

在线健康社区用户情绪识别与分析*

万欣 袁海曼

(武汉纺织大学管理学院 武汉 430200)

[摘要] 目的/意义 识别并分析在线健康社区用户情绪, 剖析疾病污名化情景下传染病社区用户的病耻感来源及情绪反应类型, 为疾病治疗、心理健康维护提供建议。方法/过程 爬取百度贴吧数据, 基于 K-means 聚类算法和 CorEx_Topic 主题模型识别社区用户的交流主题, 进行情感分析和情绪词抽取。结果/结论 社会公众和亲人朋友会从社会生活和日常生活两方面加深患者的病耻感情绪; 患者会产生“自卑内疚”“恐慌害怕”“绝望崩溃”3 种类型的病耻感情绪反应。同时从国家、社会和在线社区 3 个层面提出建议。

[关键词] 病耻感; 百度贴吧; 聚类; 主题模型; 情感分析; 在线健康社区

[中图分类号] R-058 **[文献标识码]** A **[DOI]** 10.3969/j.issn.1673-6036.2023.12.003

Recognition and Analysis of User Emotion in Online Health Community

WAN Xin, YUAN Haiman

School of Management, Wuhan Textile University, Wuhan 430200, China

[Abstract] **Purpose/Significance** To identify and analyze the emotions of users in online health communities, the paper dissects the sources of stigma and emotional reactions among users in infectious disease communities under scenarios of disease stigmatization and provides suggestions for disease treatment and psychological health maintenance. **Method/Process** The study crawls post data from Baidu Tieba, identifies the communication topics of community users based on the K-means clustering and the CorEx_Topic model, and conducts sentiment analysis and emotional word extraction. **Result/Conclusion** The public and relatives and friends will deepen the stigma of patients from both social life and daily life, resulting in three types of stigma emotion reactions: “self-inferiority and guilt”, “panic and fear” and “hopelessness and breakdown”. Additionally, advices are given from three levels; state, society and online community.

[Keywords] stigma; Baidu Tieba; clustering; topic model; sentiment analysis; online health community

1 引言

互联网技术的飞速进步与网民数量的不断增长, 促进了各类社交媒体平台的蓬勃发展, 以医疗

健康为主题的在线健康论坛、社区也随之兴起, 成为社区用户(包括医生、患者及家属等)疾病咨询、健康问题答疑、经验分享、社会支持交换的主要交流平台^[1]。基于在线健康社区中积累的大量有关社区用户对健康问题的意见、观点和情感等, 许多学者、医疗保健从业者对社区用户、患者在医生、疾病、药物、治疗等方面的感受和想法进行挖掘分析, 围绕滥用行为监测^[2]、药物不良反应识别^[3]、医疗保健产品与服务满意度调查^[4]、公共卫生政策影响分析^[5]、患者健康状态和心理状态检测

[修回日期] 2023-06-21

[作者简介] 万欣, 博士, 副教授, 硕士生导师, 发表论文 10 余篇。

[基金项目] 湖北省教育厅人文社科一般项目(项目编号: 18Y081)。

与跟踪^[6]以及流行病预防与监测^[7]等诸多研究方向展开探讨。

病耻感^[8]是对某种疾病的负面心理情绪,在传染病患者的情绪表达中尤其常见。Weiss M G 等^[9]将健康领域的病耻感定义为因经历过或预感到社会对特定健康问题个体或团体的不良评价,而形成的以拒绝、排斥、指责为特征的社会过程或相关个人经历。患者病耻感情绪将增加其疾病负担、加剧心理痛苦,还可能对疾病进程或医疗结果产生不利影响,在极端情况下还可能导致暴力。目前医疗保健领域的情感分析研究主要集中在癌症^[10]、糖尿病^[11]、抑郁症^[12]等慢性病在线健康社区中,较少对传染病社区进行情感分析研究,也少有研究关注社区用户情绪表达中的病耻感心理状态。因此,本研究以乙肝在线健康社区的用户为研究对象,综合使用主题模型、文本挖掘、情感分析等研究方法,探究严重疾病污名化背景下在线乙肝社区中用户的病耻感来源及情绪反应。

2 数据与方法

本研究对在线健康社区用户的病耻感分析框架,见图 1。分为数据采集、数据预处理、主题识别、情感分析、结果讨论 5 个部分。

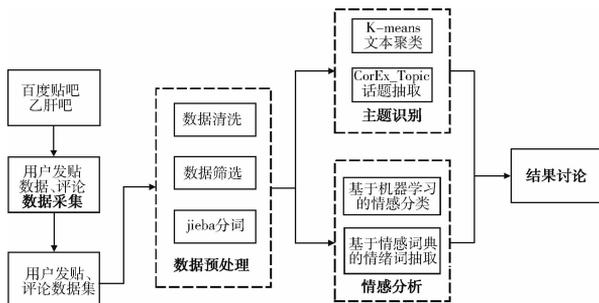


图 1 病耻感分析框架

2.1 数据来源与数据预处理

百度贴吧是全球领先的中文社区,拥有数量大、分布范围广和教育水平均衡的用户群体。本研究采用爬虫软件,采集百度贴吧小组“乙肝吧”中发布时间为 2015 年 1 月 1 日—2022 年 9 月 16 日的用户发

帖及评论数据,采集数据字段包括帖子标题、帖子内容、评论内容、发布时间、发布用户等。

为了提高后续主题挖掘和情感分析的准确性,对爬取数据进行预处理,填充缺失值,去除重复数据、无意义链接、图片、符号等,同时通过数据筛选去除无分析意义的帖子,例如征婚交友帖、广告活动帖等。预处理后得到 8 743 条发帖数据和 78 390 条帖子评论数据。根据帖子和评论数据,构建“乙肝患者用语”自定义词典,结合哈尔滨工业大学停用词表,采用 Python 中文分词库 jieba 对用户发帖数据和评论数据分词,得到后续分析文本语料库。

2.2 研究方法

2.2.1 主题分析 经过多次主题抽取分析尝试,结合文本数据特征,采用 K-means 聚类和 CorEx_Topic^[13]主题模型结合的话题抽取方法。首先,基于 K-means 进行用户发帖的文本聚类。为了提高聚类结果的准确性,使用哈尔滨工业大学讯飞联合实验室研发的中文预训练模型 RoBERTa-wwm-ext 实现用户发帖文本数据句子向量化,将每个帖子生成为一个 768 维向量,并采用截断奇异值分解算法降低向量维数,提高聚类效率和精度。根据轮廓系数确定最佳聚类数,采用 K-means 聚类算法得到 k 个聚类。其次,采用支持分层主题建模的 CorEx_Topic 主题模型进行主题抽取。该主题建模方法提供一种整合领域知识机制,即通过锚定词汇和信息瓶颈,也就是添加指定主题的锚定词汇得到解释性较高的主题,允许用户提升主题表达效果。使用聚类结果总结出的类别话题关键词作为先验知识注入 CorEx_Topic 进行主题抽取,获得主题类别及其对应话题词。最后,结合对数据的观察,将每一条发帖数据归入一个唯一主题类别,每个类别的主题由抽取出的话题词表示。

2.2.2 情感分析 首先,基于机器学习算法训练情感分类模型。随机抽取 3 000 条用户发帖数据人工标注为积极、消极、中性 3 种情感极性并进行文本数据向量化,作为机器学习算法的输入训练分类模型。在综合考虑多种分类模型的准确率后,选择基于支持向量机算法的分类模型对未分类文本数据

情感分类。对分类结果审查校对，获得情感分类结果。其次，基于情感词典抽取用户发帖和评论的情绪词。设置由基础情感词典、医学词汇词典和网络流行词典构成的自定义词典。其中基础情感词典由大连理工大学的中文情感词汇本体库中的 27 466 个情感词构成，医学词汇词典由百度词库中的医学词汇词典整理而成，网络流行词典由细胞词库中的网络流行词构成。根据分析内容对情感词典中的非当前语境情感词进行适当修改，比如删除词语“病毒”。最后，通过定制情感词典，抽取得到用户发帖及评论中能反映用户心理状态的情绪词汇，并按

照情感类别划分。

3 结果分析

3.1 用户发帖话题分析

当 k 值为 5 时，轮廓系数最大，此时 $K - means$ 算法的聚类效果最好。因此将用户发帖数据分为 5 个类别。根据聚类结果，对每个类别的发帖文本进行分词和词频统计，选择词频前 50 位中最能代表该类别的关键词作为先验知识注入主题模型。通过观察和调整，最终确定 7 组关键词词组，见表 1。

表 1 先验知识

序号	关键词							
1	帮忙	指点	求助	解答	建议	医生	检查报告	治疗
2	感受	历程	经历	分享	自愈	医院	生活	记录
3	工作	公司	入职	歧视	入职体检	面试	上班	健康证
4	孩子	母婴传播	出生	产检	怀孕	阻断	孕期	幼儿园
5	结婚	男朋友	女朋友	父母	对象	分手	离婚	接受
6	恩替卡韦	中药	效果	替诺福韦	副作用	上市	抗病毒	停药
7	恶心	失眠	腹胀	牙龈出血	蜘蛛痣	红点	疙瘩	症状

导入先验知识之后，主题平均相关性由 0.77 上升至 2.39。因此，结合先验知识，将 15 个主题分析结果总结归纳为 3 个主题 7 个话题，并根据话

题类别对聚类结果进行二次调整，保证每条用户帖子数据都被分在一个唯一类别里，从而获得最终话题分类结果，见表 2。

表 2 用户发帖主题分类

主题	描述	主题帖子数量 [n (%)]	话题	例词	话题帖子数量 [n (%)]
信息需求	社区用户对乙肝的治疗、病情分析及用药、症状方面的交流	6 292 (71.97)	病情求助	检查报告、指导、建议、复查、帮忙	3 901 (44.62)
			用药咨询	效果、国产、换药、恩替卡韦、副作用	2 154 (24.64)
			症状咨询	失眠、腹胀、恶心、出血、红点	237 (2.71)
经验分享	患病经历、治疗心得及医院、医保等相关交流	1 776 (20.31)	患病经历倾诉	分享、记录、生活、感受、历程	1 776 (20.31)
生活场景	社区用户有关人际交往和工作经历的交流	675 (7.72)	婚恋交友	结婚、男朋友、坦白、接受、嫌弃	151 (1.73)
			生育	孕期、阻断、害怕、乙肝疫苗、产检	152 (1.74)
			工作	工作、公司、入职体检、歧视、社会	372 (4.25)

分类结果显示，“信息需求”类主题占比最高，其次是“经验分享”类主题，这是因为社区中大多数用户发帖主要是为了寻求各种与疾病治疗有关的健康信息，用户的社会支持交换中信息支持交换占据重要地位，集中体现了在线健康社区作为信息支持来源的重要作用。“生活场景”类主题交流话题

显示，用户愿意在网络匿名空间与他人交流生活相关的话题，表达个人烦恼和心理状态。

3.2 用户发帖情感分析

各话题的情感极性分布，见表 3。总体上看，中性情绪在每个类别中的占比均较高，因为用户发

帖以无情感表达的提问为主。相较于其他类别,“患病经历分享”话题中用户表达积极情绪最多,考虑有两方面原因:一是很多用户会分享治愈和战胜疾病的个人历程,给予其他用户鼓励和信心;二是一些乐观的用户会发表正能量的观点,鼓励其他用户积极面对疾病,例如“今天开始抗病毒了,给自己加油,也给战友们加油”。而生活场景类别中的“婚恋交友”“生育”“工作”3个话题消极情绪的占比是积极情绪的数倍,表明在现实生活中,社区用户在这3个生活方面更易产生负面情绪。

表3 话题情感分布

主题和话题	积极 (%)	中性 (%)	消极 (%)
信息需求	1.11	94.83	4.05
病情求助	0.77	94.87	4.36
用药咨询	1.86	95.12	3.02
症状咨询	0	91.60	8.40
经验分享	21.42	59.36	19.22
患病经历倾诉	21.42	59.36	19.22
生活场景	6.84	50.37	42.79
婚恋交友	9.93	19.87	70.20
生育	11.92	64.24	23.84
工作	3.50	57.14	39.35

4 讨论与建议

4.1 用户病耻感来源

慢性乙肝极少会出现疼痛症状,因此排除因病痛带来的负面表述。分析结果显示用户在婚恋交友、工作、生育这3个现实场景交流中可能遭受较多污名,从而产生更多负面的病耻感情绪。基于这3个现实生活场景中可能涉及的社会群体,本研究提出患者的病耻感由以下两类人群引起。

4.1.1 社会公众 社会公众是疾病污名的重要主体,对传染病的刻板印象、偏见和误解是患者感知病耻感和实际病耻感的重要来源。基于对传染病的刻板印象,大部分人会产生诸如传染病会通过日常接触传播、传染病患者私生活混乱等偏见和错误认知,以打标签的形式将传染病患者隔离在人群之外,并持歧视和远离的态度,如 Leng A 等^[14]通过调查发现,64.7%的人不愿意让孩子与乙肝患者或

者携带者玩耍,51.4%的人不愿意与乙肝患者或携带者一起吃饭,42.8%的人不愿意与乙肝患者或携带者拥抱或握手。这种歧视和不公平对待的经历,形成了患者的感知病耻感和实际病耻感,并最终将这种经历带来的负面情绪内化为患者自卑、羞愧等内在病耻感。

4.1.2 亲属、朋友和同事等身边人 许多乙肝患者患病之后都会遭到亲属、朋友、同事等身边人的不自觉远离、孤立,甚至是嫌弃、埋怨和指责等不友好对待。在工作中惧怕同事知道自己的病情;在婚姻和生育中不仅要忍受亲人的不理解和指责,也要承受将疾病传染给伴侣、孩子后产生的极度愧疚、自我憎恨心理,导致乙肝患者不敢与其他人接触甚至拒绝社交,远离亲密关系。

4.2 用户病耻感情绪反应

基于大连理工大学情感词类型,将提取出的情感词划分为乐、好、怒、哀、惧、恶、惊7类,发现情绪词主要集中分布在哀、惧、恶3个类别。结合原始用户发帖文本,总结归纳乙肝患者及其家属有以下3种类型的病耻感情绪反应。

4.2.1 “自卑内疚”型 一方面,社会中的绝大多数人对乙肝患者的区别对待和歧视行为,造成了乙肝患者的自卑、自我厌弃心理,严重者甚至在疾病治愈后仍然无法改变这种心理状态,用“沉重”“孤独”“痛苦”“悲哀”“孤立”等词描述自己的心理状态。另一方面,隐瞒病情或将疾病传染给他人,会让乙肝患者产生极度愧疚自责心理,尤其是乙肝携带者孕妇,由于母婴传播是目前乙肝病毒最主要的传播途径之一,遭受了更多的指责和压力,“自责”“愧疚”“悔不当初”“自私”“凶手”等词在其帖子中反复出现。

4.2.2 “恐慌害怕”型 这种情绪反应来源于两个方面:对疾病本身可能带来的伤害的害怕、对他人的指责和歧视的恐惧。主要在3种状态用户中出现:刚刚得知自己为乙肝患者或携带者、得知亲属或周围人为乙肝患者或携带者、被迫向其他人坦白病情。其用“噩梦”“恐惧”“忐忑”“魂不守舍”“如临大敌”“心如刀割”等词语描述自己的心理状态。

4.2.3 “绝望崩溃”型 慢性乙肝治愈率低且存在反复发作的情况，导致许多患者会在长年累月的治疗过程中丧失活下来的勇气和信念，因此很多乙肝患者同时患上了抑郁症、焦虑症等精神疾病，用“苟延残喘”“折磨”“苦难”“煎熬”“一无所有”等词形容自己的人生。

4.3 启示与建议

根据以上分析，乙肝患者在社会生活和日常生活中存在严重的病耻感情绪，这不利于疾病治疗和心理健康，如果任由这种负面情绪发展，可能会产生极端行为，对社会造成不良影响。因此本研究从国家、社会和在线社区 3 个层面提出以下建议。一是国家层面，应进一步加强对用人单位的反乙肝歧视教育，强化相关法律政策的实施。尽管已经存在反乙肝歧视的法律法规，但是个别用人单位并没有正确执行，剥夺了乙肝患者作为公民的正当权益，相关部门应重视此类歧视现象，从法律层面保证乙肝患者工作和职业发展的权利。二是社会层面，有关组织应开展乙肝疾病特征、传染性、治疗方案等方面的科普，减少非乙肝患者的刻板印象和认知错误，帮助乙肝患者了解更多疾病知识，减少恐慌、害怕心理，弱化其病耻感情绪。三是乙肝在线健康社区层面，运营者应加大对正面事件的宣传，减少社区负面情绪和极端情绪的积累。乙肝治愈是一个艰难又漫长的过程，参与社区的乙肝患者既需要信息支持，又需要情感支持鼓励其继续坚持治疗。因此，社区运营者应大力推广有关疾病治疗的正面信息，邀请已经治愈的乙肝患者参与社区活动、科普答疑，保证社区持续有正面导向出现，帮助正在治疗的患者保持信心和希望。

5 结语

本研究以在线健康社区“乙肝吧”的用户生成内容为数据来源，基于医疗保健行业的情感分析研究理论，以在线健康社区中慢性乙肝患者、携带者用户为研究对象，采用主题模型、情感分类的研究方法，结合情感词分类统计分析，探究在严重疾病

污名化背景下乙肝患者的病耻感情绪来源及其情绪反应。本研究存在一定的不足之处。一是在进行话题抽取时人工导入话题词并确定了主题类别标签，主观性和对领域认知的限制可能会使用户发帖的主题分类产生一定偏差，使一些话题词被错误地归类到其他主题类别中，从而导致某些类别被模糊掉，未来将引入专家知识确定话题词与主题。二是本研究仅基于在线社区中的用户发帖数据进行分析，未来研究将进一步结合线下调查数据，作出更合理详细的解释。

参考文献

- ALLEN C J, JAMES L W, PAUL M D, et al. Online health communities an assessment of the influence of participation on patient empowerment outcomes [J]. *Information technology & people*, 2013, 26 (2): 213 - 235.
- DANIULAITYTE R, CHEN L, LAMY F R, et al. “When ‘bad’ is ‘good’”: identifying personal communication and sentiment in drug - related tweets [J]. *JMIR public health and surveillance*, 2016, 2 (2): e162.
- KORKONTZELOS I, NIKKFARJAM A, SHARDLOW M, et al. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts [EB/OL]. [2023 - 06 - 05]. <https://doi.org/10.1016/j.jbi.2016.06.007>.
- DOING - HARRIS K, MOWERY D L, DANIELS C, et al. Understanding patient satisfaction with received healthcare services: a natural language processing approach [C]. Washington DC: AMIA Annual Symposium Proceedings, 2017.
- PAUL M, DREDZE M. You are what you tweet: analyzing twitter for public health [C]. Barcelona: The International AAAI Conference on Web and Social Media, 2011.
- HERDIANSYAH H, ROESTAM R, KUHON R, et al. Their post tell the truth: detecting social media users mental health issues with sentiment analysis [EB/OL]. [2023 - 06 - 05]. <https://doi.org/10.1016/j.procs.2022.12.185>.
- GARCIA - DIAZ J A, CANOVAS - GARCIA M, VALENCIA - García R. Ontology - driven aspect - based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America [EB/OL]. [2023 - 06 - 05]. <https://doi.org/10.1016/j.futur-e.2020.06.019>.

(下转第 39 页)

疾病管理体系。

参考文献

- 1 段承阿鑫, 常峰, 路云. 我国各省城镇职工基本医疗保险门诊特殊疾病政策比较研究 [J]. 中国卫生事业管理, 2017, 34 (5): 342-345, 369.
- 2 蔡海艳, 张涛, 岳靖凯, 等. 我国部分城市门诊特殊疾病管理现状与策略研究 [J]. 中国医院管理, 2021, 41 (3): 73-76.
- 3 浙江省人力资源和社会保障厅, 原浙江省卫生和计划生育委员会, 浙江省商务厅, 等. 关于进一步完善慢性病门诊医保政策有关事项的通知 (浙人社发 [2017] 106 号) [EB/OL]. [2023-03-15]. http://ybj.zj.gov.cn/art/2019/12/11/art_1229113757_564020.html.
- 4 国家卫生健康委员会, 国家中医药管理局. 关于印发互联网诊疗管理办法 (试行) 等 3 个文件的通知 (国卫医发 [2018] 25 号) [EB/OL]. [2023-03-16]. <http://www.nhc.gov.cn/zyygi/s3594q/201809/c6c9dab0b00c4902a5e0561bbf0581f1.shtml>.
- 5 国家医疗保障局. 关于积极推进“互联网+”医疗服务医保支付工作的指导意见 (医保发 [2020] 45 号) [EB/OL]. [2023-03-16]. http://www.gov.cn/zhengce/zhengceku/202011/03/content_5556883.htm.
- 6 陈喆, 沈赟, 秦艳, 等. 互联网医院安全合理用药闭环管理体系的建立与实践效果评价 [J]. 中国临床药理学杂志, 2022, 31 (1): 15-20.
- 7 牛雨婷, 赵允伍, 王晓松, 等. 三医联动视角下老年慢性病管理发展现状及对策 [J]. 卫生经济研究, 2023, 40 (4): 48-51, 55.
- 8 沈勤, 徐越. 老年人参与“互联网+慢性病管理”意愿影响因素分析——基于 Anderson 健康行为模型的实证研究 [J]. 卫生经济研究, 2020, 37 (1): 45-48.
- 9 韩润翔, 付晶晶, 刘思奇, 等. “互联网+”医疗服务在老年人群中的应用研究进展 [J]. 护理研究, 2021, 35 (20): 3657-3660.
- 10 杨小玲, 袁丽. 互联网医疗在老年慢性病管理中的应用进展 [J]. 实用老年医学, 2021, 35 (2): 114-117.
- 11 郭一娇, 刘德阳, 唐晓蓉, 等. 处方流转的影响因素分析及平台构建的质性研究 [J]. 中国药房, 2023, 34 (11): 1307-1312.
- 12 王秀红, 王兰兰, 于晴, 等. 子女的支持与帮助对中老年人互联网使用的影响 [J]. 绿色科技, 2019 (7): 278-280.
- 13 成水芹, 张志宏, 俞雨生. 远程监控在自动化腹膜透析患者中的应用 [J]. 肾脏病与透析肾移植杂志, 2020, 29 (3): 275-279.
- 14 许春燕, 吴海波, 张彪, 等. 失能老人“互联网+居家上门护理”模式发展路径优化探析 [J]. 保险职业学院学报, 2023, 37 (1): 81-87.
- 15 杨勇, 陈少飞, 田侃. 电子处方外配与药品电子监管平台合理衔接的探讨 [J]. 中国药房, 2014, 25 (41): 3847-3849.
- 16 国家远程医疗与互联网医学中心. 2021 中国互联网医院发展研究报告 [R/OL]. [2022-05-10]. <https://www.ntmc.hc.com/#/FileDownCate?id=2835938162206031873>.

(上接第 19 页)

- 8 LINK B G, PHELAN J C. Conceptualizing stigma [J]. Annual review of sociology, 2001, 27 (1): 363-385.
- 9 WEISS M G, RAMAKRISHNA J, SOMMA D. Health-related stigma: rethinking concepts and interventions [J]. Psychology, health & medicine, 2006, 11 (3): 277-287.
- 10 CRANNELL W C, CLARK E, JONES C, et al. A pattern-matched twitter analysis of US cancer-patient sentiments [J]. Journal of surgical research, 2016, 206 (2): 536-542.
- 11 GABARRON E, DORRONSORO E, RIVERA-ROMERO O, et al. Diabetes on twitter: a sentiment analysis [J]. Journal of diabetes science and technology, 2019, 13 (3): 439-444.
- 12 BIRJALI M, BENI-HSSANE A, ERRITALI M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks [EB/OL]. [2023-06-05]. <https://doi.org/10.1016/j.procs.2017.08.290>.
- 13 GALLAGHER R J, REING K, KALE D, et al. Anchored correlation explanation: topic modeling with minimal domain knowledge [EB/OL]. [2023-06-05]. https://doi.org/10.1162/tacl_a_00078.
- 14 LENG A, LI Y, WANGEN K R, et al. Hepatitis B discrimination in everyday life by rural migrant workers in Beijing [J]. Human vaccines & immunotherapeutics, 2016, 12 (5): 1164-1171.