同时也成为谣言传播的温床.因此,快速有效地检测出社会网络谣言,对净化网络环境,维护公共安全至关重要.本文首先对谣言定义进行阐述,并描述当前谣言检测的问题及检测过程;其次,介绍不同数据获取方式并分析其利弊,同时对比谣言检测中不同的数据标注方法;第三,根据谣言检测技术的发展对现有的人工、机器学习和深度学习的谣言检测方法进行分析对比;第四,通过实验在相同公开数据集下对当前主流算法进行实证评估;最后,对社会网络谣言检测技术面临的挑战进行归纳并总结全文.

关键词: 社会网络; 谣言检测; 网络空间安全
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2020)07-1421-15
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.07.023

Social Network Rumor Detection: A Survey

GAO Yu-jun[1], LIANG Gang[1], JIANG Fang-ting[1], XU Chun[2], YANG Jin[1], CHEN Jun-ren[1], WANG Hao[3]
(1. *College of Cyber Security, Sichuan University, Chengdu, Sichuan* 610065, *China*;
2. *Information Management Center, Sichuan University, Chengdu, Sichuan* 610065, *China*;
3. *Software Engineering Institute, Chengdu University of Information Technology, Chengdu, Sichuan* 610225, *China*)

Abstract: The current social network has replaced traditional media as an important platform for information exchange. The information in social networks has the advantages of fast dissemination, wide range, and strong immediacy. However, due to the lack of effective supervision means when publishing information, the social network platform has also become a hotbed of rumors. Therefore, the rapid and effective detection of social network rumors is essential for purifying the network environment and maintaining public safety. Firstly, this article explains the definition of rumors, and the problems of current rumors detection and detection process are described. Secondly, different data acquisition methods are introduced and their advantages and disadvantages are analyzed. At the same time, different data annotation methods in rumor detection are compared. Thirdly, according to the development of rumor detection technology, analyze and compare the existing rumors detection methods of artificial, machine learning and deep learning. Fourthly, current mainstream algorithms are empirically evaluated under the same open data set through experiments. Finally, analyze and summarize the challenges faced by current social network rumor detection technology.

Key words: social network; rumor detection; cyberspace security

1 引言

社会网络爆炸式的发展使得以微博、Twitter、微信等为代表的社会网络已逐渐取代传统媒体成为人们发布和获取信息的一个重要平台[1~5].然而,社会网络平台对信息缺乏有效的监管也导致网络谣言的泛滥[6].根据新浪微博 2019 年发布的《微博辟谣 2018 年度报告》显示,三分之一的谣言始发于社会网络[7].这些信息在未经处理的情况下可能被迅速地歪曲和放大,从而误导公众.谣言无节制地在网络上传播不仅影响社

收稿日期:2019-06-19;修回日期:2019-12-25;责任编辑:孙瑶
基金项目:四川省科技厅应用基础项目(No. 2018JY0193);四川省教育厅重点项目(No. 17ZA0238, No. 18ZA0305, No. 18ZA0301);国家自然科学基金(No. 61872254)

会和谐与稳定,甚至威胁国家或地区安全.

针对社会网络中的谣言泛滥问题,学术界进行大量的研究和探索. 现有社会网络谣言检测主要分为三类:人工检测方法,基于机器学习的检测方法与基于深度学习的检测方法[8~11]. 人工检测方法准确率高,但具有明显的滞后性,无法适应社会网络中海量数据. 机器学习方法将社会网络谣言问题看作有监督学习中的二分类问题,自动化程度高,有效地弥补了人工检测方法的不足,但基于机器学习的谣言检测方法依赖于人工提取与选择特征,耗费大量的人力、物力与时间,且得到的特征向量鲁棒性[12,13]也不够健壮. 深度学习方法则比机器学习方法中通过特征工程得到的特征数据对原数据具有更好、更本质的表征性,从而能实现更好的分类效果[14].

对于社会网络谣言检测问题的现状与发展趋势,学术界也进行了归纳与总结. 中国人民大学陈燕方等人[15]对社会网络谣言检测研究进展进行了总结与分析,他们以谣言检测的工作流程为主线,对社会网络谣言检测中的数据收集,数据标注,模型选择与模型训练等各个阶段进行了细致的介绍与分析. 由于当前社会网络谣言检测技术发展的速度较为迅猛,上述研究工作已经无法全面反映社会网络谣言检测技术的现状. 英国 WarWick 大学的 Zubiaga 等人[16]将社会网络谣言检测划分为谣言检测、谣言跟踪、谣言立场分类和谣言准确性分类四个阶段,并对四个阶段应用技术的现状与发展进行了总结与归纳. 但上述工作侧重于谣言检测技术各个部分具体的实现细节,孤立地对谣言检测涉及的技术进行描述,对其各部分之间的联系缺乏必要的系统性与整体性的介绍. 同时对近年来兴起的基于深度神经网络的谣言检测方法描述较少,难以为当前的研究提供全面的参考. Cao 等人[6]在陈燕方与 Zubiaga 等人工作的基础上补充了深度学习在社会网络谣言检测问题中的应用与发展,并对当前重要的社会网络谣言数据集进行介绍与分析. 但其受限于当时技术的发展,对于深度学习在社会网络谣言的分析与总结浅尝即止,不够全面与系统.

综上所述,现有的社会网络谣言检测问题综述侧重于谣言检测模型的研究,对于谣言检测流程中的数据收集问题、数据类型与谣言的关系问题缺乏必要的分析,同时由于技术发展的限制,上述工作对当前社会网络谣言检测的最新进展特别是深度学习在社会网络谣言检测问题中的应用缺乏系统全面的总结.

本文从谣言的定义出发,以社会网络谣言检测问题面临的挑战为主线,从现有技术的发展与谣言检测的流程两个维度出发,对社会网络谣言检测过程中面临的问题与解决方法进行分析与总结. 相较于现有的研究进展总结,本文的主要贡献有:

(1)对如何从半结构化且海量数据构成的社会网络平台收集有效数据的方法进行了分析,并对现有方法可能存在的问题进行了总结;

(2)在特征工程部分,对基于机器学习方法特征向量鲁棒性较差的问题进行了归纳与总结;

(3)对于当前最新的深度学习方法在社会网络谣言检测方法的研究现状进行了补充与完善,并在相同数据集下对当前主流算法进行实证评估.

本文的整体框架如图 1 所示.

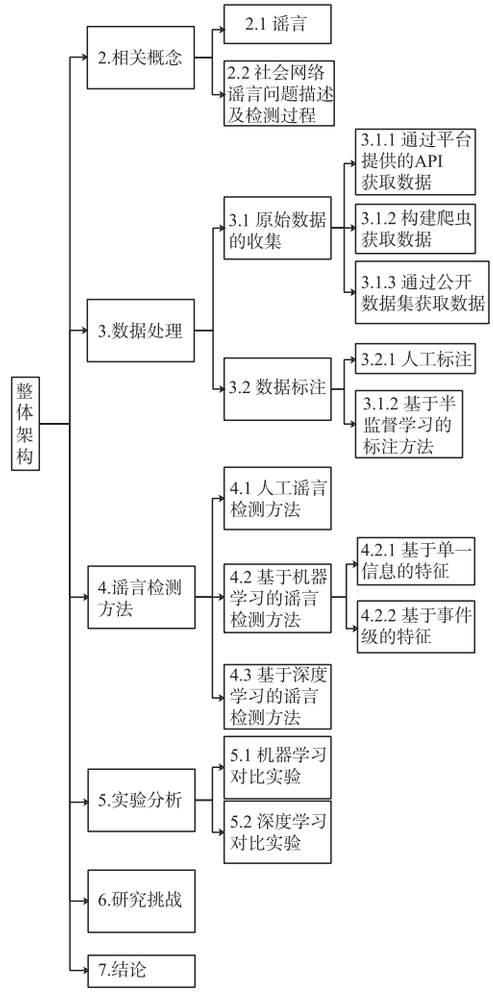


图1 本文整体架构

2 相关概念

2.1 谣言

不同文化对谣言的定义各不相同. 最新的《现代汉语词典》[17]把谣言定义为:(1)没有事实存在而捏造的话;(2)没有公认的传说;(3)民间流传的评议时政的歌谣,谚语. 在《牛津英语词典》[18]中,谣言被定义为:“目

前流传的关于不确定或可疑事实的故事或报告”. 本文将社会网络谣言定义为一种在社会网络上传播且未经验证,或已被官方证实为假,并在社会网络中流传的信息. 社会网络谣言的构成如图 2 所示,其特点是:发布门槛低、互动性强、散播速度快、散播方式和散播途径多样等.

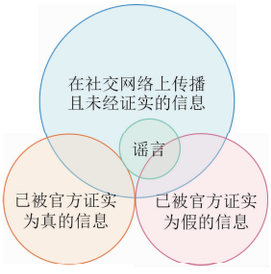


图2　社会网络谣言构成

2.2　社会网络谣言问题描述及检测过程

当前主流方法将社会网络谣言检测问题看作是有监督学习中的二分类问题[3],该问题的形式化定义如下:给定社会网络中每条推文的集合 $P=\{p_1,p_2,p_3,\cdots,p_i\}$ 和一个类别标签集合 $L=\{l_1,l_2\}$,其中,p_i 代表一条推文,l_1,l_2 分别代表谣言和非谣言这两个类别标签. 社会媒体谣言检测的任务是要学习一个分类模型 M,将推文 p_i 映射成一个类别标签 l_j,即 $M:p_i\rightarrow l_j$,模型的输入是一个包含有若干条微博的事件,输出是该事件对应的谣言或非谣言标签.

社会网络谣言检测过程通常包含:数据处理、特征选择与提取、模型训练与谣言检测四个阶段.

数据处理包括原始数据的收集与数据标注,数据收集的作用主要有两项:第一,用于构建模型训练的数据集;第二,对社会网络进行监控,获取待检测的社会网络信息. 数据标注则是根据问题及需求的不同对数据进行不同的标注.

特征选择与特征提取是从收集的原始数据中选择与构造出最能代表数据的特征向量集合. 对于机器学习方法而言,特征选择与提取的重要程度甚至超过了模型选择的重要性[19]. 因此现有基于机器学习方法的重要工作是以找到更有效的特征作为提升谣言检测准确率为主要思路. 基于深度学习的谣言检测具有很强的特征学习能力,其无需对特征进行人工提取即可得到比传统机器学习更高维、复杂、抽象的特征数据.

模型训练是指根据具体的问题场景从已有的分类模型中选择模型,并根据模型在训练数据集上的分类表现调整参数以找到一个最优模型的过程. 对于社会网络谣言问题,如何在充满噪音、且不均衡的海量数据信息中训练出准确率高的分类器是当前社会网络谣言检测问题面临的最大挑战.

谣言检测则是根据模型训练中得到的谣言分类器对社会网络中传播的信息进行信息真实性的鉴别.

3　数据处理

数据处理是社会网络谣言自动检测技术的基础,包括原始数据收集与数据标注两个阶段. 本节将对数据收集与数据标注的方法及其存在的问题进行总结与分析.

3.1　原始数据的收集

原始数据的收集是谣言检测工作的第一步. 社会网络中充斥着各种各样的信息使得获取庞大的数据集成为可能. 如表 1 所示,目前社会网络的收集方式主要有三种:通过社会网络平台提供的 API 获得,用户自己构建通用爬虫获得以及直接获取第三方提供的公开数据集.

3.1.1　通过平台提供的 API 获取数据

目前,几乎所有的社会网络平台都向用户提供了完善的 REST API 接口,方便用户从社会网络平台中获取到用户发表的微博、好友关系等社交信息[20]. Qazvinian 等人[21]根据 Snopes[22]网站上公布的谣言信息提取关键字,首先使用 Twitter search API 从 Twitter 上定时获取 Twitter 中发布的相关推文. 此后,与 Twitter 相关的研究[3,10,11,23]基本上都是参考 Qazvinian 等人的方法收集 Twitter 上数据. 2013 年,Sun 等人[24]根据微博辟谣中公布的造谣用户的账户信息,利用新浪微博 API 收集造谣用户的信息,用户发布的微博信息,用户好友信息及用户好友发布的微博信息用于新浪微博谣言检测问题的研究. 在文献[25]中,Wu 等人通过新浪微博 API 收集了从 2012 年 5 月 28 日至 2014 年 4 月 1 日期间新浪社区管理中心发布的 11466 条虚假谣言用于微博谣言的研究. Cai 等人[26]收集了新浪微博社区管理中心和微博公众号@ WeiboPiyao 在 2015 年 4 月 10 日至 2016 年 5 月 3 日期间发布的 9834 条谣言.

基于平台提供的 API 获取数据的方法优点是简单快捷,但其缺点也十分突出:

(1)受限于社会网络平台的保护策略,通过平台 API 获取的数据在数据爬取速度及爬取数量上都受到严格控制,无法满足用户研究的需求.

(2)收集的数据具有较强的先验性[15],利用 API 收集数据存在一个先决条件:需要用户提供搜索关键字,根据搜索关键字收集微博中对应用户或是对应事件的信息. 所以基于 API 的数据收集方法在社会网络谣言问题中只适用于收集模型训练中的数据集,而无法有效用于实时监控数据的收集.

3.1.2　构建爬虫获取数据

由于社会网络 API 的诸多限制,研究者们开始使用

模拟登录技术构建网络爬虫实现社会网络数据的收集. 2012 年,Yang 等人[27]对新浪微博谣言分析与检测进行了首次研究,收集了 2010 年 3 月 1 日至 2012 年 2 月 2 日的数据. Cao 等人[28]通过构建爬虫不仅收集了微博中的用户信息,文本与传播结构数据,同时也收集了微博中的多媒体数据. Guo 等人[3]编写爬虫的同时收集了 Twitter 与微博平台的数据. Dong 等人[29]选取 Facebook 作为目标网络,基于 Facebook 中的某一个用户节点,以其好友信息为拓展,对 Facebook 平台的数据进行爬取.

相较于基于平台 API 的方法,基于爬虫的方法能够快速获取满足研究者所需的数据,但是该方法在数据收集过程中也面临诸多挑战:

(1)基于爬虫的数据收集方法可能面临法律风险. 2016 年 11 月 7 日发布的《中华人民共和国网络安全法》[30]第四十一条明确规定:"网络运营者收集、使用个人信息,应当遵循合法、正当、必要的原则,公开收集、使用规则,明示收集、使用信息的目的、方式和范围,并经被收集者同意".

(2)技术复杂度高,为了从社会网络平台收集数据,用户需要使用到模拟登录,动态网页获取等相关技术,同时构建的爬虫程序需随时根据社会网络平台的安全策略的变化而进行对应的修改.

3.1.3 通过公开数据集获取数据

为了提高社会网络谣言问题的研究效率,相关领域的机构与研究者们提供了社会网络公开数据集. 比较有代表性的数据集有:北京理工大学大数据搜索与挖掘实验室张华平博士团队基于新浪微博提供的 NLPIR-Parser 数据集[31],清华大学自然语言处理实验室基于新浪微博推出的 THUCTC 数据集[32],Queensland 大学 Chen 等人[11]基于 Twitter 提供的数据集,以及 Arizona 州立大学的 Wu 等人[10]提供的基于 Twitter 的实验数据集.

公开数据集在一定程度将研究者从琐碎繁重的数据收集工作中解放出来,让研究者集中精力在谣言检测方法的研究. 但是公开数据集存在的弊端也显而易见:

(1)公开数据集中的数据也是通过 API 或是爬虫获取得到的,所以 API 或是爬虫获取数据的问题在公开数据集中依然存在.

(2)收集的数据可能无法满足用户的实际需求,公开数据是数据提供者根据自己的知识背景与经验收集的数据,在收集时无法做到面面俱到,从而满足所有用户的需求.

表 1 不同数据获取方式的对比

获取方式	来源	语言	类型	优点	缺点
API	Twitter search API [10,21]	英文	文本	方便且获取相对简单	API 授权机制要求更高,部分 API 费用高
	新浪微博 API [25~27]	中文			
爬虫	Snopes & 新浪社区管理中心[3]	英文 & 中文	图片 & 视频	可根据需求爬取各种信息	爬取数据时间成本高,限制较多
	Facebook [29]	英文	文本		
爬虫 &API	FIBODATA & Facebook [21,33]	英文 & 中文	文本	综合性强,获取数据多样性明显	花费人力物力成本相对较高
公开数据集	Cit-HepTh & College-Msg & Email & Wiki-Vote [34]	英文	文本	容易获取	数据即时性不强
	NLPIR-Parser[31]				
	THUCTC [32]				
公开数据集 & 爬虫	VMU 2015 dataset & Google & Baidu [23]	英文 & 中文	图片 & 视频	保证实验数据的权威性和多样性	相对有较大的工作量

3.2 数据标注

当前主流方法将谣言检测问题看成是有监督学习的二分类问题,为了训练出用于谣言甄别的分类器,需要对用于模型训练的数据集进行数据标注,当前面向社会网络谣言数据集的标注方法主要有人工标注和基于半监督学习标注两种方法.

3.2.1 人工标注

人工标注方法是指专人对收集的初始数据的类别(谣言或正常信息)进行标记. 为了避免认知的偏差,现有的人工标注方法通常会聘请两人及以上标注者对数据内容同时进行标注[1,17,19],并从初始数据集中选择标注结果相同的项作为最终训练数据集的候选项. 人工标注方法简单直接,但该方法耗费了大量的人力、物力与时间,而且标注的质量依赖于标注者的知识背景与经验.

3.2.2 基于半监督学习的标注方法

针对人工标注方法的问题,Wu 等人[10]首次在社会

网络谣言检测问题中引入基于半监督学习的自动标注方法,在人工标注少量数据的条件下,引入了一种叫做 CERT(Crosstopic Emerging Rumor deTection)的框架,该框架联合聚类数据、选择特征和训练分类器实现数据的分类. 基于半监督学习的自动标注方法简单且易实现,在一定程度上缓解了人工标注方法存在的问题,但该方法的先决条件太强,需要研究者能准确地估计数据分布信息. 但在实际工作中,研究者很难事先对数据做出准确的模型估计. 因此社会网络谣言检测问题中,人工标注方法依然占主导地位.

4　谣言检测方法

现有的谣言检测主要分为人工谣言检测方法、基于机器学习的谣言检测方法和基于深度学习的谣言检测方法. 本节将对这三种方法进行介绍与对比.

4.1　人工谣言检测方法

人工谣言检测方法是当前社会网络平台主流的谣言检测方式,平台将社会网络中的可疑信息交给经验丰富的编辑或是行业专家,利用编辑和专家的领域知识和经验对信息的真实性进行甄别. 当前的主流社会网络平台,如 Twitter、Facebook 与新浪微博[33~36],在其平台上都是采用人工的谣言检测方法.

Twitter 采用众包方法对平台上的信息的真实性进行鉴别. Twitter 设计了一种信息真实性判别算法,该算法能根据 Twitter 上用户对信息的评价计算平台上每一条信息的真实度. Facebook 采用人工标注与权威媒体证实相结合的方法对 Facebook 上的传播信息的真实性进行判别. Facebook 用户一旦在 Facebook 上发现可疑信息,可通过平台的接口提交其发现的可疑信息,被举报的信息其后通过权威媒体(比如 FactCheck. org[37] 或 Snopes. com[38])提供的 API 提交给该媒体的编辑,由权威媒体的编辑与专家对消息的真实性进行甄别. 新浪微博平台提供了两种不实信息检测方法,第一种是“微博辟谣”[39],“微博辟谣”是微博平台上的一个公众号,该公众号定期发布平台上发现的不实信息,凡是关注了该公众号的微博用户第一时间可以了解到微博平台中不实信息的传播情况. 第二种方法是“举报处理大厅”[40],该方法同样采用众包方法,微博用户通过“举报处理大厅”提供的接口向平台举报可疑的信息,微博平台的专家对举报的信息进行鉴别,并在平台上公布鉴别结果.

图3 展示三大社交网络平台使用的谣言检测方法.

人工谣言检测方法具有准确率高的特点,但是这种方法存在以下问题:

(1)人工谣言检测需要检测者对用户或平台举报的信息进行逐条判断,在耗费大量人力的同时也会造成信息判断的滞后性.

(a) Twitter　　(b) Facebook

(c) 举报处理大厅

图3　三大主流社会网络人工谣言检测方法示意图

(2)谣言检测的质量依赖于检测者的知识背景与经验,对个人的知识与经验要求极高,而且有可能因为个人因素而造成误判.

(3)社会网络中每天产生数以亿计的数据,单靠人力无法对所有数据进行处理,而经过筛选来判断信息有可能遗漏重要的谣言信息.

因此,设计与开发出能自动检测谣言的方法已经成为解决社会网络谣言问题的关键.

4.2　基于机器学习的谣言检测方法

早期对谣言的自动检测主要集中在利用机器学习技术来检测谣言,该类方法主要包含三个流程:(1)从训练数据集中选择并提取能够有效表征数据的特征;(2)利用选择与提取的特征在训练数据集上训练分类模型;(3)使用训练好的模型对训练数据集外的数据进行预测,经过不断的评估与优化,判断数据是否是谣言. 其谣言检测流程如图 4 所示.

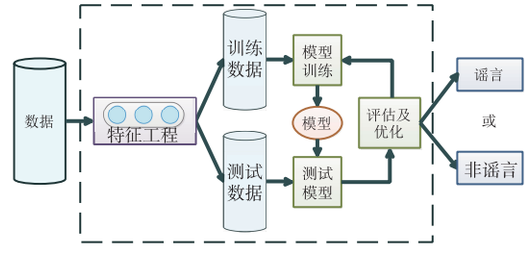


图4　基于机器学习谣言检测流程图

对于基于机器学习的谣言检测方法而言,如何选择与提取出显著的特征来表征数据对谣言检测的效果至关重要. 早在 1999 年,Waikato 大学 Mark A Hall 就在其博士论文《Correlation-based Feature Selection for Machine Learning》[19] 中指出:“选择与提取有效的特

征对于分类算法非常重要,其重要性在某种程度上甚至超过了分类模型的选择". 因此基于机器学习的谣言检测方法在某种程度上可以说是一种基于特征工程的方法. 现有的用于检测社会网络谣言的特征提取方式主要包括:(1)基于单一信息的特征提取方式,通过提取单条数据的特征来处理数据;(2)基于事件级特征提取方式,通过挖掘数据之间层次性关系来提取数据之间的潜在联系. 本节将分析基于单一信息的特征与基于事件级特征两种特征提取方式并描述其谣言检测的过程.

4.2.1 基于单一信息的特征

基于单一信息的特征提取方式是早期谣言检测中最常使用的方法,根据特征提取复杂度的不同,可分为显式特征(explicit feature)与隐式特征(implicit feature).

(1)显式特征

显式特征指的是通过直接选取即可获得的特征,包括消息文本的长度、用户的个人信息、粉丝数以及转发数等,表2为各种显式特征及其特征描述. 早在2011年,Castillo等人[8]分析与热门话题相关的微博帖子,人工地通过发布的文本内容以及引用的外部源对帖子进行可信度评估. 提出了基于消息的特征(Message-Based Features)和基于用户的特征(User-Based Features),并采用支持向量机(Support Vector Machine,SVM)、决策树(Decision Trees,DT)、决策规则(Decision Rules,DR)以及贝叶斯网络(Bayes Networks,BN)四种算法对其进行验证,其中,J48决策树算法达到了最高的89%的分类精度. 同年,Qazvinian等人[21]提出了基于内容的特征、基于网络的特征(Network-Based Features)以及基于Twitter特定模因的特征(Twitter Specific Memes Features),用来分析推文中词汇模式、词性模式以及特定于Twitter的模因中提取出的标签和URL,并采用朴素贝叶斯(Naïve Bayesian,NB)在提取基于内容的特征中得到94.1%的准确率. Yang等人[27]在文献[8,21]的基础上,提出基于客户的特征(Client-Based Features)以及基于位置的特征(Location-Based Features),以此分析用户使用的客户端程序以及所发微博事件发生的实际地点对谣言判断的作用,并在新浪微博数据集下,利用SVM分类器将准确率平均提高了5.5%. 但这些显式特征都是预定义的,且在现实生活中,一条信息是否为谣言的最终裁决者是人. 对于机器学习方法而言,单纯地通过文本,用户以及传播特征[8,21,27,41,42]等进行信息真实性的鉴别是一件非常困难的事情. 因此,研究者们引入一种动态的、潜在的隐式特征,用以提取数据之间的隐含关系.

表2 显式特征及其描述

特征	特征相关描述
基于消息/内容的特征[8,21,27]	消息长度 包含感叹号,问号,积极/消极情绪词 微笑,皱眉等表情符号 包含第一,第二,第三这类代词 大写字母的数量 流行度排名 包含标签的数量 是否是转发
基于主题的特征[8,27]	原始消息的主题 同一主题的微博数量
基于Twitter特定模因的特征[21]	Hashtag的数量 URL的数量
基于用户特征[8,27]	用户的注册年龄 用户的粉丝数量 注册用户数量(Twitter上的好友数量) 用户在过去发布的推文数量
基于用户发送特征[46]	微博中包含的URL的数量 发送文本数量 用户是个人真实姓名还是组织名称
基于用户响应特征[46]	用户在微博中评论其他微博的数量 转发微博的数量
基于位置的特征[27]	用户发送微博的地理位置
用户类型特征[46]	用户是否被验证
基于网络的特征[21,27]	使用RT @user推断转发的消息

(2)隐式特征

隐式特征指的是无法直接获取,需通过关联分析或数值计算得到的一种潜在特征,如平均情感特征、用户可信度以及质疑率等,如表3所示. Guo[3]等人提取了基于账户的特征(Account-Based Features),包含从用户简介和用户行为中提取用户可信度,可靠性和名誉等隐含信息. Wu等人[25]提出主题类型特征(Topic Type Feature)、用户类型的特征(User Type Feature)、平均情感特征(Avg Sentiment Feature)以及转发时间特征(Repost Time Feature),通过狄利克雷分布(Latent Dirichlet Allocation,LDA)[43,44]提取消息的主题,该主题在消息中的概率分布可通过式(1)求得.

$$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D}) = \prod_{i=1}^{K}p(\beta_i)\prod_{d=1}^{D}p(\theta_d)\left(\prod_{n=1}^{N}p(z_{d,n}\mid\theta_d)p(w_{d,n}\mid\beta_{1:K},z_{d,n})\right) \tag{1}$$

其中,$\beta_{1:K}$表示1到K的所有主题,β_i表示第i个主题词的分布,θ_d表示第d个消息中主题所占的比例,$z_{d,n}$表示第d个消息中第n个词的主题,$w_{d,n}$表示第d个消息中

第 n 个词.

除得到推文的主题类型之外,他们还考虑发帖者是否是已被验证的用户,并通过基于词汇的平均情绪得分来判断情绪词与谣言之间的关联,并考虑原始消息和转发消息之间的时间间隔因素. 通过基于随机游走图核(Random Walk Graph Kernel)的 SVM 检测算法在随机选取的微博数据上得到 91.3% 的准确率. 在社会网络传播的信息其实隐藏着用户的某种行为,Mendoza 等人[45]在研究智利大地震时 Twitter 中的推文变化情况发现:相较于真实信息,谣言更容易引起受众的质疑. 由此 Liang 等人[46]提出了一种基于用户行为特征的谣言检测方法,他们通过收集的微博数据发现:造谣者相较于正常信息发布者,为了逃避可能承担的惩罚以及为了快速传播谣言信息,其用户行为与普通用户存在着较大的行为差异,用户在阅读正常信息与阅读谣言信息时也存在着较大的行为差异. 在此基础上,Liang 等人[47]还提出了包括质疑率,单位时间发文数在内共计 10 条特征用于社会网络谣言的实验. 其中,质疑率表示用户所质疑的评论在所有评论中所占的比例. 实验结果表明,该方法相较于传统的基于文本、用户与传播结构特征方法,查准率与查全率的提高均超过了 15%.

表 3　隐式特征及其描述

特征	特征相关描述
平均情感特征[25]	原始信息和所有转发推文中积极/消极情绪所占的比例
基于主题的统计特征[25]	包含 URL 和 hashtags 的推文比例 主题词汇模式 主题词性模式
基于账户的特征[3]	用户的可信度、可靠度、名誉度
基于发送文本的特征[3]	事件中发送文本的统计特征
转发时间特征[25]	从原始消息到转发消息之间的时间差
基于传播结构的特征[8,27]	传播的初始微博 传播树最大子树 传播树最大/平均深度

基于单一信息的特征提取方式虽简单,但存在以下不足:

(1)依赖人工进行特征的选择,耗费人力物力的同时,得到特征向量的鲁棒性较差[48].

(2)选取的特征主要集中在从原始消息和转发消息中提取大量的词汇和语义特征,并从标记的数据中学习模型[8,21],难以全面系统地概括谣言的特点.

(3)加入用户特征虽引入了消息之间的关系且构造机器学习的特征向量也相对方便,但忽略了消息传输的内部图形结构以及该结构下用户之间的差异[25].同时,仅依赖于社交媒体平台提供的用户信息,无法真正有效地对不同平台用户发布的信息进行检测.

4.2.2　基于事件级的特征

仅仅提取单一信息的特征往往忽略了谣言之间的联系,而基于事件级特征可通过其层次性结构反映出谣言之间的潜在关联. 本节将基于事件级的特征定义为用户、消息、子事件、事件之间的层次关系特征. 如图 5 所示.

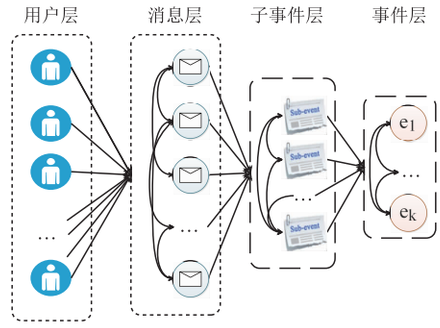


图5　基于事件级特征层次结构图

该层次结构是由用户层、消息层、子事件层以及事件层组成的多类型网络结构. 其中,事件层为 $E=\{e_1,e_2,e_3,\cdots,e_k\}$,指在特定时间、特定地点包含一定关键词的事件集合;子事件层为 $S=\{s_{k,1},s_{k,2},s_{k,3},\cdots,s_{k,n}\}$,指每个事件中子主题的集合;消息层为 $M=\{m_{n,1},m_{n,2},m_{n,3},\cdots,m_{n,i}\}$,指用户发出的原贴以及转发贴的集合. 层内链接反映同一层级内实体之间的关系,而层间链接则反映了不同层级之间的关系. 2012 年,Gupta 等人[49]提出了一种基于事件图优化(Event Graph-based Optimization)的可信度分析方法. 根据事件重要程度的不同赋予不同的分数,同时,通过对新事件层次化关系之间使用正则化更新事件可信度得分来增强基本的可信度分析. 在数百万条推文的数据集上,参考 Castillo 等人[8]用四种机器学习算法进行实验,得到高于文献[8]方法 14% 的准确率,说明基于事件的层次化结构优于基本的基于单条推文的可信度分析方法. 此后,Sun 等人[24]引入一种新的基于多媒体的特征(Multimedia-Based Feature),加入了图片的特征,并根据该项特征来判断微博信息中包含的图片是否是过去图片. 采用朴素贝叶斯、贝叶斯网络、神经网络以及决策树对新特征进行验证,发现该特征在贝叶斯网络中可获得 85% 的准确率. 由于不同主题事件中不同层级或层内消息在谣言检测中的潜在联系也是不同的,因此,Jin 等人[50]首次引入子事件层,提出了一种分级传播模型(Hierarchical Propagation Model),用以对从消息级到事件级新闻可信度进行评估. 该模型由事件、子事件和消息组成三层可信度网络,并利用这些实体之间的语义和社会

关系建立联系,同时将该网络的可信度传播过程表示为图的优化问题,用以求出迭代算法的全局最优解. 在两个数据集该模型的准确率提高了6%以上,F-score[51]提高了16%以上.

结合谣言的层次结构虽然可弥补基于单条推文特征的一些不足,但其本质还是通过人工选择并提取特征. 因此,仍存在机器学习中特征提取的通病:

(1)难以获得高维、复杂、抽象的特征数据.

(2)试图用一套通用的特征集合表征社会网络不同平台不同语言中的全部信息,训练出来的谣言分类器容易陷入"过拟合"状态[52],模型准确度不高.

(3)所有的实验都在研究者自己选择的数据集上进行实验,并不能有效地体现出新提出的特征在不同平台不同数据集下对谣言检测的作用.

4.3 基于深度学习的谣言检测方法

由于传统机器学习的谣言检测方法依赖特征工程需要耗费大量的人力、物力与时间来选择合适的特征向量,因此,研究者们尝试在社会谣言问题检测中引入深度学习的方法. 深度学习具有很强的特征学习能力,其模型学习的特征比传统机器学习算法中通过特征工程得到的特征数据对原数据具有更好的,更本质的代表性,从而能实现更好的分类效果[14]. 本节以基于深度学习的谣言检测技术的发展为线索,深入分析并总结了现有的基于深度学习的谣言检测方法.

微博中的信息是一种与时间密切相关的时序数据,而循环神经网络(Recurrent Neural Network,RNN)[53,54]在时间序列和句子等变长序列信息建模方面显示出了强大的功能. 2016年,Ma等人[55]首次将循环神经网络引入到谣言检测中,通过对文本序列数据进行时间维度上的建模分析得到谣言上下文信息随时间变化的隐式特征. 加入长短期记忆(Long-Short-Term Memory,LSTM)[56,57]以及门控循环单元(Gated Recurrent Unit,GRU)[58]等额外的隐藏层,解决了在长序列训练过程中,随着RNN层数的加深而造成的梯度消失与梯度爆炸问题[59,60],从而提高谣言检测的准确度. 在微博数据集上,加入双层GRU的循环神经网络准确率为88.1%,在Twitter数据集上,其准确率高达91.0%,都超过了基础tanh-RNN与加入一层LSTM/GRU的谣言检测准确率. 图6为基于循环神经网络的谣言检测的流程图. 首先,针对每个事件收集相关帖子,对输入的事件文本数据得到tf-idf值矩阵,再将高维的词袋模型向量通过词嵌入的方式转成低维空间的向量表示,得到输入值. 然后,将该值输入到RNN模型中,通过循环神经网络捕获文本序列的相关语义特征,由于基础的隐藏层没有门控单元,在t时刻向前反向传播的过程中,存在梯度消失(大部分情况下)或者梯度爆炸的情况,使得该结构难以捕捉长距离依赖,为缓解基础模型带来的缺陷,在隐藏层加入门控单元LSTM/GRU,通过门(gate)机制控制隐藏层中的信息流动,保留了文本间的语义信息,以提高谣言检测的准确度. 最后,通过sigmoid激活函数输出分类标签,预测是否是谣言.

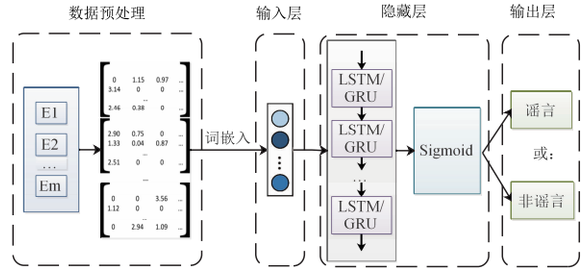


图6 基于循环神经网络的谣言检测流程,E代表事件

然而,在谣言爆发的初期,无法获取足够的标记数据用来训练模型,因此,为能够尽早地检测出社会网络中的谣言,Chen等人[52]提出结合循环神经网络和变分自编码器(Variational Auto-Encoder)[61]的无监督学习模型来学习社会网络用户的网络行为,由于正常数据与异常数据在降维过程中存在着显著的差异[62],因此利用模型得到输出值和输入的目标值之间的误差与指定阈值进行比较,判断其是否是谣言. 其中,RNN与自编码器(Auto-Encoder,AE)的结合模型如图7所示.

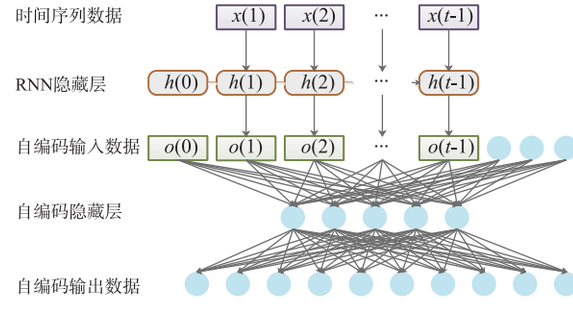


图7 RNN与AE结合的模型

该模型主要分成两个模块进行层次训练,分别为RNN模块和AE模块. 首先将收集到的不同时间节点的微博数据进行清洗后,建立特征工程,通过微博内容提取是否有图片,是否有转发,是否是积极态度等15个特征,传入RNN模块,并在时间维度上进行训练;然后将该模块的输出结合发博时间,发博来源等其余特征送入AE模块,通过AE实现无监督的异常检测,通过一系列的矩阵映射将输出重构成与输入形状相同的结构;最后,使用欧几里得范式计算AE模块输入的目标值和输出值之间

的误差,并与设定的阈值比较,从而判断该推文是否是谣言. 该模型实现了单隐藏层和多隐藏层结构,两层模型的准确率分别为 92.49% 和 89.16%. 但该模型只在新浪微博的谣言数据下进行实验,并不能很好地验证出其在不同平台数据下的适应性. 因此,Wen 等人[23]设计了一个基于神经网络的模型,该模型采用了跨语言、跨平台的有限元分析方法,利用不同平台和语言之间的信息相似性和一致性来验证谣言. Ajao 等人[63]利用卷积神经网络(Convolutional Neural Networks,CNN)和长短期循环神经网络模型(Long-Short Term Recurrent Neural Network Models)来检测并分类 Twitter 上发布的虚假新闻. 该方法无需任何人工提取外部特征的步骤即可直观地识别与谣言相关的特征.

传统的基于深度学习的谣言检测方法摆脱了人工构建特征工程的方式. 然而,天然的端到端结构难以把握谣言信息中的关键成分,模型训练缺乏可控性,训练时间长且模型复杂. 因而引入注意力机制(Attention Mechanism)[30,64]进行谣言检测. 注意力机制最早提出于视觉图像[65]领域,该方法借鉴了人类的注意力思维方式,模仿人类对图片不同地方的观察侧重点,用以对图像不同位置施加不同的权重,从而决定更重要的部分,并提高该部分的权重,降低噪声部分的权重. 2014 年,Bahdanau 等人[66]首次将注意力机制引入自然语言处理领域,该工作首先通过对 Encoder 部分的输入和隐藏状态值经过循环神经网络进行编码,从而输出中间向量,再由 Decoder 部分将中间向量借助另一个循环神经网络解码成输出向量.

基于注意力机制在谣言检测领域的应用,Chen 等人[11]提出一种基于注意力机制的循环神经网络模型 CallAtRumors(Call Attention to Rumors),加入注意力机制从重复、不断变化的推文中提取出隐式与显式的谣言特征,用于对社会网络信息序列中选择关注度高的信息进行检测,在模型训练中,采用交叉熵损失函数和双重随机正则化[67]相结合的方法,对输入字矩阵的每个元素进行校正,其损失函数如式(2)所示:

$$L = -\sum_{t=1}^{\tau}\sum_{i=1}^{C} y_{t,i}\log y'_{t,i} + \lambda\sum_{i=1}^{K}\left(1-\sum_{t=1}^{\tau} a_{t,i}\right)^2 + \gamma\varphi^2 \quad (2)$$

其中,y_i表示独热标签向量(one hot label vector),y'_i表示在 t 时刻的二分类概率向量,τ 表示总时间,C 表示输出类的数目,其数值为 2(表示谣言或非谣言),λ 表示注意力分配系数,γ 表示权值系数,φ 代表所有模型参数.

该模型在 Twitter 与新浪微博上分别取得 88.63% 和 87.10% 准确率. Jin 等人[1]在此基础上加入图片这一特征,使用循环神经网络来学习文本和社会背景(social context)相结合的表示;使用卷积神经网络训练提取图像的视觉特征;使用注意力机制对视觉特征和共同的文本/社会背景特征分配不同权重. 融合了文本、图像和社会背景特征对 Twitter 和新浪微博数据集进行谣言检测,但其在两个数据集上的准确率分别为 78.8% 和 68.2%,难以保证谣言检测的效果. 因此,Guo 等人[3]提出了一种结合社会信息(social information)的层次神经网络(HSA-BLSTM)方法用于谣言检测. 首先建立了表示学习的层次双向长短时记忆模型(Hierarchical Bi-directional Long Short-term Memory Model),然后通过注意力机制将社会背景整合到网络中,最后在新浪微博和 Twitter 中进行实验,分别取得 94.3% 和 84.4% 的准确率. 与 Guo 等人[3]类似,Liao 等人[68]通过采用两层带有注意力机制的双向 GRU 网络从微博内容和时间层面分别获取微博序列的隐藏层表示和时间段序列的隐藏层表示,从而在事件的特征表示中融入了时间段内各微博间的时序信息. 此外,还针对各个时间段提取了局部用户特征及文本潜在特征,并将这些特征融入到时间段中,进一步捕获这些特征随时间变化的隐藏层状态值,最终得到 96.8% 的谣言检测准确率. 但该方法依赖人工对事件进行时间段划分,在花费人力及时间的基础上还可能带来信息的丢失. 为通过区别原贴和转发贴来检测谣言,Xu 等人[69]考虑原帖内容、转发帖的扩散情况以及用户信息三方面,提出一个融合神经谣言检测(Merged Neural Rumor Detection,MNRD)模型,通过基于内容的注意力机制的原贴编码和基于扩散的注意力机制的转发编码分别学习从原贴和转发中提取高层次的特征表示,通过用户特征编码器对用户信息进行编码,以获取用户可靠性和社会影响力,结合这些特征对谣言进行检测. 在新浪微博数据集上取得 94.4% 的准确率.

基于注意力机制的循环神经网络模型不仅具有很强的特征学习能力,同时能捕获谣言中的重要语义成分,但其仍存在以下不足:

(1)对数据的需求量大,当样本数据较少时,训练出来的分类器仍存在分类偏倚[70]问题.

(2)模型训练周期更长,训练出的模型可解释性差.

(3)需要 GPU 来高效优化矩阵运算,对 GPU 的要求较高.

5　实验分析

为对现有主流谣言检测算法进行客观评价,对比实验所用的算法选择如表 4 所示. 本文实验数据集选择公开数据集 Chinese_Rumor_Dataset[71],实验环境如表 5 所示.

表 4 算法选择及基本原理

方法	算法	基本原理
机器学习	逻辑回归(Logistic Regression,LR)[36]	根据现有数据对分类边界线建立回归公式,以此进行分类.
	随机森林(Random Forest,RF)[20,65]	利用随机的方式将许多决策树组合成一个森林,并通过投票方式决定测试样本的最终类别.
	支持向量机(Support Vector Machine,SVM)[8,25]	寻找一个超平面对给定的包含正例和反例的样本集合进行分割.
	决策树(Decision Trees,DT)[8,24]	根据数据的属性采用树状结构建立决策模型.
	朴素贝叶斯(NaÏve Bayesian,NB)[21,24]	基于贝叶斯定理和特征条件独立性假设的多分类.
深度学习	循环神经网络(Recurrent Neural Network,RNN)[52,55]	以序列数据为输入,在序列的演进方向进行递归,且所有节点(循环单元)按链式连接.
	长短期记忆(Long-Short-Term Memory,LSTM)[3,55]	通过门控机制使基本 RNN 不仅能记忆过去的信息,同时还能选择性遗忘一些不重要的信息,从而对长期语境等关系进行建模,为解决基本 RNN 存在的长期依赖问题而专门设计,是 RNN 的一种变体.
	门控循环单元(Gated Recurrent Unit,GRU)[55,68]	基于 LSTM 算法思想在保留长期序列信息下减少梯度消失问题,通过更少的门控单元及参数,实现与 LSTM 相差无几的训练效果.
	长短期记忆 + 注意力机制(LSTM + Attention)[3,11,66]	在 LSTM 的模型上加入 Attention 层,通过 Attention 机制将注意力集中在对当前任务更重要的向量上,即给不同向量分配不同权值.
	门控循环单元 + 注意力机制(GRU + Attention)[11,68]	在 GRU 的模型上加入 Attention 层,减少门控单元及参数的基础上实现更好的权值分配

表 5 实验环境

CPU	IntelI CoreI i7-9700K CPU @ 3.60GHz
GPU	NVIDIA GeForce GTX 1080Ti
内存	32GB

表 6 特征及其描述

类别	特征	特征描述
文本特征	text_length	文本的长度
	mark_num	文中!,? 等标签的数量
用户特征	friends	好友数
	followers	粉丝数
	comments	评论数
	likes	点赞数
传播特征	reports	转发数
	mention_num	文中@ 符号的数量
模因特征	url_num	文中 URL 的数量
	hashtag_num	文中 Hashtag 的数量

5.1 机器学习对比实验

早期基于机器学习的谣言检测方法侧重于特征的提取. 如表 6 所示,本文参考[8,21,25,46],选择微博信息中的“文本特征”、“用户特征”、“传播特征”、“模因特征”四大类特征,用于表征模型中的微博数据. 图 8 显示通过皮尔逊相关系数(Pearson Correlation Coefficient,

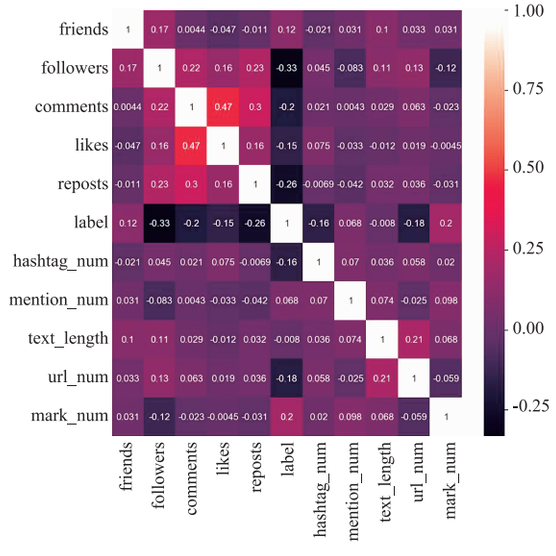


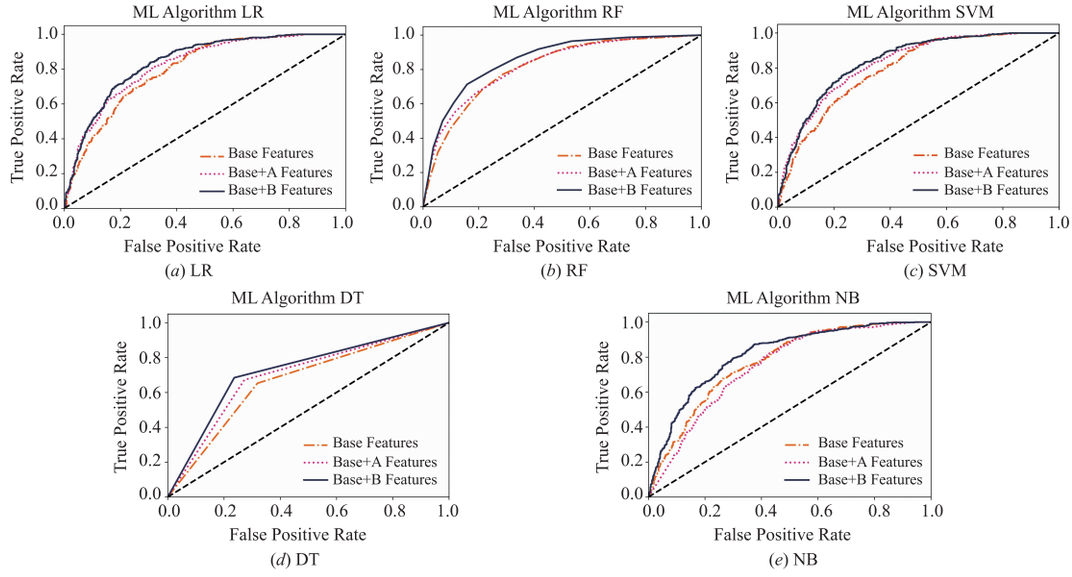
图8 多特征热力图

PCCs)[72]度量数据集中各个特征与谣言之间的相关程度. 为了验证模型、特征与谣言检测的关系,本文将特征分为三组:Base、Base + A、Base + B. 其中,Base 为根据相关系数选择的基准特征,Base + A 在基准特征中引入了参考文献[27]中提出的 location(微博时间发生地点)与 source(微博使用终端类型)两个特征. Base + B 借鉴参考文献[47]加入了质疑率、单位时间关注好友数目两个特征. 三组特征在五种主流机器学习算法下

的 ROC 曲线如图 9 所示.

图 9 表明,在相同算法下,加入新特征后谣言检测准确率均有所提升;除决策树外,在相同特征下,不同算法的 AUC 值[73]相差不足 0.03. 其中,造成决策树算法准确率普遍下降的原因在于其训练模型的同时对基线特征进行再次筛选,导致用于表征微博特征向量的特征数目减少,因此效果略逊于其他算法.

上述实验结果表明,对基于推理的机器学习方法而言,如何选择显著的特征向量表征数据是提高谣言分类器的关键,而模型选择对于谣言识别性能的影响则相对较小.



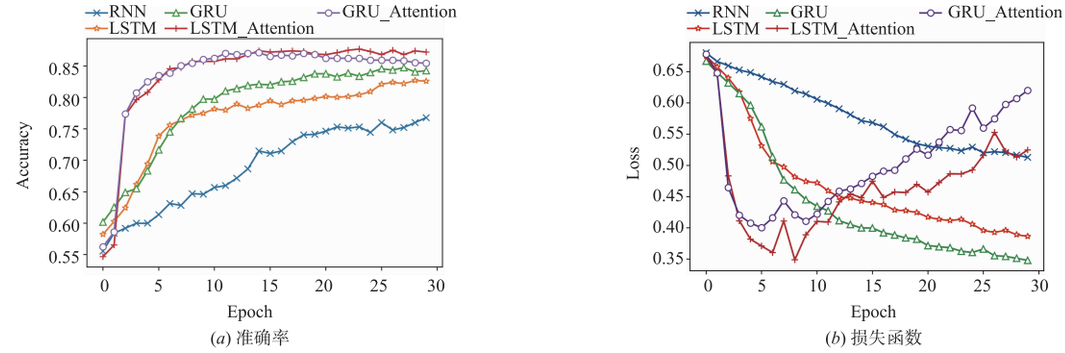
(a) LR　(b) RF　(c) SVM　(d) DT　(e) NB

图9　三组特征在不同机器学习算法下的ROC曲线

5.2　深度学习对比实验

基于深度学习的谣言检测方法无需进行特征工程,算法具有极强的特征学习能力. 为更加客观公正地对五种基于深度学习的谣言检测算法进行评估,本文选择当前研究中主流的五种深度学习算法,并全部选择单层隐藏层,以防止由于模型过于复杂而造成的过拟合现象. 五种算法均通过反复调参,经过 30 ×(3385 ×0.8/32)次迭代(Iteration),得出的实验结果如图 10 所示.

图 10 清晰地显示了五种深度学习算法的准确率(Accuracy)及其损失函数值(Loss)随轮次(Epoch)变化的情况. 相比之下,基础 RNN 只有 78% 的准确率,且其损失函数值高达 0.50;在隐藏层中加入 LSTM 及 GRU 后的准确率分别为 79% 和 81%,两者的损失函数值则分别比传统 RNN 降低了 0.06 和 0.13;而 LSTM + Attention 和 GRU + Attention 两种算法的准确率均达到 85% 以上,但由于加入 Attention 机制也使得模型更加复杂,从而造成过拟合现象. 因此其损失函数值从第 5 个轮次开始呈现上升趋势,并一度超过另外三种算法的损失函数值. 造成以上结果的原因主要在于数据样本数量不足且数据分布不均匀,难以支撑深度学习算法中所需的大量数据用以训练及测试模型的要求. 同时,模型复杂度的增加也导致模型极其容易造成过拟合现象.



(a) 准确率　(b) 损失函数

图10　五种深度学习算法的学习曲线

综上,基于结构的深度学习方法相较基于推理的机器学习方法在性能上有了较大的提高. 原因在于微博信息从本质上是一种时序数据,而深度学习算法可以提取丰富的潜在特征,并因此进一步捕获这些特征随时间变化的隐藏层状态值,提升谣言检测的准确率. 而传统的机器学习算法则缺乏对时序数据进行处理的功能. 因此,即使在小样本情况下,基于深度学习的谣言检测算法整体表现仍优于基于机器学习的谣言检测算法,这也是近年来谣言检测领域的研究者不断在各种深度学习算法中探索的重要原因之一.

6 研究挑战

尽管社会网络谣言检测技术已经取得了长足的发展,并在各种社会网络平台的应用中取得了不错的效果. 但社会网络谣言检测问题依然存在如下挑战:

(1)对谣言以及非谣言数据的类型缺乏必要的分析,缺少灵活、准确的信息表征手段,试图使用一套通用的特征集合表征社会网络中的全部信息,训练出来的谣言分类器容易陷入"过拟合"的状态[52]. 由于社会网络安全策略的限制,现有方法主要采取基于关键字或基于监控用户两种采集方式,具有非常强的同质性,难以真实有效地获取社会网络中不同类别的信息以及不同社会网络中用户的行为. 因此,即使训练出来的分类器在训练数据集合中具有较好的表现,在实际应用中也存在检测率低,误报率较高的问题.

(2)缺乏应对突发、海量的社会网络信息的自适应处理能力,现有方法试图对社会网络中所有信息或是所有主题信息的真实性进行鉴别,检测时延大、效率低. 在面对新发布的信息时,由于缺少必要的线索,存在检测的"冷启动"问题[74]. 在面对这类信息时,现有的谣言检测手段往往会显得束手无策,无法在谣言信息造成较大影响之前就进行有效的预防,而只能检测出当前已经在社会网络中广为传播的谣言信息. 因此,如何在谣言造成较大影响之前检测出谣言是当前社会网络谣言检测领域最迫切需要解决的问题.

(3)缺少有效应对训练数据集存在的不均衡与小样本问题的技术手段,导致训练出来谣言分类器存在较为严重的"分类偏倚"问题[70]. 由于社会网络中正常信息远远多于谣言信息,以及社会网络中的基于个人隐私保护的安全保护策略,现有方法收集的数据集中正常信息与谣言信息的数量具有严重的不均衡性. 基于这种数据集训练出来的谣言检测模型在判别时倾向于数据量大的数据类别,导致谣言检测模型检测结果误报率较高,甚至检测模型完全失效,无法真正有效地应用在不断变化的社会网络平台中.

综上所述,现有的社会网络谣言检测方法在面对不断变化的社会网络信息时存在着新的挑战. 研究与开发实时、准确与自适应性强的社会网络谣言检测技术仍然是网络安全与网络舆情领域的迫切需求.

7 结论

自媒体时代下的谣言检测已经刻不容缓. 如何在海量信息中准确高效地检测出谣言信息,是净化网络环境,维护网络空间安全亟待解决的关键问题之一. 本文以社会网络谣言检测问题面临的挑战作为主线,从谣言检测技术的发展历程与谣言检测流程两个维度出发,分别针对现有谣言检测领域相关模型、技术与方法进行归纳与总结. 本文不仅对研究者有借鉴学习作用,还将对网络舆情监测和引导具有重要的实际应用价值.

参考文献

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

[2] 汪林玉,谷科,余飞,等. 基于个人意愿的社会网络团体结构与信息检测方案[J]. 电子学报,2018,46(4):886 – 895.
WANG L-Y,GU K,YU F,et al. Socialcommunity structure and information detection scheme based on personal willingness[J]. Acta Electronica Sinica,2018,46(4):886 – 895. (in Chinese)

[3] GUO H,CAO J,ZHANG Y,et al. Rumor detection with hierarchical social attention network[A]. Proceedings of the 27th ACM International Conference on Information and Knowledge Management[C]. ACM,2018. 943 – 951.

[4] 吴奇,陈福才,黄瑞阳,等. 基于语义路径的异质网络社区发现方法[J]. 电子学报,2016,44(6):1465 – 1471.
WU Q,CHEN F-C,HUANG R-Y,et al. Communitydetection in heterogeneous network with semantic paths[J]. Acta Electronica Sinica,2016,44(6):1465 – 1471. (in Chinese)

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

[6] CAO J,GUO J,LI X,et al. Automatic rumor detection on microblogs: A survey[J]. arXiv Preprint,2018,arXiv:1807. 03505.

[7] 微博发布 2018 年辟谣报告-中国互联网联合辟谣平台[EB/OL]. http://www. piyao. org. cn/2019-02/03/c _ 1210053804. htm. [2019-05-12].

[8] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[A]. Proceedings of the 20th International Conference on World Wide Web[C]. ACM, 2011. 675 – 684.

[9] 姜维,张重生,殷绪成. 基于深度学习的场景文字检测综述[J]. 电子学报,2019,47(5):1152 – 1161.
JANG W, ZHANG C-S, YIN X-C. Deeplearning base scene text detection:a survey[J]. Acta Electronica Sinica, 2019, 47(5):1152 – 1161. (in Chinese)

[10] WU L, LI J, HU X, et al. Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media[G]. 2017. 99 – 107.

[11] CHEN T, WU L, LI X, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[A]. Trends and Applications in Knowledge Discovery and Data Mining[C]. Springer, 2018. 40 – 52.

[12] CANDÈS E J, LI X, MA Y, et al. Robust principal component analysis? [J]. Journal of the ACM (JACM), 2011, 58(3):11.

[13] 李康,李亚敏,胡学敏,等. 基于卷积神经网络的鲁棒高精度目标跟踪算法[J]. 电子学报,2017,45(9):2087 – 2093.
LI K, LI Y-M, HU X-M, et al. Arobust and accurate object tracking algorithm based on convolutional neural network[J]. Acta Electronica Sinica, 2017, 45(9):2087 – 2093. (in Chinese)

[14] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep Learning[M]. MIT Press, 2016.

[15] 陈燕方,李志宇,梁循,等. 在线社会网络谣言检测综述[J]. 计算机学报,2018,41(7):1648 – 1677.
CHEN Y-F, LI Z-Y, LIANG X, et al. Review onrumor detection of online social networks[J]. Chinese Journal of Computers, 2018, 41(7):1648 – 1677. (in Chinese)

[16] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumors in social media: A survey [J]. ACM Computing Surveys, 2018, 51(2):32:1 – 32:36.

[17] 邓国峰,唐贵伍. 网络谣言传播及其社会影响研究[J]. 求索,2005,(10):88 – 90.
DENG G-F, TANG G-W. Internet rumor communication and its social impact research[J]. Seeker, 2005, (10):88 – 90. (in Chinese)

[18] MCARTHUR T, LAM-MCARTHUR J, FONTAINE L. Oxford Companion to the English Language[M]. OUP Oxford, 2018.

[19] HALL M A. Correlation-based feature selection for machine learning[A]. Proceedings of the Seventeenth International Conference on Machine Learning[C]. ACM, 2000. 359 – 366.

[20] KWON S, CHA M, JUNG K, et al. Prominent features of rumor propagation in online social media[A]. IEEE 13th International Conference on Data Mining[C]. USA: IEEE, 2013. 1103 – 1108.

[21] QAZVINIAN V, ROSENGREN E, RADEV D, et al. Rumor has it: Identifying misinformation in microblogs[A]. Proceedings of the Conference on Empirical Methods in Natural Language Processing[C]. USA: ACM, 2011. 1589 – 1599.

[22] Snopes. com. The Definitive Fact-Checking Site and Reference Source For Urban Legends, Folklore, Myths, Rumors, and Misinformation [EB/OL]. https://www.snopes.com/. [2019-05-14].

[23] WEN W, SU S, YU Z. Cross-Lingual Cross-Platform Rumor Verification Pivoting on Multimedia Content[OL]. https://arxiv.org/abs/1808.04911. 2018.

[24] SUN S, LIU H, HE J, et al. Detecting event rumors on Sina Weibo automatically[A]. ISHIKAWA Y, LI J, WANG W, et al. Web Technologies and Applications[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. 120 – 131.

[25] WU K, YANG S, ZHU K Q. False rumors detection on Sina Weibo by propagation structures[A]. IEEE 31st International Conference on Data Engineering[C]. USA: IEEE, 2015. 651 – 662.

[26] CAI G, BI M, LIU J. A novel rumor detection method based on labeled cascade propagation tree[A]. The 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)[C]. USA: IEEE, 2017. 2185 – 2194.

[27] YANG F, LIU Y, YU X, et al. Automatic detection of rumor on Sina Weibo[A]. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics[C]. USA: ACM, 2012. 13.

[28] CAO J, JIN Z, ZHANG Y. MCG-ICT at Media Eval 2016 verifying tweets from both text and visual content[A]. MediaEval 2016 Workshop[C]. MediaEval, 2016. 1 – 3.

[29] DONG S, FAN F-H, HUANG Y-C. Studies on the population dynamics of a rumor-spreading model in online social networks[J]. Physica A: Statistical Mechanics and its Applications, 2018, 492:10 – 20.

[30] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998 (11):1254 – 1259.

[31] Contribute to NLPIR-team/NLPIR Development by Creating an Account on GitHub[OL]. https://github.com/NLPIR-team/NLPIR/projects. 2019.

[32] THUCTC:一个高效的中文文本分类工具[EB/OL]. http://thuctc.thunlp.org/.[2019-05-14].

[33] ZHANG Y,LI Z,GAO C,et al. Mobile social big data:Wechat moments dataset,network applications,and opportunities[J]. IEEE Network,2018,32(3):146 - 153.

[34] LIM S,HAO J,LU Z,et al. Approximating the k-minimum distance rumor source detection in online social networks[A]. The 27th International Conference on Computer Communication and Networks (ICCCN)[C]. USA:IEEE,2018. 1 - 9.

[35] 柏文言,张闯,徐克付,等. 一种融合用户关系的自适应微博话题跟踪方法[J]. 电子学报,2017,45(6):1375 - 1381.
BAI W-Y,ZHANG C,XU K-F,et al. Aself-adaptive microblog topic tracking method by user relationship [J]. Acta Electronica Sinica,2017,45(6):1375 - 1381. (in Chinese)

[36] GIASEMIDIS G,SINGLETON C,AGRAFIOTIS I,et al. Determining the veracity of rumours on Twitter[A]. International Conference on Social Informatics[C]. Springer,2016. 185 - 205.

[37] FactCheck. org[EB/OL]. https://www. factcheck. org/.[2019-05-14].

[38] Snopes. com. The Definitive Fact-Checking Site and Reference Source for Urban Legends,Folklore,Myths,Rumors,and Misinformation[OL]. https://www. snopes. com. 2019.

[39] 微博辟谣的微博_微博[EB/OL]. https://weibo. com/weibopiyao? refer_flag = 1005055013_&is_all = 1.[2019-05-12].

[40] 微博社区管理中心[EB/OL]. https://service. account. weibo. com/? bottomnav = 1&wvr = 6&is_redirected = 1.[2019-05-20].

[41] WANG A H. Don't follow me:Spam detection in Twitter[A]. 2010 International Conference on Security and Cryptography (SECRYPT)[C]. Springer,2010. 1 - 10.

[42] RATKIEWICZ J,CONOVER M,MEISS M,et al. Detecting and tracking the spread of astroturf memes in microblog streams[J]. arXiv Preprint,2010,arXiv:1011. 3768.

[43] BLEI D M. Introduction to Probabilistic Topic Models[OL]. https://www. seas. harvard. edu/courses/cs281/papers/blei-2011. pdf. 2019.

[44] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3(1):993-1022.

[45] MENDOZA M,POBLETE B,CASTILLO C. Twitter under crisis:can we trust what we RT? [A]. Proceedings of the First Workshop on Social Media Analytics[C]. New York,NY,USA:ACM,2010. 71 - 79.

[46] 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(3):99 - 108.

[47] LIANG G,YANG J,XU C. Automatic rumors identification on Sina Weibo[A]. The 12th International Conference on Natural Computation,Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)[C]. USA:IEEE,2016. 1523 - 1531.

[48] JIANG W,CHEN B,HE L,et al. Features of rumor spreading on WeChat moments[A]. MORISHIMA A,ZHANG R,ZHANG W,et al. Web Technologies and Applications[C]. Cham:Springer International Publishing,2016. 217 - 227.

[49] GUPTA M,ZHAO P,HAN J. Evaluating event credibility on twitter[A]. Proceedings of the 2012 SIAM International Conference on Data Mining[C]. SIAM,2012. 153 - 164.

[50] JIN Z,CAO J,JIANG Y,et al. News credibility evaluation on microblog with a hierarchical propagation model[A]. IEEE International Conference on Data Mining[C]. USA:IEEE,2014. 230 - 239.

[51] HUANG W,YAN H,LIU R,et al. F-score feature selection based Bayesian reconstruction of visual image from human brain activity[J]. Neurocomputing,2018,316:202 - 209.

[52] CHEN W,ZHANG Y,YEO C K,et al. Unsupervised rumor detection based on users' behaviors using neural networks[J]. Pattern Recognition Letters,2018,105:226 - 233.

[53] MIKOLOV T,KARAFIÁT M,BURGET L,et al. Recurrent neural network based language model[A]. Eleventh Annual Conference of The International Speech Communication Association[C]. INTERSPEECH,2010. 1 - 4.

[54] SCHUSTER M,PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing,1997,45(11):2673 - 2681.

[55] MA J,GAO W,MITRA P,et al. Detecting rumors from microblogs with recurrent neural networks[A]. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence[C]. IJCAI,2016. 3818 - 3824.

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

[57] GUPTA A,LAMBA H,KUMARAGURU P,et al. Faking Sandy:characterizing and identifying fake images on Twitter during Hurricane Sandy[A]. Proceedings of the 22nd International Conference on World Wide Web[C]. USA:ACM,2013. 729 - 736.

[58] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv Preprint, 2014, arXiv:1409.1259.
[59] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks[A]. International Conference on Machine Learning[C]. USA: ACM, 2013. 1310 – 1318.
[60] HANIN B. Which neural net architectures give rise to exploding and vanishing gradients? [A]. Advances in Neural Information Processing Systems[C]. The MIT Press, 2018. 582 – 591.
[61] LI Y, PAN Q, WANG S, et al. Disentangled variational auto-encoder for semi-supervised learning[J]. Information Sciences, 2019, 482: 73 – 85.
[62] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey [J]. ACM Computing Surveys (CSUR), 2009, 41(3): 15: 1 – 15: 58.
[63] AJAO O, BHOWMIK D, ZARGARI S. Fake news identification on twitter with hybrid cnn and RNN models[A]. Proceedings of the 9th International Conference on Social Media and Society[C]. USA: ACM, 2018. 226 – 230.
[64] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[A]. Advances in Neural Information Processing Systems[C]. USA: ACM, 2014. 2204 – 2212.
[65] JIN Z, CAO J, ZHANG Y, et al. Novel visual and statistical image features for microblogs news verification[J]. IEEE Transactions on Multimedia, 2017, 19 (3): 598 – 608.
[66] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv Preprint, 2014, arXiv: 1409.0473.
[67] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[A]. Proceedings of the 32Nd International Conference on International Conference on Machine Learning[C]. JMLR. org, 2015. 2048 – 2057.
[68] LIAO X, HUANG Z, YANG D, et al. Rumor detection in social media based on a hierarchical attention network [J]. SCIENTIA SINICA Information, 2018, 48 (11): 1558 – 1574.
[69] XU N, GUANDAN C, MAO W. MNRD: a merged neural model for rumor detection in social media[A]. The 2018 International Joint Conference on Neural Networks[C]. USA: IEEE, 2018. DOI: 10.1109/IJCNN.2018.8489582.
[70] TOLOSI L, TAGAREV A, GEORGIEV G. An analysis of event-agnostic features for rumour classification in twitter [A]. Tenth International AAAI Conference on Web and Social Media[C]. AAAI Press, 2016. 151 – 158.
[71] thunlp/Chinese _ Rumor _ Dataset [EB/OL]. https://github.com/thunlp/Chinese_Rumor_Dataset. [2019-11-22].
[72] BENESTY J, CHEN J, HUANG Y, et al. Pearson correlation coefficient[A]. Noise Reduction in Speech Processing[G]. Springer, 2009. 1 – 4.
[73] LOBO J M, JIMÉNEZ-VALVERDE A, REAL R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology and Biogeography, 2008, 17(2): 145 – 151.
[74] YANG P, MAO K, ZHONG X, et al. Service recommendation with case-based reasoning[A]. IEEE 12th International Conference on Networking, Sensing and Control [C]. USA: IEEE, 2015. 631 – 635.

作者简介

高玉君　女,1995 年 10 月出生,江西吉安人. 四川大学网络空间安全学院硕士研究生. 主要研究方向为谣言检测与网络安全.
E-mail: yj631@foxmail.com

梁　刚(通讯作者)　男,1976 年 5 月出生,四川成都人. 博士、副教授、硕士生导师. 现为四川大学网络空间安全学院副教授. 主要研究方向为网络安全、智能计算.
E-mail: lianggang@scu.edu.cn