DOI:10.13878/j.cnki.jnuist.2017.06.003

金志威[1,2]　曹娟[1,2]　王博[3]　王蕊[3]　张勇东[1,2]

融合多模态特征的社会多媒体谣言检测技术研究

摘要

以微博为代表的社会媒体的蓬勃发展在加速信息交流的同时,也促使虚假谣言信息迅速在社会网络上传播,造成严重的后果.自动谣言检测问题受到了国内外学术界、产业界的广泛关注.围绕社会多媒体谣言检测这一问题,本文总结了融合多模态特征的谣言检测相关技术.首先从基本概念出发,阐述了谣言的定义和社会多媒体的特点,给出了社会多媒体谣言检测问题的定义.针对谣言检测面临的多模态特征抽取和模型构建两大难点,分别总结和归纳了各种类型的特征及其提取方法和不同的机器学习检测模型.这些特征和算法是检测谣言的基本手段,也是接下来研究的基础,可为进一步谣言检测的研究提供参考.

关键词

谣言检测;社会媒体计算;多媒体计算;深度学习;多模态特征融合;新闻认证

中图分类号 TP393.092

文献标志码 A

收稿日期 2017-08-28

资助项目 国家自然科学基金(61525206);中国电科创新基金(16105501);中国电科联合基金(20166141B08020101)

作者简介

金志威,男,博士,主要研究方向为数据挖掘、多媒体计算.jinzhiwei@ict.ac.cn

张勇东(通信作者),男,博士,教授,主要研究方向为多媒体计算、视频内容分析与理解.zhyd@ict.ac.cn

1 中国科学院计算技术研究所 智能信息处理实验室,北京,100190

2 中国科学院大学,北京,100049

3 中国电子科技集团公司电子科学研究院创新中心,北京,100041

0 引言

随着 Web2.0 时代的到来,各种社会媒体应运而生.以微博为代表的社会媒体通过开放平台鼓励用户自己生产内容(User Generated Content,UGC),并通过社交网络进行发布、分享、交流和传播.这种基于社会媒体发布、分享多媒体内容的社交行为方式成为人们生活中不可或缺的一部分,对社会产生了巨大的影响.

社会媒体平台以其开发与便捷性,极大地促进了新闻信息的快速交流,成为当今社会人们获取信息资源的重要手段.根据中国互联网络信息中心(CNNIC)2017 年 1 月发布的第 39 次《中国互联网络发展状况统计报告》[1]表明,截止 2016 年 12 月,我国网民规模已达 7.31 亿,其中 84%的网民通过互联网获取新闻.对媒体工作者而言,社会媒体也是重要的新闻线索来源:根据 2011 年的统计数据,超过 80%的社会重大新闻第一手信息来源于微博[2].

然而,社交平台在加速信息公开的同时,也带来了谣言等虚假信息的泛滥.由于普通用户的媒介素养参差不齐,造成 UGC 新闻普遍存在着虚假、差错、欠准确等问题.在缺乏有效的新闻认证技术以及“抢新闻”、“追热点”的心态下,大量公众人物和主流媒体无意间推转相关虚假新闻,成为很多网络谣言和虚假报道的推波助澜者,严重损害了他们的媒体公信力.据《中国新媒体发展报告(2013)》[3]统计的 2012 年的 100 件微博热点舆情案例中,有 1/3 的热点事件出现了谣言.国外的网络谣言问题同样不容乐观.在 2016 年美国总统大选期间,大量谣言在 Facebook、Twitter 上广泛传播,甚至被指控严重影响了美国大选结果[4].

网络谣言的广泛传播会侵害到个体和社会的发展,对个体情感、社会经济、政治稳定发展方面产生严重的负面影响.2013 年的 10 大假新闻之一“深圳 90 后女孩当街给残疾乞丐喂饭感动路人”,严重伤害了公众的感情;2011 年响水县“爆炸谣言”引发十几万人大逃亡,4 人遇难,严重危害社会稳定;2013 年 1 条据称来自美联社的 Twitter 消息说,“白宫发生 2 起爆炸,美国总统奥巴马受伤”,导致美国股指暴跌,短时间内市值蒸发了 2 000 亿美元,产生巨大经济损失.

当前,世界各国纷纷采取措施推动互联网谣言检测的技术研究与应用.在美国,2017 年初,企业代表 Facebook 在该平台上线了一个“虚假标签”模块供用户手动举报,若有多名用户举报则该条消息会

自动显示“虚假新闻”的标签予以提醒.在英国,媒体机构代表 BBC 将要成立核实组,重点打击网络媒体上的虚构性及有误导性的新闻,此计划已获得 2.9 亿英镑的项目支持.欧盟也于 2014 年初分别成立了 2 个叫做“PHEME”和“REVEAL”的网络谣言自动检测计划,前者由英国谢菲尔德大学带领 15 个研究机构共同承担,主要侧重网络内容可信度计算的理论研究;后者由多家企业联合承担,主要侧重网络谣言检测的产业化.在中国,受中宣部的委托,2013 年底,新华社联合中国科学院计算机研究所研发了一个互联网新闻认证系统[5].

由于社会媒体上的信息数量巨大、非结构性、不完备、噪声多等特点,自动化地检测谣言仍然面临着许多挑战.首先,无法仅仅基于文本内容来有效检测谣言.因为谣言多是蓄意捏造出来误导大众的报道,通常手段是将虚假信息糅杂在部分真实情况中,很难仅根据内容判定其真假;同时,谣言在话题选择、语言风格等方面千变万化,这导致了传统的基于人工特征的、针对某一类特定数据的文本分析算法无法有效检测出社会媒体谣言,必须借助社交网络上的用户参与、内容传播链路、多媒体内容等多种辅助信息来提高谣言检测准确率.而这又带来谣言检测的第 2 大挑战:如何有效地利用这些大规模、异构的、跨模态的辅助信息来检测谣言.

针对社会媒体谣言检测的挑战和发展,在厘清社会媒体谣言检测相关概念后,本文重点介绍了基于多模态融合的方法检测谣言的关键技术,特别是从特征抽取和模型构建 2 个方面展开阐述,对谣言检测问题中的多模态特征以及特征融合方法进行阐述.

1 社会多媒体谣言检测概念

相对于各界对谣言检测问题的关注度而言,社会多媒体谣言检测技术在研究领域的发展才刚刚起步,且出现了一些理解上的偏差.如一些研究团队通过媒体报道声称目前的谣言检测精度已经达到 90% 以上,甚至已经解决等,给研究者们造成很多困惑和误解.究其原因,主要在于对于谣言检测问题理解上的不同,如什么是谣言,谣言检测的类型等.另一方面,基于社会多媒体的谣言检测必然不能脱离社会媒体自身的特点进行孤立地研究.为此,本节首先厘清谣言检测问题的定义,再结合社会多媒体的定义和特点,综合阐述了社会多媒体谣言检测的相关概念,最后给出谣言检测问题的严格形式化定义.

1.1 谣言的定义

谣言,又称作“虚假传言”、“虚假新闻”等,在传统社会心理学上被定义为“真实值不确定或者故意伪造的报道或声明”[6-7].而在实际研究与应用中,多数研究者从谣言的“故意伪造”这个角度出发,将权威渠道证实确实是伪造虚构的消息认定为谣言[8-13].基于该定义,在标注谣言时从 Snopes.com[4]、微博谣言举报平台[12-13]等权威渠道获知每条消息是否为谣言,能够快速得到大量权威标注数据.该定义无法判断预测性、情感性等类型谣言的真伪,因为这类谣言往往还不能够证伪.

这种客观定义的谣言,由于其具有标注权威准确、数据易收集的特点,被谣言自动检测界广泛采用.鉴于本文关注于如何检测有害谣言并防止其继续传播造成危害,本文后续所有谣言都是指客观定义的谣言.

1.2 社会多媒体的定义

针对社会多媒体的谣言检测技术,需要充分挖掘社会媒体的特征,利用其提供的多种资源.社会多媒体通常被定义为“支持个体参与、社区形成和社会交互的在线多媒体资源”[16].该定义指出了社会媒体的 3 个核心要素:多媒体内容、网络用户以及用户与媒体内容之间的交互(图 1).

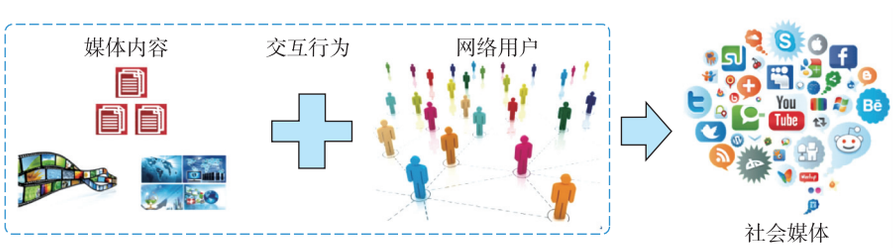


图 1 社会媒体的组成

Fig. 1 The content of social media

1)多媒体内容.社会媒体网络上的内容由多种不同模态的内容组成,主要包括文字、图片、视频、语音等.与传统单一模态媒体相比,在社会媒体上发布的内容通常包含一种以上的内容形式,从而增加了内容表现力,使其能够得到更广泛的传播和关注.

2)网络用户.在社会多媒体中,网络用户既是内容的生产者,又是内容的消费者,是社会多媒体的一个非常重要的组成部分.社交媒体平台允许用户编辑的特点,使得用户从信息的被动接受者成为一个主动的贡献者.用户的广泛参与使得大量的 UGC 内容出现在社会媒体平台上,极大地促进了社会多媒体内容的繁荣.如果将网络用户理解为数据感知器,社会多媒体实际上是由用户所见、所听、所说、所想组成的.

3)用户与多媒体内容的交互.用户和媒体内容是社会媒体中的 2 个基本元素,通过交互行为,孤立的各个元素间形成了相互连接的网络:①用户之间的交互,包括"加好友"、"关注"、"收听"等方式构建成一个庞大的用户社交网络,也正是多媒体内容传播的网络;②多媒体内容通过标签、话题、超链接等形式构建相互连接,形成不同的内容子话题,这些连接关系对分析多媒体内容有重要作用;③用户对多媒体内容进行上传、评论、转发、标注等操作与其进行交互,促使用户和多媒体内容之间建立了丰富的社会关系.

与传统的单一模态、孤立的内容分析相比,社会多媒体在内容和用户交互上具有多模态性和互联性,如何利用这些特性进行高效的谣言检测成为当前研究的重点.

1.3 谣言检测形式化定义

给定一个新闻事件 e,其包含了 n 条相关微博消息 $M=\{m_1,m_2,\cdots,m_n\}$ 以及对应每条消息的发布用户 $U=\{u_1,u_2,\cdots,u_n\}$.每条消息 m_i 由一组表示消息文本、图片等内容的属性表示.每个用户 u_i 由用户名、年龄、注册时间等一系列代表用户的属性表示.现定义谣言检测问题如下:

定义 1(谣言检测) 给定新闻事件 e,其包含了微博消息集合 M 以及对应用户集合 U,谣言检测任务定义为预测该事件是否为虚假事件,即学习预测函数 $:e\rightarrow\{0,1\}$ 满足:

$$(e)=\begin{cases}1, & \text{如果 } e \text{ 是谣言},\\ 0, & \text{其他情况}.\end{cases} \tag{1}$$

从定义 1 可以看出我们把谣言检测问题定义为一个基于内容和用户的二分类问题.谣言检测的目标即为学习分类预测函数 来区分谣言事件和真实事件.下面介绍谣言检测的一般性方法.这里主要涉及到 2 个方面的研究重点,一是如何有效地表示谣言事件的特征,二是如何利用这些特征来检测谣言.为此,从特征抽取和模型构建 2 个方面展开介绍.特征抽取研究如何从文本、图片、用户等事件包含的丰富的多媒体内容中抽取出有效信息,并把它们表示成结构化的数学形式.在此基础上,模型构建基于这些特征表达利用机器学习模型来检测谣言.近年来,一些基于深度神经网络的方法将特征抽取与模型学习整合到一个端到端的网络中,本文也将对这些工作进行介绍.

2 谣言检测特征抽取

传统的新闻报道通常只包含新闻本身的内容,而在社会媒体上,新闻消息会附带有其他社会属性的内容,这些辅助内容能够用来提高谣言的特征表达性.如图 2 所示的一则谣言消息中,就包含了文本内容(包括文字描述、话题和外部链接等)、图片内容(2 张图片)和一些社交内容(转发、评论等).为此,将介绍如何从消息内容和社交属性 2 个方面提取有效特征来表达新闻消息.

2.1 内容特征

新闻事件 e 其包含的微博消息集合 M 描述了新闻事件的关键信息.主要包含以下几个方面的属性:

1)文本内容:主体的一段话来描述新闻事件.通常有能够体现作者观点和立场的重要结论,或支持性描述.

2)图片/视频:有些消息会通过附图片/视频的方式给文字描述提供视觉支撑.

3)其他内容:社会媒体特有的语言交流方式会产生额外的内容信息,比如话题(##)、用户提醒(@)、超链接(URL)、表情符号等.

基于这些原始的内容属性,各种各样的内容特征被提取出来以区分谣言特性.通常这些特征可以分为文本特征和视觉特征 2 大类.下面介绍这 2 类特征的主要抽取方法.

2.1.1 文本特征

谣言通常是蓄意捏造的,有误导大众意图的虚假信息而不是客观的事实报道,因而它们通常包含着一些观点性或者煽动性的语言,即所谓的"标题党",来引诱大众关注和传播.例如,文献[14]通过分

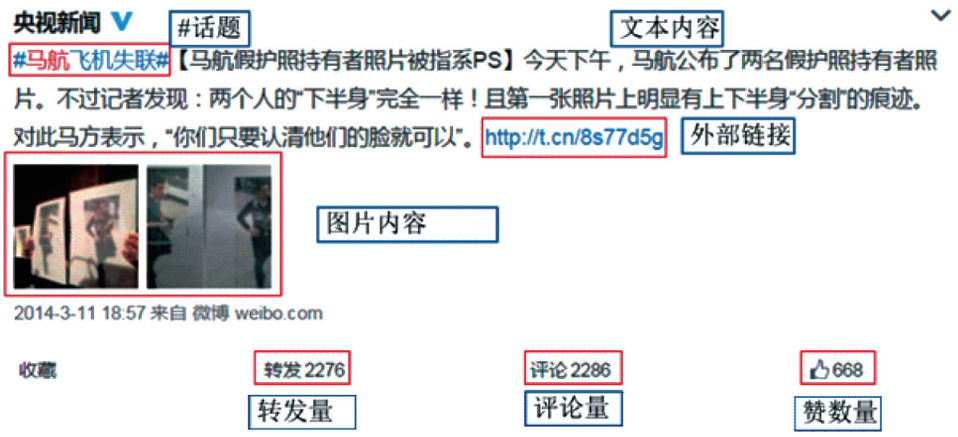


图 2　微博上的谣言消息示例,它包含了来自内容和社交属性的特征

Fig. 2　A rumor example from Weibo, which contains content features and social context features

析大量谣言信息流发现谣言在语言模式上具有"求真性"和"质疑性"2 大类语言模式.所以,可以通过抽取语言学特征来描述谣言消息与真实消息的不同特点.

文本特征通常从文本内容的不同组织维度上抽取,包括字、词、句、消息、消息集合等.为了更加全面地描述文本内容,现有的研究工作不仅提出了一般性的文本特征,也结合平台特点提出了领域相关的文本特征.

一般性的文本特征是指在其他自然语言处理任务中被广泛应用的一类特征.常见的语言特征有:

1)词法特征:单个字级别的或单个词级别的语言特征,包括总字数、总词数、不同单词个数、每个词平均长度等[8].

2)句法特征:句子级别的语言特征,包括关键词频数(n-grams 模型和词袋模型[17])、标点符号类型和数目,以及词性标注等.

3)主题特征:主题级别的语言特征,例如对整个文档集构建主题模型(topic model[18]),还有提取的消息话题特征、消息的情感倾向特征等.

领域相关的文本特征是指跟发布平台、消息类型有关的一些特征,比如外部链接、应用图片数量、消息长度等[19].其他的一些语言特征也能一定程度上捕捉文本的写作风格用来检测谣言,比如谎言检测特征[20].

2.1.2　视觉特征

视觉内容在谣言产生和传播方面有着重要的作用.一方面,图片等视觉内容在社交网络上广泛存在.受限于单条微博的字数限制,越来越多的微博消息通过图片形式辅助传递信息.文献[13]指出超过 51.6%的微博带有图片.另一方面,图片对于新闻信息的传播具有重要影响.相比于纯文本内容,图片能够生动形象地描述具体场景,吸引到更多的注意力.统计发现,平均而言,带有图片的微博获得的转发量是不带图片微博的 11 倍(191 比 16)[13].如此巨大的差距体现了图片在信息传播过程中的重要作用.基于上述分析,很有必要综合利用图片等视觉内容辅助进行谣言检测.

视觉特征指从以图片视频等视觉内容为中心抽取的一组特征,根据特征抽取方式的不同,视觉特征大致可以分为以下 3 类:图片相关特征、视觉内容特征以及深度学习特征.

1)视觉统计特征

视觉统计特征通常直接从图片附属的属性抽取特征而对其具体视觉内容不做分析.在文献[7]中定义了一个特征来描述用户是否包含头像,用来评估该用户的可信度.文献[21]中定义了一个微博级的"has multimedia"特征来描述微博是否包含有多媒体信息这一状态.Gupta 等[22]提出一种分类方法来识别飓风发生期间的各类虚假图片.文献[10]发现虚假新闻更有可能包含之前已经发布过的过时图片,因此他们定义了图片发布时间延迟这一特征,并用搜索引擎发现和获取原始图片的发布时间.Boididou 等[23]提出了一项验证多媒体使用(Verifying Multimedia Use)的任务,以致力于自动预测包含多媒体内容的微博是否为假.文献[13]提出 7 种统计特征,描述微博事件中图片大小、图片比例、图片热点等特点.

2)视觉内容特征

传统的基于内容的图片视觉特征从视觉语义的角度描述了图片内容.而针对谣言检测这一任务,我们通常并不关心图片是否描述了某一特定对象或者场景.我们需要从区分谣言事件的角度分析图片在真假事件中不同的分布特点.如图 3 所示,通过观察真假 2 个不同事件中的热门图片,可以发现,真新闻里图片更多,差异性更大,而假新闻里,图片多样性更差[13].因此,在视觉特性上,文献[13]提出 5 个能够准确描述图片视觉分布的特征:

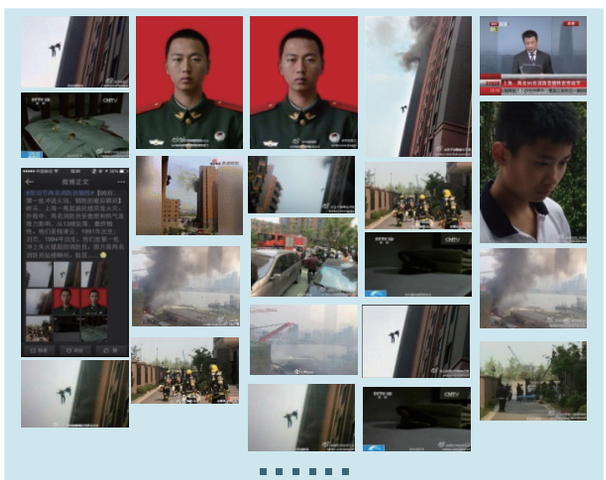
①视觉清晰度特征(visual clarity score)度量 2 个图片集的分布差异.一个是指定新闻事件中的图片集(事件集),另一个是包含所有图片的全集.这个特征背后的逻辑很简单:如果一个事件集和全集中的图片分布差距很大,那么这个事件很有可能是真实事件.这是基于真实事件中包含大量原创性图片的假设.可以通过构建 2 个语言模型来计算这一特征,即分别对事件集和全集构建视觉词汇语言模型.视觉清晰度就定义为这 2 个模型之间的 KL 散度,图片集的语言模型可以用视觉词袋模型得到.

②视觉一致度特征(visual coherence score)描述了同一事件中的图片是否具有一致性.相关的图片通常会具有相似的视觉外观,通过计算视觉一致度,能够量化出同一事件中的图片管理程度.这里定义视觉一致度为事件内任意图片对相似度的平均值.

③视觉相似性直方图(visual similarity distribution histogram)从更加精细的粒度上衡量图片集的一致性程度.该特征是基于事件中所有图片的相似度矩阵计算的.首先计算两两图片之间的相似度得到相似度矩阵,然后将矩阵量化就能得到对应的直方图.

④视觉多样度(visual diversity score)度量了指定新闻事件图片集中的视觉上的差异程度.和视觉一致度相比,这个特征直接计算了图片的多样性分布特点,而且更加强调代表性的图片.我们定义一个图片的多样度为该图片到排在其之前的图片中的最小的距离.视觉一致度计算的是整个图片集上相似度的算术平均,而视觉多样度计算的是不相似度的加权平均.在社会多媒体网络上,可通过图片获得的转发量来排序图片.因此视觉多样度打分能够加重这些代表性图片的权重,减少事件中噪音图片的干扰.

⑤视觉聚类度(visual clustering score)从图片聚类的角度衡量了图片的视觉分布特点.它被定义为图片集中聚类得到的类簇的个数.我们采用分层聚合聚类算法自底向上地将相似图片聚集成类.相比于其他聚类算法,如 *K*-means,该算法不需要事先指定聚类个数,而能根据数据分布特点自动聚集出若干个类.设定相同的参数下,该算法能够揭示出图片集的多样性特点.我们移除了数量小于 3 的小类,并把剩下的类的个数记作视觉聚类度.



a. 真实事件案例:天津消防员救火牺牲

b. 谣言事件案例:美军基地检测到马航MH370雷达信号

图 3　图片在真假新闻案例中的不同分布

Fig. 3　The distributional difference of images in real and rumor events

3)深度学习特征

近年来,以卷积神经网络(Convolutional Neural Networks,CNN)为代表的深度神经网络算法在视觉表征学习上展示出了远超传统浅层模型的优良效果.对于很多计算机视觉任务,包括图片分类[24-25]和对象检测[26-27],CNN 都明显优于传统的手工构造的特征方法.在谣言检测方面,文献[28]提出利用 CNN 来学习谣言图片中的复杂语义特征.一个典型的

CNN 包含了一系列卷积层和全连接层.一个深度卷积神经网络通常包含了数以百万计的参数,这些参数在模型训练的过程中得以学习.比如,AlexNet 就包含了超过 6 000 万的参数[2].要训练这样一个复杂的神经网络通常需要大量的标注样本,而现有的虚假图片数据集太小,不能满足直接训练的需求,因此文献[28]提出利用深度迁移学习来解决特征学习和标注数据集缺乏的难题.

2.2 社交特征

社会媒体最大的特点之一就是广泛的互联性,主要包括 3 个方面的互联关系.一是用户之间的交互:社会媒体用户通过“加好友”、“关注”、“收听”等方式构建成一个庞大的社交网络,多媒体内容正是通过该网络进行快速传播;二是媒体内容的交互:多媒体内容通过标签、话题、超链接等形式构建相互连接,这些连接关系对分析多媒体内容有重要作用;三是用户与媒体内容的交互:用户对多媒体内容进行上传、评论、转发、标注等操作与其进行交互,促使用户和多媒体内容之间建立了丰富的社会关系.如转发过同一个视频的用户之间存在联系,由同一个用户上传的图片和视频之间存在联系等.

因此,社会媒体上的谣言检测,除了直接抽取谣言的内容特征外,还需要充分挖据这些互联关系网络中形成的各类特征.下面分别从用户网络、内容网络和交互网络 3 个方面介绍基于社交属性的谣言检测特征.

2.2.1 基于用户的社交特征

谣言传播过程中,可能存在大量“水军”推波助澜,或者一些恶意账户故意捏造、传播.前文也分析过不同类型的账户对大众具有不同的可信度.因此利用用户画像的方法抽取基于用户的特征能够帮助提高谣言检测准确率.基于用户的社交特征是指描述用户在社交网络中传播信息时展现出来的特点.从不同的粒度看,这些特征可以分为 2 大类:个体特征和群组特征.

1)个体特征

个体特征是指针对单个用户的各项统计指标中抽取出来,用来分析该特点用户可信度的一系列特征.主要包括注册时间、用户名类型、年龄、性别、粉丝数、关注数、发布微博数等[8].

2)群组特征

群组特征描述的是在信息传播过程中具有相似性的某个用户群体的整体特征[9].抽取该类特征时的一个基本的假设就是传播谣言的社区和传播真实消息的社区各不相同并且有不同的特点.群组特征通常是从个体特征聚合而来的,例如认证用户的比例、平均粉丝数等[29-30].

2.2.2 基于内容的社交特征

新闻事件在社会媒体上传播的过程中,不同的用户会通过转发、评论的方式表达各自的观点、情感倾向,例如质疑原文真实性的态度、反感的情绪表达等.这些来自社交网络的反馈信息在谣言检测中具有重要的价值.通过抽取基于内容的设计特征,能够有效捕捉这些反馈情感和特征.从考察的不同角度和粒度出发,基于内容的社交特征大致可以分为 3 类:消息级的内容特征、群组级的内容特征和时间片级的内容特征.

1)消息级特征

消息级特征为每条转发或评论的微博抽取特征来描述单条消息.因此上文中提到的各种内容特征提取方法和一些基于词嵌入的模型方法[31]都可以用来提取消息级特征.文献[21,30]采用基于主题模型的方法(LDA)来抽取每条消息的话题特征.

2)群组级特征

内容的群组级特征基于“群体智慧”的思想,从大量消息中总结出谣言检测特征.这些特征通常是通过聚合消息级特征产生的.文献[8]中列举了大量的群组级特征,通过在这些特征上构建决策树来检测谣言事件.文献[11]通过聚类的方式将描述相同话题的消息聚合在一起抽取特征.

3)时间级特征

内容的时间级特征考察的是随着时间的变化事件中消息的特征变化情况[30].无监督的深度神经网络方法(反馈神经网络 RNN)被用来学习消息流随着时间变化的特征[31-32].文献[29]通过考察随时间变化的消息数量变化曲线,抽取特征刻画谣言消息特征.

2.2.3 基于交互网络的社交特征

谣言消息在社交网络上的传播可以形成转发传播树,另一方面参与传播的用户也潜在地隐含在一个用户社交网络中.通过抽取特征来描述这些关系网络就形成了基于交互网络的谣言检测特征.文献[29]通过网络度和聚类系数来描述传播网络和社交网络特征.文献[21]提出一种基于核方法的 SVD 模型来描述简化后的转发树.

图 4 总结了社会媒体谣言检测中常用的各类特征.

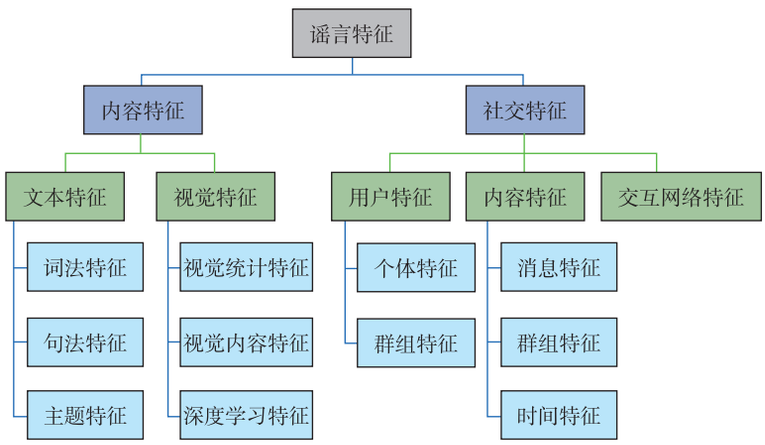


图 4　社会媒体谣言检测特征

Fig. 4　Detection features for rumor on social media

3　谣言检测模型构建

从社会媒体上抽取出谣言检测的大量特征后,如何构建模型分类谣言成为研究的关键.从对特征的不同利用方式出发,目前主要有 2 种模型构建方法:基于特征分类的方法和基于传播的方法.下面以一些典型应用案例出发介绍这 2 类方法.

3.1　基于特征分类的谣言检测模型

谣言检测问题本质上是一个二分类问题,抽取出大量特征后,可以直接对特征进行传统的机器学习建模得到分类器来进行谣言分类.Castillo 等[8]首先应用分类算法决策 Twitter 上新闻事件的真假.他们提取了来自文本内容、用户以及传播等多方面的特征,并比较了这些特征在决策树、SVM 等多种常用分类器上的新闻认证效果.针对中文微博的新闻认证通常也遵循了同样的思路,文献[9-10]提出了几个新的特征来增强中文微博的谣言检测效果,同样采用逻辑回归等传统分类器进行分类.Wu 等[21]提出一种混合 SVM 分类器来检查微博上的谣言.该分类器利用一种随机行走的核方法(random walk graph kernel)来描述单条微博的复杂转发树,并与通常的 RBF 核结合,更加准确地描述了微博传播的特征,取得了良好的谣言检测结果.为进一步整合消息级和群组级特征,Jin 等[33]从消息级和群组级 2 个层次分别进行分类器学习再采用类似于 stacking 的集成学习方法检测多媒体谣言.

3.2　基于传播的谣言检测模型

传统的基于特征的分类算法孤立地分析单条微博或者单个新闻事件的可信度,而忽略了不同微博和事件具有广泛的关联.为此,基于传播的方法被提出来从整体上评估整个内容网络中各消息的真假.该类算法的核心是内容网络的构建和可信度传播算法.具体而言,该类算法通过定义微博间的连接关系将时间相关的所有内容连接成一个可信度传播网络;随后,不同消息的可信度在一定约束条件下在该网络上彼此影响和传播直到收敛.不同消息的初始可信度值可以通过基于分类的方法学习得到,因此该方法往往比简单的分类方法具有更好的认证准确率和稳定性.

设计可靠、合理的可信度传播算法是基于传播的新闻认证方法的关键.不同对象的可信度初值在内容网络上的传播过程可以看作是一种半监督的网络学习模型.作为一种有效的图学习方法,半监督图学习的理论已被广泛的研究和应用[34-35].该类算法的目标是在保持已有标注数据和网络结构一致性的前提下,预测未标注数据的类别.

Gupta 等[22]构造了一个包含用户、微博消息和事件的可信度传播网络,将不同实体基于相似度连接在一起.基于半监督学习的思想,他们用了一种启发式的迭代算法来求解可信度的传播结果.

基于特征的分类算法通常将事件中涉及到的每条信息当成孤立的对象,而没有考虑到内容之间可能存在的内在关系.另一方面,根据标签、话题、超链接等形成的内容网络往往稀疏且噪音多,不能满足谣言检测的需要.文献[11]注意到除了事件级的关联之外(即 2 条消息是否描述了同一个谣言事件),同一事件下的消息还会在社交网络上形成不同子事件.如图 5 所示,在“深圳最美女孩当街为乞讨老人喂饭”这一谣言事件中,随着事件进展,社交网络上

图 5 “深圳最美女孩”谣言事件中的子事件

Fig. 5 Different sub-events in the rumor “a kind girl in Shenzhen helps a homeless old man”

出现了不同的讨论重点,形成了不同的子事件.

每个子事件有不同的可信度,子事件之间也存在一定关联.与孤立地计算每条消息的可信度相比,综合考虑子事件的可信度以及子事件之间的依赖关系能够更加准确地判断新闻事件的真假.为此,文献[11]提出一种分层的内容网络,它能够从微博消息、子事件和事件 3 个不同粒度全面地考察新闻事件,构建更加真实的可信度传播网络.其中子事件通过聚类算法将语义相似的微博消息聚合而成.

对于一个新闻事件来说,一个分层的内容网络由 3 层网络(消息层、子事件层和事件层)以及它们之间的边组成.如图 6 所示,该网络中有 3 种在上节中定义的实体:消息 m、子事件 s 和事件 e,以及 4 种类型的边:消息到子事件之间的边($g(m_i,s_j)$)、子事件到事件之间的边($p(s_i,e_j)$)、消息之间互联的边($f(m_i,m_j)$)以及子事件之间互联的边($h(s_i,s_j)$).各边的权重都定义为该边 2 个定点的函数.通过子事件聚类,消息连接到对应的子事件.

该网络中各类型的边权重计算方法如下:

1) 消息-消息.在可信度传播网络中,消息间的边权值决定了每条消息是如何影响其他消息的可信度的.假定相似的消息很大程度上具有相似的可信度值,这样,2 条消息越相似,它们之间的边权重就越大.考虑到微博是 140 字以内的短文本,可利用 Jaccard 系数来计算 2 条消息的 unigram 序列之间的相似度.同时考虑 2 条消息的情感值极性,定义不同情感倾向的消息之间的边权值为 0,相同情感倾向的消息之间的边权值正比于 2 条消息的内容相似度.

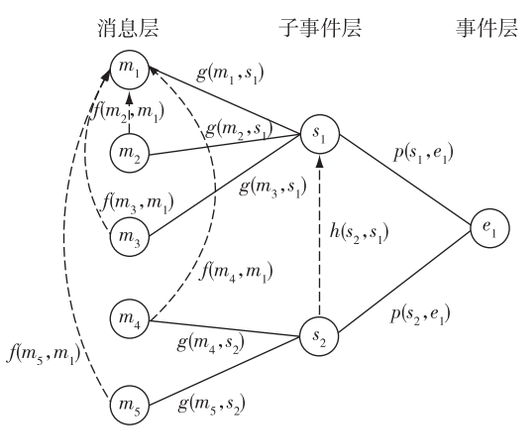


图 6 3 层的内容网络结构示意

Fig. 6 Structure of a 3-layer content network

2) 子事件-子事件.同理,相似子事件之间应该有更强的关联性.用每个子事件聚类中心代表该子事件,这样,通过计算 2 个聚类中心之间的余弦距离,可以得到子事件之间的关联度.

3) 消息-子事件.定义一条消息对所在子事件的影响来自 2 个方面:一是消息与子事件的一致程度,二是消息在子事件中的重要程度.其中一致性可由文本相似度来刻画,重要性由媒体转发量来刻画.

4) 子事件-事件.子事件对事件的影响同样也由相似度和转发重要程度 2 个方面决定.

通过把不同实体在该分层网络上的可信度传播过程定义为一个图优化问题,定义损失函数后,利用梯度下降算法可以得到该函数的迭代解,从而得到各实体的最终可信度值.

4 小结

社会媒体由于其开放性、实时性和交互性,成为当今社会人们发布、获取、传播信息的重要渠道.然而由于缺乏有效监管,大量虚假谣言信息的泛滥不仅损害媒体公信力,还有可能造成重大的经济、政治损失,破坏网络舆情环境和社会稳定.针对自动化谣言检测这一问题,本文首先阐述了谣言的各种定义以及社会媒体的特性,并以此给出谣言检测的明确定义.针对谣言检测面临的特征抽取和模型构建2大难题,文章总结概括了现有工作中的各种方法.具体而言,从网络谣言的内容和社交属性2个方面出发,介绍了谣言检测中应用的5大子类的特征.这些特征全面描述了谣言的文本、视觉内容和社交化属性,为构造有效的谣言检测算法提供了基础.在谣言检测模型方面,文章总结了现有工作中的2大类算法.基于特征的分类方法简单有效,但受限于人工构造的特征以及模型表达能力,通常效果不是最优的.基于传播的算法能够有效利用谣言的社会属性构建内容网络来检测谣言.本文总结的各类特征方法提供了构建一个有效谣言检测算法的指南,同时也为进一步研究提供了参考.

参考文献

References

[1] 中国互联网络信息中心.中国互联网络发展状况统计报告[R].2017
China Internet Network Information Center. Statistical report on the development of Internet in China[R].2017
[2] 刘琼.中国网络新闻可信度研究[D].武汉:华中科技大学新闻与信息传播学院,2011
LIU Qiong. Study on China's Internet news credibility [D]. Wuhan: Journalism and Information Communication School, Huazhong University of Science and Technology,2011
[3] 唐绪军.中国新媒体发展报告[M].北京:社会科学文献出版社,2013
TANG Xujun. Annual report on development of new media in China[M]. Beijing: Social Sciences Academic Press,2013
[4] Jin Z W, Cao J, Guo H, et al. Detection and analysis of 2016 US presidential election related rumors on twitter [C]// International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation,2017:14-24
[5] Zhou X, Cao J, Jin Z W, et al. Real-time news certification system on Sina weibo[C]// Proceedings of the 24th International Conference on World Wide Web, 2015: 983-988
[6] Allport G W, Postman L. The psychology of rumor[M]. New York: Heney Holt and Company,1947
[7] Gupta M, Zhao P X, Han J W. Evaluating event credibility on twitter[C]// Proceedings of the SIAM International Conference on Data Mining,2012:153-164
[8] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]// Proceedings of the 20th International Conference on World Wide Web,2011:675-684
[9] 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,2012:13
[10] Sun S Y, Liu H Y, He J, et al. Detecting event rumors on Sina weibo automatically[C]// Asia-Pacific Web Conference: Web Technologies and Applications,2013:120-131
[11] Jin Z W, Cao J, Jiang Y G, et al. News credibility evaluation on microblog with a hierarchical propagation model [C]// IEEE International Conference on Data Mining, 2014:230-239
[12] Jin Z W, Cao J, Zhang Y D, et al. News verification by exploiting conflicting social viewpoints in microblogs[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,2016:2972-2978
[13] 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,2017,19(3):598-608
[14] Zhao Z, Resnick P, Mei Q Z. 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
[15] Morris M R, Counts S, Roseway A, et al. Tweeting is believing? Understanding microblog credibility perceptions [C]// ACM Conference on Computer Supported Cooperative Work,2012:441-450
[16] Naaman M. Social multimedia: Highlighting opportunities for search and mining of multimedia data in social media applications[J]. Multimedia Tools and Applications, 2012,56(1):9-34
[17] Fürnkranz J. A study using *n*-gram features for text categorization[J]. Austrian Research Institute for Artificial Intelligence,1998,3:1-10
[18] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine Learning Research, 2003, 3: 993-1022
[19] Potthast M, Kiesel J, Reinartz K, et al. A stylometric inquiry into hyperpartisan and fake news[J]. arXiv e-print,2017,arXiv:1702.05638
[20] Afroz S, Brennan M, Greenstadt R. Detecting hoaxes, frauds, and deception in writing style online[C]// IEEE Symposium on Security and Privacy,2012:461-475
[21] Wu K, Yang S, Zhu K Q. False rumors detection on Sina weibo by propagation structures[C]// IEEE International Conference on Data Engineering,2015:651-662
[22] Gupta A, Lamba H, Kumaraguru P, et al. Faking Sandy: Characterizing and identifying fake images on twitter during hurricane Sandy[C]// Proceedings of the 22nd International Conference on World Wide Web, 2013: 729-736
[23] Boididou C, Papadopoulos S, Dang-Nguyen D, et al. Verifying multimedia use at mediaEval 2015[C]// MediaEval

Workshop,2015:235-237
[24] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems,2012:1097-1105
[25] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition[J].arXiv e-print, 2014,arXiv:1409.1556
[26] Girshick R,Donahue J,Darrell T,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:580-587
[27] Szegedy C,Toshev A,Erhan D.Deep neural networks for object detection[J].Advances in Neural Information Processing Systems,2013:2553-2561
[28] Jin Z W,Cao J,Luo J B,et al.Rumor image detection with effective domain transferred deep networks[J].ACM Transactions on Multimedia Computing,Communications and Application(accepted)
[29] Kwon S,Cha M,Jung K,et al.Prominent features of rumor propagation in online social media[C]// IEEE International Conference on Data Mining,2013:1103-1108
[30] Ma J,Gao W,Wei Z Y,et al.Detect rumors using time series of social context information on microblogging websites[C]// ACM International Conference on Information and Knowledge Management,2015:1751-1754
[31] Ruchansky N,Seo S,Liu Y.CSI:A hybrid deep model for fake news[J].arXiv e-print,2017,arXiv:1703.06959
[32] Ma J,Gao W,Mitra P,et al.Detecting rumors from microblogs with recurrent neural networks[J].International Joint Conference on Artificial Intelligence, 2016: 3818-3824
[33] Jin Z W,Cao J,Zhang Y Z,et al.MCG-ICT at MediaEval 2015:Verifying multimedia use with a two-level classification model[J].Media Eval,2015
[34] Zhu X J,Ghahramani Z.Learning from labeled and unlabeled data with label propagation[R].CMU Technical Report,CMU-CALD-02-107,2002:19-26
[35] Zhu X J,Ghahramani Z,Lafferty J.Semi-supervised learning using Gaussian fields and harmonic functions[C]// Twentieth International Conference on International Conference on Machine Learning,2003:912-919

Rumor detection on social media with multimodal feature fusion

JIN Zhiwei[1,2] CAO Juan[1,2] WANG Bo[3] WANG Rui[3] ZHANG Yongdong[1,2]
1 Key Laboratory of Intelligent Information Processing,Institute of Computing Technology,CAS,Beijing 100190
2 University of Chinese Academy of Sciences,Beijing 100049
3 Innovation Center,China Academy of Electronics and Information Technology,Beijing 100041

Abstract Social media,such as microblogs,has developed rapidly nowadays,which accelerates the information diffusion on the Internet.However,numerous false rumors fostered on social media are spreading widely on the social network and can result in serious consequences.It has become a huge concern in research and industry areas to detect rumors automatically on social media. Focused on the rumor detection task, this paper summarizes the approaches of multimodal fusion on this problem.Starting from the basic concepts,we give formal definitions of rumors and introduce the characteristics of social media.We summarize the studies on rumor detection into two major parts,i.e.,extracting effective multimodal features to identify rumors and constructing robust models to detect rumors.For each of the research aspects,we give detailed introduction based on existing studies.This paper can be served as a basic guidance to build state-of-the-art rumor detection models and a reference for future researches.

Key words rumor detection;social media computing;multimedia computing;deep learning;multimodal feature fusion;news verification