

# 突发公共卫生事件中谣言识别研究\*

孙冉<sup>1</sup> 安璐<sup>1,2</sup> (¹武汉大学信息管理学院 湖北 430072; ²武汉大学信息资源研究中心 湖北 430072)

**摘要:** [目的/意义]揭示社交媒体环境下识别谣言过程中的关键要素和谣言识别机制,识别突发公共卫生事件中的谣言微博,研究及评估影响谣言识别的重要特征,有助于准确识别网络谣言、维护健康的网络生态环境。[方法/过程]文章抽取谣言微博的用户特征、时间特征、微博文本结构特征、文本语义特征和微博传播特征,结合MAIN理论模型,采用二元逻辑回归方法从信息内容、信息模态、信息源角度对谣言的影响因素深入研究,利用神经网络模型提取文本语义特征,构建融合文本语义特征的多特征谣言识别模型,并通过XGBoost算法计算不同特征在谣言识别中的重要性。[结果/结论]正向评论情感度、用户发布微博数、用户影响力越大,则是谣言的可能性越小。谣言识别模型的准确率达到0.984,其中,文本语义特征的重要性最高。

**关键词:** 谣言识别 突发公共卫生事件 神经网络模型 XGBoost

## Research on Rumor Identification in Public Health Emergency

Sun Ran<sup>1</sup> An Lu<sup>1,2</sup>

(<sup>1</sup>School of Information Management, Wuhan University, Hubei, 430072;

<sup>2</sup>Center for Studies of Information Resources, Wuhan University, Hubei, 430072)

**Abstract:** [Purpose/significance] This study aims to explore the significance of factors on the rumor identification in the social media environment and identify rumors in public health emergencies. We also evaluated the important features that affect the identification of rumors to help the cyber security department accurately identify rumors and maintain a healthy network ecological environment. [Method/process] We extracted user features, time features, structure features, text semantic features and propagation features in microblog entries. We combined with the MAIN theoretical models, and used binary logistic regression method to deeply research the influence factors of rumors from the perspective of modality, information content, information sources. We built a multi-feature based rumor identification model that integrated the semantic feature extracted by neural network model. XGBoost algorithm was used to calculate the importance of different features in rumor identification. [Result/conclusion] The higher the positive emotional value of comment, the number of microblog entries posted by users, and greater the influence of users, the lower the possibility that the microblog entry is a rumor. The value of the accuracy of rumor recognition model is 0.984. The semantic features of text are the most important.

**Keywords:** rumor identification public health emergency neural network model XGBoost

\* 本文系教育部哲学社会科学研究重大课题攻关项目“提高反恐怖主义情报信息工作能力对策研究”(项目编号:17JZD034)、国家自然科学基金面上项目“危机情境下网络信息传播失序识别与干预方法研究”(批准号:72174153)、国家自然科学基金重大课题“国家安全大数据综合信息集成与分析方法”(批准号:71790612)和国家自然科学基金创新研究群体项目“信息资源管理”(批准号:71921002)的研究成果之一。

## 1 引言

伴随着重大突发公共卫生事件的爆发,网络平台上谣言四起,可能会引发大众产生恐慌和负面情绪,如何有效识别重大突发公共卫生事件中的谣言是抑制谣言传播、降低谣言危害的前提。尤其是突发公共卫生事件、恐怖事件等特定情境下,网络谣言的传播会对事件处置、疫情防控、经济与社会发展等方面造成极大的破坏。根据《2019年网络谣言治理报告》显示,医疗健康、食品安全、社会科学三类是网络谣言的高发区,而在2019年微信平台共生产17881篇辟谣文章,辟谣文章阅读量达到1.14亿万次<sup>[1]</sup>。如何运用大数据技术治理网络谣言、提升疫情的应急处置能力,已经成为各级政府和相关部门急需解决的任务。由此本文拟解决三个问题:(1)谣言和非谣言信息在传播效果方面是否存在显著差异?(2)社交媒体环境下各因素对识别谣言的影响作用是否显著?(3)如何从微博用户自生成数据中自动抽取谣言信息的若干关键特征,基于深度学习和机器学习模型构建多特征融合的谣言预测模型,协助有关部门快速识别谣言?

本文选取突发公共卫生事件中的网络谣言为研究对象,探究权威线索、从众线索、模态线索等影响因素跟谣言识别之间的关系,并结合神经网络模型和机器学习算法探究微博用户特征、微博文本特征、微博传播特征、时间特征在谣言识别问题上的表现,构建融合文本语义特征的谣言识别模型。对尽早识别网络谣言、有效抑制其传播、维护健康的网络生态环境具有重要的现实意义。

## 2 相关研究

### 2.1 MAIN模型

由Sundar提出的Modality-Agency-Interactivity-Navigability (MAIN)模型阐述了与模态、代理、交互和导航相关的四种承载力,可以用来进行社交媒体信息可信性评估<sup>[2]</sup>。该模型的基本内涵是:嵌入在四种承载力中的某些线索触发的特定启发式会影响用户对信息质量的评价,进而影响用户对信息可信度的判断。其中一个例子就是模态线索(如文本、听觉、视听)。当年轻用户不确定信息的真实性时,他们可能会依赖基于模态的启发式,同时超链接上的文字本身也可能触发不同的启发式。但现在一些社交媒体用户可能会使用虚

假或者与事实描述不符的图片、视频或外部链接来帮助证明其内容的真实性,从而误导其他用户。

代理承载力可以触发权威启发式,或者将资源归因到更大的用户群中并触发潮流启发式。例如,有实验证明用户更倾向于认为专业信息源发布的内容比一个门外汉发布的信息内容更可信<sup>[4]</sup>。MAIN模型理论认为被大多数人所认可的新闻内容会触发潮流启发式,例如,被其他用户所称赞的信息内容将会是可信的<sup>[5]</sup>。与正面评价相比,负面评价往往导致用户对信息可信度的评价较低<sup>[6]</sup>。通过对评论情感倾向的建模,祖坤琳等<sup>[7]</sup>对谣言分类模型进行了改进。权威、身份和从众线索在一定程度上影响信息源的可信度感知,尤其是权威线索对信息源可信度感知的影响最为强烈<sup>[8]</sup>。信息来源的权威性、微博的评论、点赞和转发数通过触发用户的从众启发式思考,对判定微博内容的可信度和传播效果有一定的补充作用,同时挖掘微博文本内容(如文本语义特征)的深层特征来分析文本内容的可信度。

### 2.2 网络谣言识别

早期研究多从社会学和心理学角度对谣言进行定义:“谣言是在模棱两可、存在危险或潜在威胁的情况下出现的未经证实却广泛流传的信息,能帮助人们了解模糊的情况并适应可能会出现危机”<sup>[9]</sup>、“谣言是一种特定信仰或话题的主张,通常通过口口相传,没有可靠的依据,具有重要性和模糊性”<sup>[10]</sup>。社交媒体的流行行为研究网络谣言产生和传播提供了数据来源,贺刚等<sup>[11]</sup>将微博谣言定义为在微博社区环境下,以各种渠道传播能引起用户兴趣的事物的未经证实的诠释。本文在上述谣言定义的基础上将事实依据,并且未经官方证实的情况下能引起广泛流传的信息定义为谣言。由于谣言的内容、发布用户以及它的传播有别于非谣言,传统的方法多采用SVM<sup>[12]</sup>、决策树<sup>[13]</sup>、随机森林<sup>[14]</sup>等分类器来分类谣言和非谣言,谣言识别研究多围绕三个方面进行展开:(1)谣言用户特征研究,信源用户粉丝数量太多或者太少都会导致信息的不可信<sup>[15]</sup>,同时用户的行为特征是识别谣言的隐藏线索<sup>[16]</sup>;(2)谣言内容特征研究,微博的主题分布特征、情感特征能在谣言识别中发挥重要作用<sup>[17]</sup>;(3)谣言传播特征研究,包括谣言传播路径和谣言的评论内容、评论用户等。较多的学者基于微博转发、评论特征,同时,评论的多维信息有助于判断微博内容的关注度和可信度。Waddell等<sup>[6]</sup>通过在线实验测试用户评论对新闻可信度的影响,发现负面

评论相对于正面评论会降低信息的可信度。Castillo等<sup>[18]</sup>通过提取推特中热点话题的消息特征、用户特征、主题特征和传播特征,构建J48决策树模型来预测推特的可信度。近年来则开始采用基于神经网络模型的方法来预测谣言<sup>[19]</sup>。

综上所述可以看出,大多数学者采用了基于微博内容特征、用户特征和传播特征进行分析,但是传播特征主要采用微博的转发、点赞特征,而未对评论的内容特征和情感特征进行深入分析,而这些信息能够辅助谣言的识别。另外,在方法的选择上,多数谣言预测研究选择一种机器学习或者神经网络的方法进行研究,尝试融合多种机器学习/深度学习方法的研究相对较少。由此可见,相关研究往往忽视事件内容本身的特征,这也反映出谣言识别研究侧重单一方向,尝试融合多种方法的研究相对较少。

### 3 研究方法

本研究拟结合神经网络模型和极端梯度提升(XGBoost)模型将微博用户特征、微博文本特征等深度融合在一起,即将微博文本输入Bert模型进行向量表示,再输入RNN/CNN等神经网络算法中得到文本属于谣言的概率,将其作为一个特征值和其他各个离散特征作为XGBoost分类器的输入,同时选择多种基线模型对预测结果的准确率指标进行比较。同时,本文基于MAIN理论模型,将特征体系中的用户影响力特征、文本结构特征、传播特征等作为影响识别谣言的权威线索、模态线索、从众线索,探索社交媒体环境下各因素对识别谣言的影响作用是否显著。研究框架如图1所示。

#### 3.1 特征构建

谣言识别特征体系包含微博用户特征和微博文本特征、微博传播特征等,如表1所示。

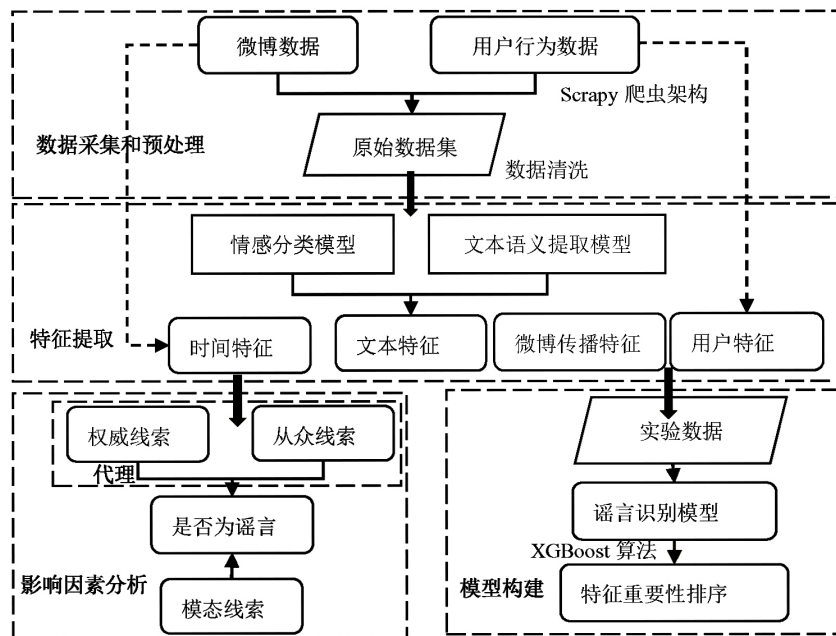


图1 突发公共卫生事件中谣言识别模型构建流程

表1 谣言识别的特征体系及特征值

特征描述			特征值
用户特征	用户基本属性	是否认证	是否
		用户所在地区	北京/上海/广州……
		会员	是否充值会员
		用户性别	男女
	用户行为特征	用户影响力	粉丝数与(关注数+粉丝数+1)比值
		用户发布微博总数	微博总数
		创作积极性	发布微博日均值
文本特征	内容特征	互动积极性	转发/评论数平均数
		情感倾向特征	积极/中性/消极
		话题类别	医疗类、健康防疫类、政府管控类、社会类、人物类、其他类
	结构特征	语义特征	神经网络模型判断微博为谣言的概率
微博传播特征	传播影响	url/图片/视频/艾特/表情/hashtag	分别判断微博文本中是否包含这些特征
		评论质疑度	针对每条原创微博,先判断单条评论包含的质疑词个数,然后将点赞数作为权重,计算总评论的质疑度
	评论内容特征	评论情感度	针对每条原创微博,先对单条评论进行情感倾向识别,然后将点赞数作为权重,计算总评论的情感度
		评论情感度	针对每条原创微博,先对单条评论进行情感倾向识别,然后将点赞数作为权重,计算总评论的情感度
时间特征	时间跨度特征	微博与话题源头发布的时间差	同一话题中,微博发布的时间和最先发布该话题微博的时间差
	所在时间段	深夜/清晨/上午/中午/下午/晚上	是/否
	是否为节假日	法定节假日以及周末	是/否
	星期	周一-二/三/四/五/六/日	是/否



### 3.1.1 用户特征

(1)用户基本属性。用户的基本属性包含是否认证、会员、所在地区、用户性别、用户影响力、用户发布微博总数六个维度,可以通过爬取的微博字段获取;其中会员分为是否开通了微博会员,所在地区通过微博用户资料中的地理信息获取,最终选择的特征值包括中国34个省级行政区,以及“海外”“其他”共36个特征值。由于用户的粉丝数和关注数可能会影响微博流行度<sup>[20]</sup>,本文用粉丝数和关注数来表示用户影响力  $U_{infu}$ <sup>[3]</sup>,其计算公式如公式(1)所示:

$$U_{infu} = \frac{n_{fans}}{n_{fans} + n_{followers} + 1} \quad (1)$$

其中,  $n_{fans}$  和  $n_{followers}$  分别为用户的粉丝数和关注数。

(2)用户行为特征。用户行为特征包括事件近期创作积极性和近期互动度两个二级指标,其中近期创作积极性是指用户在突发事件发生前30天截止到数据采集的后30天内(2019年12月1日-2020年4月31)发布或者转发与事件相关的微博条目数;近期互动度在此定义为在事件发生前30天截止到数据采集的后30天内用户发布的所有微博的互动量,包括微博的转发、评论、点赞数。一般来说,网络水军为了散播谣言,会大量转发、评论、回复他人<sup>[21]</sup>,因此本文限定时间区间为谣言数据采集前后30天主要是为了判断用户在短期内的热度,如果热度过高,则其发布谣言的倾向更高。

### 3.1.2 文本特征

文本特征包括文本结构特征、语义特征和情感倾向特征:(1)情感倾向特征,本文拟采用百度AI开放平台的情感倾向分析功能得到原创微博的情感倾向。(2)话题类别特征,新型冠状病毒疫情事件中谣言的主题可以分为医疗类、生活类、教育类、地点类、社会类<sup>[22]</sup>,再根据实验数据中微博的话题类别情况,本文将微博的话题类别分为医疗类、健康防疫类、政府管控类、社会类、人物类、其他类,其中健康防疫类微博多与伪医学知识有关,医疗类微博则主要是与感染病例、疫情发展等。(3)文本语义特征,大多数研究只对文本内容所包含的主题词或情感词等统计数据作为特征,没有深入挖掘内容中包含的语义特征,Bert可以提取文本语义特征。本文拟采用Bert训练微博文本得到向量矩阵作为CNN/RNN卷积层的输入,将谣言预测的概率作为一个特征值。(4)文本结构特征,用户在发布微博时,通常会在文本内容中加入各种符号加强语义的表达,如

hashtag、url、图片、视频来传递更多的信息<sup>[23]</sup>,因此本研究将是否有链接、哈希标签、图片、视频、提及(@)、表情纳入微博文本结构特征中。

### 3.1.3 微博传播特征

在谣言的传播过程中,大多采用微博的转发数、评论数和点赞数作为微博流行度的表征,而观察谣言微博数据可以发现,谣言微博的评论中会出现对抗性声音,比如谣言/造谣/传谣/真的吗?/假的吧?/不会吧?/抵制谣言/传谣/不信谣等质疑词,这些评论信息能够辅助判断微博是否为谣言。通过对微博的评论内容识别其质疑度和情感倾向能对原始微博进行“打分”。因此本文拟考虑微博评论内容特征和微博传播影响特征,微博评论内容特征包括评论质疑度和评论情感度,评论质疑度  $C_{doubts}$  计算公式如公式2所示。

$$C_{doubts} = \frac{\sum_i^N n_{i\_likes} * n_{i\_words}}{N_{comments}} \quad (2)$$

其中,  $n_{i\_likes}$  是每条原创微博中第  $i$  条评论的点赞数,  $n_{i\_likes}$  是每条原创微博中第  $i$  条评论的质疑词个数;  $N_{comments}$  每条原创微博的评论总数,  $N$  是每条原创微博的总评论数。评论情感度  $S_{score}$  则首先计算每条微博评论的情感倾向,再对所有评论的情感倾向进行累加,如公式3所示。其中,  $S_{i\_comments}$  为每条原创微博中第  $i$  条评论的情感倾向。

$$S_{score} = \frac{\sum_i^N n_{i\_likes} * S_{i\_comments}}{N_{comments}} \quad (3)$$

### 3.1.4 时间特征

当微博发布时间的不同时,能接收到信息的用户数量也不同,因此本研究将微博发布时间段、星期、节假日考虑在内,微博发布所在时间段根据微博用户作息规律将全天划分为深夜(00:01-6:00)、清晨(6:01-8:30)、上午(8:31-12:00)、中午(12:01-14:00)、下午(14:01-18:00)、晚上(18:01-24:00)六个阶段<sup>[24]</sup>。节假日包括法定节假日和周末。在同一谣言事件中,一般先发布的微博容易引起较大的关注,而且在辟谣的过程中,大众对谣言的关注度会逐渐倾斜。因此本文引入时间差特征,将其定义为在一个谣言事件中,微博与最先发布谣言的微博之间的时间差  $T$ ,其计算公式如公式4所示,其中,  $T_{weibo}$  和  $T_{(first\_weibo)}$  分别为同一谣言事件中,微博发布的时间和最先发布谣言微博的时间。

$$T = T_{weibo} - T_{first\_weibo} \quad (4)$$



### 3.2 二元逻辑回归模型

本文采用的二元逻辑回归模型可以用来描述当各自变量变化时,因变量的发生概率会发生怎样变化。因变量 $Y$ 是指用户发布的微博是否为谣言,为二分类变量(取值为0或1),比值( $Y=1$ 与 $Y=0$ 的概率之比)可以通过公式(5)中的 $logit$ 形式表达出来:

$$\ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad (5)$$

其中, $\beta_0$ 是常数项,表示自变量取值全部为0时,比值的自然对数值;本文主要研究模态线索、从众线索、权威线索等对识别谣言的影响作用是否显著,因此 $x_i$ 表示选择的部分变量:用户影响力、评论质疑度和情感度等;自变量的回归系数 $\beta_i(i=1, \dots, n)$ 表示自变量 $x_i$ 每改变一个单位,比值比(OR)的自然对数值改变量, $e^\beta$ 表示自变量 $x_i$ 每变化一个单位,阳性结果出现概率与不出现概率的比值是变化前的相应比值的倍数。目前均采用最大似然法来解决方程的估计和检验问题。

### 3.3 XGBoost 模型

本研究拟融合神经网络模型的判断结果和离散特征作为XGBoost分类器的输入,同时为了验证模型的先进性,可以选择多种基线对比算法对预测结果的准确率、召回率和F1值进行比较。同时对特征进行不同组合评估模型效果,能够展示谣言预测中哪些特征最为重要。特征重要性是通过将数据集中的每个属性进行计算并排序而得出,其原理是一次随机从数据集去掉数据的某一个特征,计算其性能指标的下降程度,变化越大,则代表特征就越重要。

本研究采用XGBoost算法进行特征重要性排序,在此将谣言预测问题看成一个二分类问题,若某条微博为谣言,则设置为1,否则设置为0。XGBoost的输入是特征向量的组合,其基学习器是CART回归树。对于多维特征向量 $x_i$ ,则XGBoost的输出 $\hat{y}_i$ 如公式(6)所示。

$$\hat{y}_i = \sum_{m=1}^M f_m(x_i), f_m \in F \quad (6)$$

其中, $M$ 是CART树的棵数,即为XGBoost模型中输入的 $n\_estimators$ 参数。 $F$ 表示所有可能的CART树, $f_m(x_i)$ 表示CART树 $m$ 的分类结果。

## 4 实验过程

以微博平台为例,本研究以“微博辟谣”(https://weibo.com/weibopiyao)和“社区管理平台”(https://service.account.weibo.com/)上的谣言话题为研究对象。本文采

集的微博谣言包含28个谣言话题,如罗玉凤感染新冠肺炎、被感染隔离的宝宝向医生求抱抱、北京进行大面积消毒、女子未戴口罩遭阻拦转身跳河、格力复工发现疫情、顺丰快递员拦截包裹卖口罩、11万人将从欧洲进京等。根据不同的谣言事件通过具体的检索词对谣言和相应的辟谣微博的内容、发布者信息、传播用户信息以及评论内容进行抓取。目前已爬取谣言微博2995条,并将谣言事件中的辟谣微博(2689条)当作非谣言数据集。其中原创微博的转发评论有近9万条,并爬取了发布原创微博的5259个用户的基本信息,以及这些用户近期(2020年1月1日-3月31日)发布的微博共454万余条。

在疫情初期,有关新冠肺炎病毒传播途径、穿毛衣更容易吸附病毒等谣言较多;随着国内疫情越来越严重,医用口罩逐渐成为生活必需品并且供不应求,有关如何鉴别口罩、口罩的佩戴方式等谣言层出不穷,特别是顺丰快递员拦截包裹卖口罩事件在社交媒体平台上引发热议,后续经过证实是寄件人捏造。另外,还有诸如北京全市大面积消杀、瑞德西韦能治疗新冠肺炎等谣言出现。此外,诸如汤姆克鲁斯去世、屠呦呦落选院士、西安大明宫万达杀人截肢等与疫情无关的谣言也引发了广泛传播。在国内疫情逐渐稳定后,境外输入病例成为大众的关注热点,11万人从欧洲到北京、澳籍华人女子返京后拒绝隔离等谣言的出现进一步引发大众的恐慌和愤怒。

### 4.1 微博传播特征分析

为研究谣言微博和辟谣微博在传播效果上的差异性,本文将谣言微博和辟谣微博的转发、评论、点赞数输入到SPSS软件中进行T检验。其中,谣言微博的转发(5.13)、评论(10.28)、点赞数(32)的均值均低于辟谣微博(18.19、21.95、194.09),尤其是点赞数和转发数,这可能是因为辟谣微博一经官方发布,大众会自发地进行点赞、转发,从而避免其他用户被谣言所误导。在传播效果的差异性上看,谣言微博和辟谣微博在转发上的差异有统计学意义( $p=0.009<0.05$ ),但是在点赞( $p=0.61>0.05$ )和评论( $p=0.116>0.05$ )上的差异性并不显著。

### 4.2 谣言识别影响因素分析

本文采用二元逻辑回归方法对谣言识别的影响因素进行分析,并采用基于最大似然估计的向前逐步回归法进行自变量筛选,实验结果如表2所示,表中的变

量均呈现出0.05水平的显著性( $p<0.05$ )。采用ROC曲线对回归模型拟合效果进行判断,ROC曲线下的面积为0.876,说明模型拟合效果较好。经过筛选后不在方程中(对因变量影响不显著)的变量有评论质疑度、图片、提及、点赞数、转发数、评论数、创作积极性、性别。

对于是否为会员、是否认证等二分类变量,OR值的含义为赋值较高的微博是谣言的风险为赋值较低的微博是谣言的风险的多少倍,例如在MAIN理论的模态线索(文本结构特征)中,包含视频、表情的微博比不包含这些因素的微博是谣言的风险分别为1.82、1.20倍,而包含链接、hashtag的微博比不包含这些因素的微博是谣言的风险分别为0.43、0.22倍,也就是风险更低。对于情感值、话题类别等多分类变量,需要指定某一分类为参照,从而将其变换成哑变量。以话题类别特征为例,医疗类微博相对于健康防疫类微博是谣言的风险会增加( $OR=2.21$ , 95% CI: 1.51–3.25),而社会类微博是谣言的风险最大。对于连续型变量可直接观察其回归系数值来判断对因变量的影响,例如评论情感度、用户发布微博数、用户影响力的回归系数值B分别为-0.45、-0.36、-0.56,并且均呈现出0.01水平的显著性( $p<0.01$ ),这意味着这些变量会对微博是否为谣言产生显著的负向影响关系。

表2 二元逻辑回归模型检验结果

自变量	B	Sig.	OR(95%CI)	自变量	B	Sig.	OR(95%CI)
情感值				时间段			
负		0	1.00	深夜		0	1.00
中性	0.67	0	1.95(1.34–2.82)	清晨	-0.41	0.06	0.66(0.43–1.01)
正	0.79	0	2.20(1.87–2.60)	上午	-0.08	0.61	0.92(0.68–1.26)
话题类别				中午	-1.05	0	0.35(0.25–0.49)
健康防疫类		0	1.00	下午	-1.65	0	0.19(0.14–0.26)
医疗类	0.80	0	2.21(1.51–3.25)	晚上	-0.06	0.69	0.94(0.71–1.26)
社会类	1.50	0	4.49(3.20–6.28)	星期			
政府管控类	0.16	0.38	1.17(0.82–1.68)	星期一		0	1.00
人物类	-0.03	0.88	0.97(0.65–1.45)	星期二	0.24	0.08	1.28(0.97–1.67)
其他类	1.12	0	3.06(2.18–4.30)	星期三	0.89	0	2.30(1.71–3.07)
评论情感度	-0.45	0	0.64(0.54–0.75)	星期四	-1.14	0	0.32(0.24–0.43)
url	-0.85	0	0.43(0.35–0.52)	星期五	-0.02	0.91	0.98(0.69–1.40)
视频	0.6	0	1.82(1.30–2.56)	星期六	-0.05	0.86	0.96(0.57–1.59)
表情	0.19	0.04	1.20(1.01–1.43)	星期日	0.89	0	2.43(1.50–4.07)
hashtag	-1.52	0	0.22(0.18–0.26)	发微博总数	-0.36	0	0.70(0.63–0.78)
节假日	0.73	0	2.08(1.51–2.85)	用户影响力	-0.56	0	0.57(0.42–0.77)
是否会员	0.24	0	1.27(1.08–1.50)	是否认证	-0.69	0	0.50(0.41–0.61)

在权威线索中,用户发微博总数和用户影响力的OR值分别为0.70、0.57,意味着用户发微博总数和用户影响力每增加一个单位,Y(用户发布的微博是谣言的概率)的减少幅度分别为0.70、0.57倍,认证用户发布谣言的概率比未认证的用户发布谣言的概率更低,但是会员用户发布谣言的概率相对于非会员的概率则更高,可能是微博认证系统对用户的要求更加严格。在从众线索中,评论质疑度对谣言的概率没有显著影响,但是评论情感度越偏向正,则是谣言的概率越低。在时间特征中,深夜发布的微博是谣言的概率最高,清晨、上午、中午、下午、晚上发布的微博为谣言的概率分别是深夜发布微博为谣言概率的0.66、0.92、0.35、0.19、0.94倍。星期日发布的微博是谣言的概率相比其他时间发布的微博是谣言的概率高。

4.3 谣言预测模型

本文分别构建随机森林、决策树、XGBoost、神经网络模型(CNN/RNN)等预测模型,采用python中sklearn包中的train\_test\_split()函数随机划分训练集和测试集,并用准确率、召回率和F1值来评估模型的分类效果。各模型分类结果如表3所示,经过对模型对比后发现,添加了文本语义特征的XGBoost模型在准确率、召回率、F1值均最高。随后再利用XGBoost算法进行谣言预测中特征重要性排序,结果如下页图2所示。

表3 模型评估结果

模型	准确率	召回率	F1值
随机森林	0.844	0.901	0.871
决策树	0.847	0.861	0.854
XGBoost	0.899	0.924	0.911
Bert_CNN	0.951	0.926	0.938
Bert_RNN	0.959	0.930	0.945
Bert_CNN+XGBoost	0.984	0.987	0.986
Bert_RNN+XGBoost	0.947	0.972	0.960

从实验结果上看,经过Bert和CNN模型得到的文本语义特征在谣言识别中的重要性最高(0.446),其次是用户特征(0.176)和时间特征(0.174)。将重要性为0的特征去掉后得到的特征重要性排序如下页图3所示,除了语义特征外,用户是否认证在谣言识别中的重要性最高(0.055)。在用户特征中,用户所在地为上海(0.012)和香港(0.009)的特征重要性较高,其次是用户创作积极性、用户影响力等特征;在文本结构特征中,hashtag的特征重要性最高(0.047),其次是提及(0.009);在

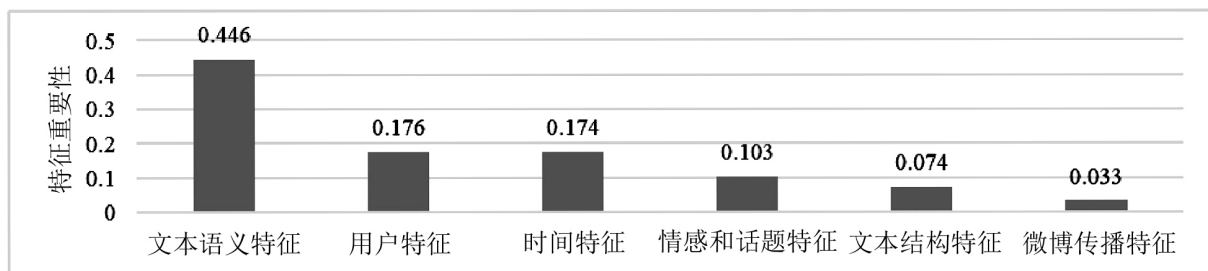


图2 特征重要性排序

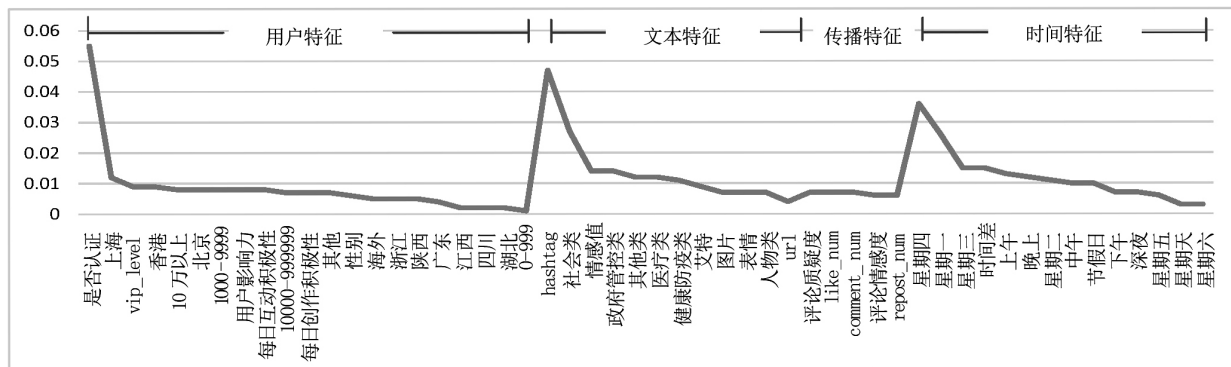


图3 特征重要性排序(除文本语义特征外且重要性不为0)

时间特征中,星期四(0.036)特征重要性最高,其次是星期一(0.026)和时间差特征(0.015);传播特征的重要性均较低,在内容特征中,社会类、微博文本的情感值等特征的重要性较高。虽然在影响因素分析中评论质疑度、图片、提及、用户创作积极性等对判断是否为谣言并不显著,但是在谣言识别中具有一定的重要性。

## 5 结语

本文针对谣言识别问题,分析社交媒体环境下哪些因素在识别谣言方面具有显著影响力,发现评论情感度、用户影响力、用户发布微博数等变量越高,则微博是谣言的风险越小,相比其他时间段,在星期日以及深夜发布的微博是谣言的风险更大。同时本文从用户基本属性特征、用户行为特征、文本特征、微博传播特征、时间特征等方面,采用XGBoost算法评估特征对谣言识别的重要性。实验发现文本语义特征的重要性最高。随后,本文探讨了辟谣微博和谣言微博的传播特征,发现谣言微博的转发、评论、点赞数的均值均低于辟谣微博,并且在转发数上具有显著差异。谣言识别

的实验结果表明,模型准确度达到了0.984,高于其他基线模型,能够较好地识别谣言。本研究的不足之处在于:本文选择的谣言数据和辟谣数据较为均衡,但是现实环境中,谣言微博和普通微博存在较大的不平衡性,这可能导致有些特征在数据不均衡的情况下效果变差的问题。

致谢:感谢图书情报国家级实验教学示范中心为本研究提供的实验支持!

## 参考文献

- [1] 人民网. 2019年网络谣言治理报告[EB/OL]. (2019-12-26) [2020-04-20]. <http://society.people.com.cn/n1/2019/1226/c1008-31524533.html>.
- [2] Sundar S S. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility[M]//Miriam J Metzger, Andrew J Flanagin. Digital Media Youth, and Credibility. Cambridge, MA: The MIT Press, 2008: 73-100. doi: 10.1162/dmal.9780262562324.073.
- [3] 张仰森, 彭媛媛, 段宇翔, 等. 基于评论异常度的新浪微博谣言识别方法[J]. 自动化学报, 2020, 46(8): 1689-1702.
- [4] Lee J Y, Sundar S S. To tweet or to retweet? That is the question



- for health professionals on Twitter[J]. Health Communication, 2013, 28(5): 509-524.
- [5] Waddell T F. What does the crowd think? How online comments and popularity metrics affect news credibility and issue importance[J]. New Media & Society, 2018, 20(8): 3068-3083.
- [6] Waddell T F, Sundar S S. # thisshowsucks! The overpowering influence of negative social media comments on television viewers[J]. Journal of Broadcasting & Electronic Media, 2017, 61(2): 393-409.
- [7] 祖坤琳, 赵铭伟, 郭 凯, 等. 新浪微博谣言检测研究[J]. 中文信息学报, 2017, 31(3): 198-204.
- [8] Lin X, Spence P R, Lachlan K A. Social media and credibility indicators: the effect of influence cues[J]. Computers in Human Behavior, 2016, 63: 264-271.
- [9] DiFonzo N, Bordia P. Rumor Psychology: Social and Organizational Approaches[M]. Washington, DC: American Psychological Association, 2007.
- [10] Allport G W, Postman L. The Psychology of Rumor[M]. Oxford, UK: Henry Holt, 1947.
- [11] 贺 刚, 吕学强, 李 卓, 等. 微博谣言识别研究[J]. 图书情报工作, 2013, 57(23): 114-120.
- [12] Lukasik M, Cohn T, Bontcheva K. Classifying tweet level judgments of rumours in social media[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Sheffield: The Association for Computational Linguistics, 2015: 2590-2595.
- [13] Ma J, Gao W, Wong K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017: 708-717.
- [14] 曾子明, 王 婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例[J]. 情报学报, 2019, 38(1): 89-96.
- [15] Westerman D, Spence P R, Van Der Heide B. A social network as information: the effect of system generated reports of connectedness on credibility on Twitter[J]. Computers in Human Behavior, 2012, 28(1): 199-206.
- [16] Liang G, He W, Xu C, et al. Rumor identification in microblogging systems based on users' behavior[J]. IEEE Transaction on Computational Social System, 2015, 2(3): 99-108.
- [17] 首欢容, 邓淑卿, 徐 健. 基于情感分析的网络谣言识别方法[J]. 数据分析与知识发现, 2017, 1(7): 44-51.
- [18] Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]. International Conference on World Wide Web. New York, NY: ACM, 2011, 675-684.
- [19] 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, 2018: 40-52.
- [20] Yu D, Chen N, Ran X. Computational modeling of Weibo user influence based on information interactive network[J]. Online Information Review, 2016, 40(7): 867-881.
- [21] 仲丽君, 杨文忠, 袁婷婷, 等. 社交网络异常用户识别技术综述 [J]. 计算机工程与应用, 2018, 54(16): 13-23.
- [22] 姚艾昕, 马 捷, 林 英, 等. 重大突发公共卫生事件谣言演化与治理策略研究[J]. 情报科学, 2020, 38(7): 22-29.
- [23] Hamidian S, Diab M T. Rumor detection and classification for twitter data[J]. arXiv preprint arXiv:1912.08926, 2019.
- [24] 安 璐, 易兴悦, 孙 冉. 恐怖事件情境下微博影响力的预测及演化[J]. 图书情报知识, 2019(4): 52-61, 81.
- [作者简介] 孙 冉, 女, 1997 年生, 武汉大学信息管理学院博士研究生。
- 安 璐, 女, 1979 年生, 武汉大学信息资源研究中心数据管理与知识服务研究室主任, 武汉大学信息管理学院教授, 博士生导师。
- 收稿日期: 2020-12-20

## 欢迎订阅

### 2022年《情报资料工作》杂志

- 中国社会科学情报学会学报
- CSSCI 来源期刊
- 全国中文核心期刊
- 中国社会科学院 AMI 核心期刊
- “复印报刊资料”重要转载来源期刊
- 邮发代号 82-22 全年定价 288 元