

基于情感分析的突发公共卫生事件微博评论用户画像构建^{*}

张彩红

李 想 周雪侠 潘 玮

(蚌埠医学院公共卫生学院 蚌埠 233030)

(蚌埠医学院卫生管理学院 蚌埠 233030)

〔摘要〕 介绍基于情感分析的突发公共卫生事件微博评论用户画像构建技术路线, 将计算得到的情感倾向度与其用户特征进行相关分析, 提出可选取与用户评论情感倾向度具有相关性的性别、年龄、粉丝数、点赞数、简介、标签 6 个用户特征为内容构建突发公共卫生事件微博评论用户画像。

〔关键词〕 网络舆情; 情感分析; 用户画像; 突发公共卫生事件

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.03.003

User Portrait Construction of Weibo Comments on Public Health Emergencies Based on Sentiment Analysis ZHANG Caihong, School of Public Health, Bengbu Medical College, Bengbu 233030, China; LI Xiang, ZHOU Xuexia, PAN Wei, School of Health Management, Bengbu Medical College, Bengbu 233030, China

〔Abstract〕 The paper introduces the technical route of constructing user portraits for Weibo comments on public health emergencies based on sentiment analysis, analyzes the correlation between the calculated sentiment orientation and user characteristics, and proposes that six user characteristics, such as gender, age, number of followers, number of likes, introduction and label, which are related to the emotional tendency of user comments, could be selected as content to construct user portraits of Weibo comments on public health emergencies.

〔Keywords〕 network public opinion; sentiment analysis; user portrait; public health emergency

1 引言

突发公共卫生事件发生后, 相关舆情高度集中。目前, 网络社交媒体已成为网络舆情传播的重要平台^[1]。充分挖掘微博用户针对突发公共卫生事件话题相关评论的主观信息, 识别其情感倾向及用

户特征, 能够为舆情管理部门开展舆情监控和引导工作提供决策支持和信息参考, 对消除社会恐慌具有重要意义。

情感倾向反映网络用户对网络事件的态度, 影响网络舆情的走向^[2]。情感倾向可基于情绪知识^[3]、传统机器学习^[4]和词向量^[5]等多种方法进行判定。网络用户对于突发公共卫生事件的评论, 其

〔修回日期〕 2022-11-14

〔作者简介〕 张彩红, 硕士研究生; 通信作者: 潘玮, 博士, 副教授, 发表论文 12 篇。

〔基金项目〕 安徽高校人文社会科学重点研究项目“基于多源数据融合的医药领域技术创新机会识别机制研究”(项目编号: SK2020A0355); 安徽省大学生创新创业计划项目“基于情感分析的突发公共卫生事件网络用户画像研究”(项目编号: S202110367051); 蚌埠医学院 512 人才项目“优秀青年教师”(项目编号: by51201311)。

情感倾向往往表现出一定集群性，即性别、年龄、地区等特征相似的网络用户有着较为接近的情感倾向^[6]。在突发公共卫生事件发生过程中提前识别出具有特定情感倾向的微博评论用户特征，进而制定具有针对性的舆情引导策略，对防控舆情风险具有重要作用。

用户画像是通过数据挖掘技术，从海量数据中获取的描述用户的标签，广泛运用于推荐系统和产品运营，近年来被逐渐引入医疗健康领域^[7]。但将用户画像应用于突发公共卫生事件网络舆情治理的研究较少。因此，本研究运用网络爬虫技术获取突发公共卫生事件相关微博评论数据及评论用户特征数据，对评论数据进行情感分析，计算其情感倾向度；将计算得到的情感倾向度与用户特征进行相关分析，探索与其存在相关关系的用户特征集合，并将评论用户的情感倾向及与其存在相关关系的用户特征共同构成用户画像，用于提升网络舆情监控和引导的精确性，见图 1。

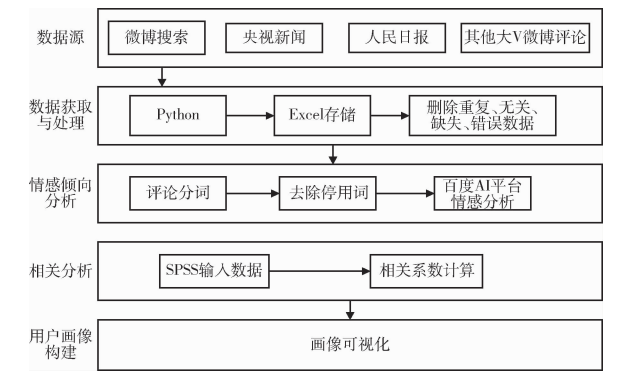


图 1 用户画像构建技术路线

2 研究方法与数据来源

2.1 研究方法

2.1.1 网络爬虫 网络爬虫能够批量检索和下载数据，对大数据环境下的网络信息采集及分析发挥着巨大的推动作用。其获取网络信息的基本步骤为：根据用户需要确定初始网页地址（uniform resource locator，URL）；请求网页数据，获取响应内容并对其进行解析；存储到数据库或保存为特定格式的文件。本文利用网络爬虫技术获

取研究数据的步骤如下。（1）采用 Python 程序设计语言编写爬虫代码，调用 requests 库对新浪微博网页端 URL 发起请求。（2）获得响应后，将请求的数据信息返回 json 格式，使用 json.loads（）方法解析数据，提取所需评论数据及评论用户个人主页信息，包括性别、年龄、粉丝数、评论点赞数、简介、标签、关注数以及昵称长度等用户特征数据。（3）将爬取到的数据保存到 Excel 文件中。

2.1.2 文本处理 微博用户评论数据具有语言不规范、半结构化的特点，在对其进行分析之前需经过分词处理。Jieba 分词是当前效果较好的中文分词器，主要支持精确模式、全模式和搜索引擎模式 3 种分词模式，同时还支持繁体分词及自定义词典。本文利用 Python 中的 Jieba 分词工具对新浪微博用户评论数据进行分词，依据哈尔滨工业大学停用词表去除停用词，并将处理好的评论文本命名为 text。

2.1.3 情感倾向分析 情感倾向分析指对包含主观信息的文本进行情感处理、分析、归纳和推理的过程^[8]。百度人工智能（artificial intelligence，AI）开放平台是国内较为常用的情感倾向分析人工智能平台，以应用程序接口（application programming interface，API）或开发库工具（software development kit，SDK）等形式对外共享其在语音、图像、自然语言处理（natural language processing，NLP）等领域积累的技术^[9]。用户可以直接调用这些 API 对中文文本进行情感分类，准确度较高。本文应用百度 AI 开放平台，选择创建应用，获取 API KEY 及 Secret KEY。在 Python 3.9 环境下，输入 API KEY、Secret KEY，利用 client.sentimentClassify（text）命令，对命名为 text 的评论文本进行情感倾向分析，对所有评论文本的情感倾向进行分类。

2.1.4 相关分析 相关分析是研究两个或两个以上处于同等地位随机变量间相关关系的统计分析方法。通常采用相关系数分析变量之间的相互关联性，相关系数定义为：

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

其中， r 为相关系数， X_i 和 Y_i 分别为两个分析变量的数值， r 的取值范围在 $(-1, +1)$ 之间。 r 的绝对值越大，变量间的相关程度越大。

运用 SPSS 25.0 软件对新浪微博用户评论数据情感倾向度与用户特征进行相关分析。在进行相关分析之前，需对用户特征文本数据进行编码，见表 1。

表 1 用户特征文本数据编码情况

序号	用户特征	文本	编码
1	性别	男	0
		女	1
2	昵称	昵称名	昵称名长度
		(如: abcde)	(如: “abcde” 的编码为 “5”)
3	简介	有简介	1
		无简介	0
4	标签	有标签	1
		无标签	0
5	年龄	出生日期	2022—“出生年”

2.2 数据来源

在微博高级搜索中，选择重大突发公共卫生事件，如重大传染病防控等突发公共卫生事件的全部热门微博，检索时间为 2020 年 1 月 23 日—4 月 8 日，选取微博用户评论数大于 1 000 条的 10 条微博作为数据来源，见表 2。

表 2 数据来源情况

序号	微博编号	发布者	评论数	发布时间
1	01200124	央视新闻	2 175	2020 年 1 月 24 日
2	02200131	环球时报	1 303	2020 年 1 月 31 日
3	03200214	人民日报	2 303	2020 年 2 月 14 日
4	04200215	央视新闻	2 170	2020 年 2 月 15 日
5	05200317	人民日报	2 088	2020 年 3 月 17 日
6	06200317	央视新闻	1 314	2020 年 3 月 17 日
7	07200331	胡锡进	1 515	2020 年 3 月 31 日
8	08200407	人民日报	1 400	2020 年 4 月 7 日
9	09200407	头条新闻	1 142	2020 年 4 月 7 日
10	10200408	头条新闻	2 236	2020 年 4 月 8 日

爬取上述 10 条微博的评论数据及评论用户个人主页信息，包括性别、年龄、粉丝数、评论点赞数、简介、标签、关注数以及昵称长度等用户特征数据。受平台对每条微博评论数据爬取数量的限

制，共得到评论数据 3 408 条，评论用户特征数据 3 145 条。使用 Excel 软件对重复、无关的评论数据，以及缺失、错误的用户特征数据进行删除。最终得到评论数据 3 367 条，评论用户特征数据 3 107 条。

3 结果

3.1 情感倾向度计算

调用百度 AI 平台的自然语言处理 API，计算所有用户评论文本的情感倾向度。若一个用户发表多条评论，则分别计算该用户每条评论的情感倾向。情感倾向度 (Q) 计算公式为：

$$Q = 1 \times P_1 + (-1) \times P_2 \tag{2}$$

其中， P_1 为文本内容属于积极类别的概率， P_2 为文本内容属于消极类别的概率，当 $Q \geq 0$ 时，用户情感为正向；当 $Q < 0$ 时，用户情感为负向。

以评论“多保持积极乐观的心态，一起加油”为例，得到以下结果：positive_prob: 0.999 89, confidence: 0.999 756, negative_prob: 0.000 108, sentiment: 2。该结果表示文本内容属于积极类别 (positive_prob) 的概率为 0.999 89，属于消极类别 (negative_prob) 的概率为 0.000 108，分类的置信度 (confidence) 为 0.999 756。则该用户的情感倾向度为 $0.999\ 89 - 0.000\ 108 = 0.999\ 782$ ，情感为正向。带有正向情感的用户评论其情感倾向度计算结果均大于 0，带有负向情感的用户评论其情感倾向度计算结果均小于 0，可认为此情感倾向度计算方法可行。

3.2 情感倾向度、用户特征相关分析

对微博评论情感倾向度与性别、年龄、粉丝数、评论点赞数、简介、标签、关注数以及昵称长度 8 个用户特征分别进行相关分析。在进行相关分析之前，首先对微博评论情感倾向度进行正态性检验，显著性水平确定为 0.05，柯尔莫哥洛夫正态性检验结果 $P < 0.05$ ，微博评论情感倾向度不符合正态分布，因此选择斯皮尔曼相关系数进行相关分析，见表 3，微博用户情感倾向度与性别、年龄、粉丝数、评论点赞数、简介和标签 6 个用户特征之间存在相关关系。

表 3 相关分析结果

序号	用户特征	From	To	Correlation	P
1	性别	Gender	Emotion	0.115	0.000**
2	年龄	Age	Emotion	-0.025	0.041*
3	粉丝数	Fans	Emotion	0.057	0.001**
4	评论点赞数	Like	Emotion	-0.116	0.000**
5	简介	Intro	Emotion	0.034	0.038*
6	标签	Tag	Emotion	0.192	0.000**
7	关注数	Attention	Emotion	-	0.085
8	昵称长度	Nickname	Emotion	-	0.129

注：* 表示 $P<0.05$ ，** 表示 $P<0.01$ 。

3.3 用户画像构建

随机选取的 3 个微博评论用户为例，分别构建突发公共卫生事件微博评论用户画像（ID = 001、002、003），见图 2。用“+”表示正相关，“-”表示负相关，则上述 3 例用户画像中用户特征与其情感倾向度的相关关系，见表 4。此 3 例用户画像的用户特征与其情感倾向度的相关关系与表 3 相关分析结果一致。

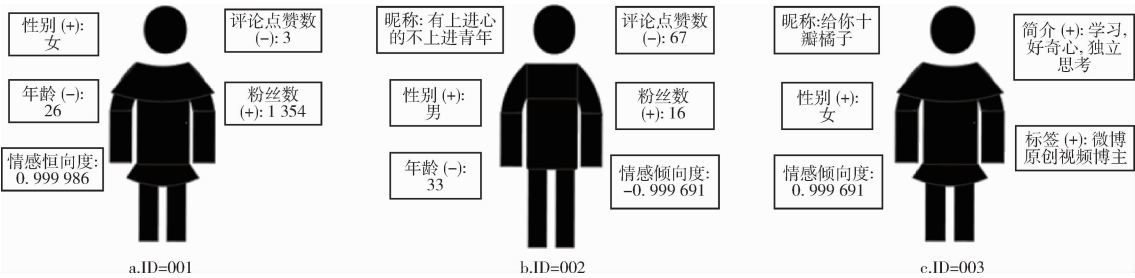


图 2 突发公共卫生事件微博评论用户画像举例

表 4 用户特征与情感倾向度的相关关系

画像 ID（情感倾向度）	性别	年龄	粉丝数	评论点赞数	简介	标签
001（0.999 986）	女（+）	26（-）	1 354（+）	3（-）	-	-
002（-0.923 855）	男（+）	33（-）	16（+）	67（-）	-	-
003（0.999 691）	女（+）	-	-	-	有（+）	有（+）

4 讨论

4.1 女性更倾向于发表正向评论，男性更倾向于发表负向评论

性别与情感倾向度成正相关。相对于男性，女性更倾向于发布正向评论，这可能是因为在面对突发公共卫生事件时，男性更理性，更关注事件本身存在的问题，具体表现为男性用户更倾向于对反映突发公共卫生事件防控过程中出现问题的报道发表评论，因此更倾向于发表负向评论；而女性更感性，更关注感人瞬间，具体表现为更倾向于对反映人性光辉的报道发表评论，因此更倾向发表正向评论。

4.2 用户年龄越大越倾向于发表负向评论，年龄越小越倾向于发表正向评论

年龄与情感倾向度成负相关。年龄对用户情感倾向的影响机制可能与性别类似，即年龄越大的用户越理性，更倾向对反映出现问题的报道发表评论，因此更倾向于发表负向评论；年龄越小的用户越感性，更倾向于对反映人性光辉的报道发表评论，因此更倾向发表正向评论。

4.3 用户粉丝越多越倾向于发表正向评论，粉丝数越少越倾向于发表负向评论

粉丝数与情感倾向度成正相关。一方面可能是

考虑到自身发表评论的影响较大,粉丝多的用户发表评论更为谨慎。另一方面自 2013 年国家开展“净网”行动以来,用户开始自觉规范自己的网络行为,社会责任感不断增强,对待突发公共事件在舆论表达上变得规范和有序,在态度立场上更加理性和专业^[10-12]。基于以上两方面原因,粉丝数越多的用户越倾向于发表正向言论。

4.4 有简介、标签的用户更倾向于发表正向评论,无简介、标签的用户更倾向于发表负向评论

简介、标签与情感倾向度成正相关。这可能与网络用户真实身份信息展示程度有关。对于用户在网络舆情事件中的参与行为,在实名制网络平台中要比非实名制平台中理性得多^[13]。简介和标签在一定程度上能够展示用户的身份信息,有简介和标签的用户在发表评论时会更理性,也会更注意评论的措辞,因此有简介和标签的用户更倾向于发表正向评论。

综上所述,就突发公共卫生事件微博评论用户特征而言,年龄大、粉丝数少、无简介、无标签、男性的用户更倾向于发表负向评论。因此,在突发公共卫生事件发生时,一方面可以通过构建用户画像对可能发表负向评论的用户进行识别,从而在一定程度上实现网络舆情的精准监控。另一方面,可以根据用户特征采取相应措施进行分类引导,如针对男性用户可及时推送事件发展动态及问题解决方案,使其全面、客观、准确了解事件发展全过程;针对无简介、标签的用户在发表评论时显示其网络地址等。

5 结语

本研究通过对突发公共卫生事件相关新浪微博用户评论进行情感分析,计算出用户评论的情感倾向度,并通过相关分析方法探索用户评论情感倾向度与用户特征之间的相关性,进而以用户评论情感倾向度以及性别、年龄、粉丝数、点赞数、简介、标签 6 个用户特征为内容构建突发公共卫生事件微博评论用户画像,并对用户评论情感倾向度与上述 6 个用户特征之间的相关关系进行讨论。由于微博舆情复杂多变,本研究仅进行了初步工作,且需要满足以下条件:微博用户的个人信息均是真实的;数据获取时间段内没有微博评论被删除或删除微博的数量对

本研究的总体工作没有影响。在后续研究中,将在对用户评论情感倾向度与用户特征进行相关分析的基础上,进一步以情感倾向度为因变量,以用户特征为自变量进行多因素分析,并融合用户评论中的表情、图片等信息进行多模态数据情感分析。

参考文献

- 1 张柳,王晰巍,李玥琪,等.信息生态视角下微博舆情生态性评价指标及实证研究[J].情报理论与实践,2022,45(3):35-41.
- 2 蒋知义,马王荣,邹凯,等.基于情感倾向性分析的网络舆情情感演化特征研究[J].现代情报,2018,38(4):50-57.
- 3 BARBOSA L, FENG J. Robust sentiment detection on Twitter from biased and noisy data [C]. Cambridge: The 23rd International Conference on Computational Linguistics, 2010.
- 4 BAKLIWAL A, FOSTER J, VANDER PUIL J, et al. Sentiment analysis of political Tweets: towards an accurate classifier [C]. Stroudsburg: NAACL Workshop on Language Analysis in Social Media, 2013.
- 5 刘勘,袁蕴英.基于词向量的微博情感倾向分类研究[J].图书情报工作,2018,62(15):92-101.
- 6 任中杰,张鹏,兰月新,等.面向突发事件的网络用户画像情感分析——以天津“8·12”事故为例[J].情报杂志,2019,38(11):126-133.
- 7 王若佳,严承希,郭凤英,等.基于用户画像的在线健康社区用户流失预测研究[J].数据分析与知识发现,2022,6(Z1):80-92.
- 8 沈洪洲,马巧慧,张一涵.可预测性灾害中网民情感传播特征研究[J].南京邮电大学学报(社会科学版),2022,24(4):46-57.
- 9 谢秋逸,周年兴,轩源,等.基于百度 AI 开放平台的北京市共享住宿顾客满意度及需求层次研究[J].南京师大学报(自然科学版),2021,44(1):64-70.
- 10 靖鸣,王勇兵.新浪大 V 传播行为的变化与思考——以突发公共事件为例[J].现代传播(中国传媒大学学报),2016,38(5):69-75.
- 11 崔鹏,张巍,何毅,等.突发公共事件网络舆情演化及政府应对能力研究[J].现代情报,2018,38(2):75-83,95.
- 12 安璐,吴林.融合主题与情感特征的突发事件微博舆情演化分析[J].图书情报工作,2017,61(15):120-129.
- 13 汪春香,徐立青,赵树成.影响食品安全网络舆情网民行为的主要因素识别研究——基于模糊集理论 DEMATEL 方法[J].情报杂志,2015,34(3):138-143.