基于爬虫技术获取新型冠状病毒（2019-nCoV）贴吧数据的网络舆情分析及应对策略

耿 辉[1,2]，马 茂[2]，张 勇[3]，尹小妹[4]，徐安定[5*]，吕 军[1*]

1. 暨南大学附属第一医院临床研究部（广州 510630）
2. 西安交通大学第一附属医院体检部（西安 710061）
3. 西安交通大学第一附属医院国资科（西安 710061）
4. 西安交通大学第一附属医院儿科（西安 710061）
5. 暨南大学附属第一医院神经内科（广州 510630）

【摘要】新型冠状病毒感染的肺炎是一种急性感染性肺炎，2019 新型冠状病毒具有很强的传染性。自 2019 年年底发现到 2020 年年初快速扩散到全球数十个国家和地区，引起疫区大众严重的心理恐慌，因此本文使用爬虫获取新型冠状病毒贴吧数据进行舆情分析，其目的是快速获取大量可靠、完整的信息数据，以分析大众心理健康状况，并据此做出应对策略，以期推动疫区大众健康心理干预的研究和发展。

【关键词】新型冠状病毒；爬虫；舆情分析；应对策略

Analyzing internet public sentiment of the novel coronavirus (2019-nCoV) base on Python crawler from Post Bar and its countermeasures

Hui GENG[1,2], Mao MA[2], Yong ZHANG[3], Xiao-Mei YIN[4], An-Ding XU[5*], Jun LYU[1*]

1. Department of Clinical Research, The First Affiliated Hospital of Jinan University, Guangzhou 510630, China;
2. Physical Examination Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China;
3. Department of State-owned Assets Division, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China;
4. Department of Pediatrics, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China;
5. Department of Neurology, The First Affiliated Hospital of Jinan University, Guangzhou 510630, China;
**Corresponding author: Jun LYU, E-mail: lyujun2019@163.com; An-Ding XU, E-mail: tlil@jnu.edu.cn*

【Abstract】Novel coronavirus pneumonia (NCP) is an acute infectious pneumonia. The 2019 novel coronavirus, which is highly infectious, and has been rapidly spreading to the world by the end of 2019 and beginning of 2020. Countries and regions, causing serious psychological panic among the people in the epidemic

DOI: 10.12173/j.issn.1004-5511.2020.02.04
基金项目：国家社会科学基金一般项目（16BGL183）；陕西省软科学研究计划一般项目（2017KRM211）；西安交通大学第一附属医院院基金（2019RKX-05）
* 通信作者：吕军，研究员，博士研究生导师，E-mail: lyujun2019@163.com;
徐安定，主任医师，博士研究生导师，E-mail: tlil@jnu.edu.cn

area, so this article uses crawler to obtain novel coronavirus post bar data for public opinion analysis, the purpose of which is to quickly obtain a large number of reliable and complete information data to analyze their mental health and this makes a response strategy, with a view to promoting the research and development of psychological health intervention in the affected areas.

【Keywords】 2019–nCoV; Crawler; Public opinion analysis; Coping strategies

随着移动互联网技术的迅猛发展，智能手机持续普及，人们已习惯于手机快速查询各类资讯，在这个信息爆炸的时代，媒体形态比以往任何时候都要丰富，数据规模、类型呈几何式增长，但是数据价值普遍较低，信息越多，人们的思维越混乱，越难以辨别真伪，因此为了从海量的信息数据里获取有价值的数据，衍生了网络爬虫，研究人员通过设定获取信息源的规则获取到有价值的网络数据，再进行数据清洗、加工，构建信息调查的信息数据基础，本文就从当前热点新型冠状病毒贴吧获取到帖子以及回帖信息，掌握当前人们对新型冠状病毒的心理状态，并对其进行分析，尝试从积极的角度做出应对策略。

1 新型冠状病毒贴吧数据分析

截止 2020 年 2 月 3 日新型冠状病毒吧共有主题数 17 340 个，贴子数 581 104 篇，平均每分钟增加 1.2 个主题，65 篇回帖，是目前最火热的几个热议贴吧之一，该吧作为关注病毒的百度网友了解实时信息和灌水帖的前沿阵地，回帖最真实，具有典型的代表性。

2 Python爬取贴吧数据过程

首先分析贴吧网站分析网站结构，第一步获取初始的 URL: http://tieba.baidu.com/f?kw= 新型冠状病毒 &ie=utf-8；第二步根据初始的 URL 爬取该页面并获得新的主题帖的 URL，例如获得主题帖“疫情之下，记录 / 交流真实的生活”的 URL 为 : http://tieba.baidu.com/p/6470674116，爬取当前 URL 地址信息，解析网页内容，这里就只获取主题和回帖内容，关键 python 语句为“content['title'] = li.find('a', attrs={"class": "j_th_tit"}).text.strip() # .strip() 以及 print(li.find('a', attrs={"class": "j_th_tit"}).text.strip())”将网页有价值信息数据存储到 txt 等文件中；第三步设置爬虫的停止条件为爬完指定页面数后停止，因贴吧过大，陈旧的数据不必提取，这里就提取了 1 月份的贴吧数据。

利用 python 爬虫的高效便捷性，爬取一月份数据共耗时 6 分钟左右，导出为 txt 格式文件，以下为数据示意图（图 1）

为更好的了解帖子中的情绪反应，尝试

图1 python程序爬取贴吧帖子示意图
Figure 1. Python program crawl post

用 Python 做了中文情感分析，情绪分析结果显示总负面情绪为 58.44%，正面情绪为 41.56%，能够保持正常较好情绪的语言占 32.81%，恐惧情绪占 28.75%，厌恶情绪占 22.50%，乐观情绪占 8.75%，悲哀和惊恐情绪比例均是 2.5%，愤怒情绪占 2.19%，人们普遍抱有负面情绪，在严重传染病大流行期间，民众的心理情绪本身比较脆弱，加上疾病流行期间的传言与猜测，国家政府机关采取非常时期的应急措施，街道的冷清，甚至超市的萧条，因宣传出现误解引起的抢购风潮，再加上层层口罩手套防护的路人，微信、新闻媒体等不断对疾病进行渲染宣传，也易使人感受到生活的紧张气氛，衍生恐慌、悲观、厌恶等负面情绪，如图 2 所示。

词云分析也显示，大众关心每日新型冠状病毒确诊病例增长情况，发热门诊疑似病例人数和科学家能否将药物及时研制成功，国家对疫区的管控政策，中央领导对目前物资如口罩供应等情况的回应，其中明显有负面词汇出现，如：害怕、求助、死亡等（图 3）。

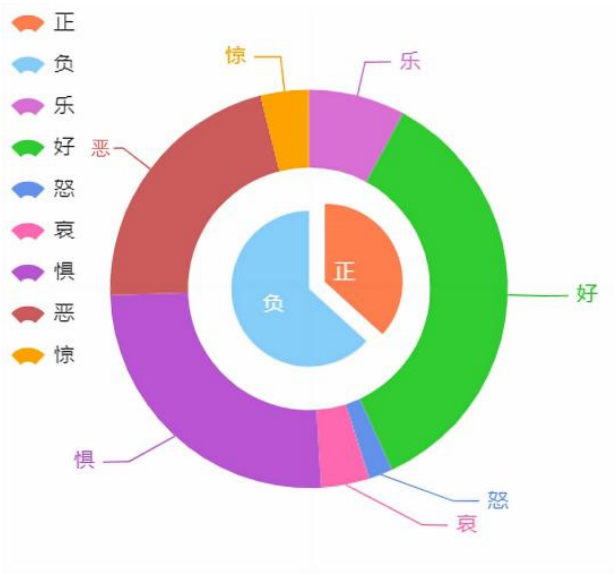

图2 贴吧中文情感分析得分
Figure 2. Chinese sentiment analysis score in post

图3 词云分析结果
Figure 3. Word cloud analysis results

3 讨论

新型冠状病毒感染对民众健康具有高度威胁性，其潜伏期短、起病急骤、病情凶险、目前无特效治疗手段，在严重威胁大众生命安全的同时，亦会对引起人们变现出恐惧、厌恶等情绪 [1]，对贴吧帖子的中文情感分析和词云分析我们也看到，新型冠状病毒的陌生性、不确定性和多变性，引起了民众普遍的负面情绪，而情绪持久性的紧张，会引起以焦虑、抑郁等为主要表现的应激相关精神障碍 [2]，因此，需做出以下应对策略：

3.1 加强民众医学知识普及工作

对新事物缺乏认知，成为民众恐惧的根源。抖音、快手等新媒体未来会成为快速普及医学科普知识和健康教育的重要宣传途径，在这一过程中新媒体应学习“丁香园”的做法，以医生的专业知识为背景，在公众号推出众多优质科普文章，面对新冠状病毒疫情，“丁香园”制作了疫情地图，来源为各地卫健委的实时报道，确诊、疑似、治愈、死亡人数一目了然，打通了全国疫情的“任督二脉”，普通民众也能分析疫情发展的趋势，大大延缓了因不确定性而带来的恐慌情绪蔓延。如果没有媒体的正确引导，极易出现各种抢购狂潮，这些密集抢购的人群中，发生输入性新型冠状病毒肺炎第二代病例概率大大增加，这些第二代病例的出现，对社会、社区防疫工作带来巨大困难，是医务人员潜在的职业暴露和医院感染的风险来源 [3]。因此，加强民众的医学知识普及工作迫在眉睫，并且抖音、快手等新媒体是普及医学知识的先锋队。

3.2 加强民众的心理疏导及人文关怀工作

在各地纷纷启动重大突发公共卫生事件一级响应的大背景下，地方政府部门应积极搭建心理专家为主的专业人士与大众沟通的渠道，开通以社区为中心的线上和线下心理咨询门诊，减轻乃至消除民众的负面情绪。

如无法改变当前非人性化的隔离模式，应充分尊重居家隔离人员对医疗信息、社会交往、亲友探访、学习工作及休闲娱乐等方面的心理需求，作为社区管理人员应尤其关注这一部分隔离的心理健康状况。对未产生负面情绪的隔离人员进行预防，建立以各种兴趣爱好为主的微信群，正向引导情绪、宣泄消极情绪。对心理反应较重的隔离人员，应定期进行心理疏导，同时联系医院进行线上心理咨询，必要时在做好防护的情况下开展线下心理咨询[4]。

3.3 重要信息专业化、公开化、及时化、去个性化

近年来，互联网和移动终端迅速发展，决定了传播广泛的信息更多的出现在以数字技术为代表的新媒体，其最大特点是打破了媒介之间的壁垒，消融了媒体介质之间地域、行政之间，甚至传播者与接受者之间的边界，例如微博及其回复帖子等，其快速及时高效使其成为媒体宣传重要信息的宠儿。同时，新媒体也可以做到将受众细分，个性化的推送其需要新闻，这就意味着，受众的信息来源如果只通过新媒体，就有可能接收到片面的信息。因此，在政府启动重大突发公共卫生事件的响应下，应建立媒体统一的对外宣传平台，例如设立各地网络新闻发言人制度，这些网络新闻发言人的服务对象是大众，日常进行时效性强的权威信息发布工作，进行突发事件舆论引导，清楚传达政府意见的同时也要便于公众接受，尤其重视民众心理情绪的安抚。

另外随着老龄化的加剧，社区是老人的第一道防疫阵地。本次疫情从各地公布的死亡病例详细信息分析来看，中老年人所占比例较高，更需要社区和家庭紧密配合，做好防疫。调查显示社区中 69.08% 的老年人使用手机获取健康信息，但仍有一些高龄、健康意识和健康自评情况低的老年人很少使用手机获取健康信息[5]，需要社区管理人员及时以多种方式提供教育资源，如电话、上门随访等，安抚老人紧张情绪的同时提高老年人健康防护意识，并为老人提供一些健康有趣的爱好，如书法、绘画、摄影、集邮等，使他们老有所乐。同时在社区卫生服务中心开设预检分诊，对发热病人进行逐一排查，减轻社区交叉感染的可能性，消除社区大众的恐慌情绪。

参考文献

1 李进，李文全，聂滨，等. 输入型难辨性新型冠状病毒感染的肺炎一例及传播模式分析 [J/OL]. 华西医学：1-4[2020-02-04]. DOI: 10.7507/1002-0179.202002001. [Li J, Li WQ, Nie B, et al. Analysis of a case of pneumonia and the transmission mode of imported indistinct coronavirus infection[J/OL]. West China Medical Journal: 1-4[2020-02-04].]

2 王琛，王旋. 新型冠状病毒感染的流行、医院感染及心理预防 [J/OL]. 全科护理 :1-2[2020-02-04]. [Wang C, Wang X. Epidemic of new coronavirus infections, nosocomial infections and psychological prevention[J/OL]. Chinese General Practice Nursing: 1-2[2020-02-04].]

3 徐明川，张悦. 首批抗击新型冠状病毒感染肺炎的临床一线支援护士的心理状况调查 [J/OL]. 护理研究 :1-3. [2020-02-04]. [Xu MC, Zhang Y. Psychological survey of the first-line clinical front-line support nurses to combat new coronavirus-infected pneumonia[J/OL]. Chinese Nursing Research: 1-3. [2020-02-04].]

4 王文凯，于海涛，曹霞，等. 社区老年人使用智能手机获取和利用健康信息的调查分析 [J]. 中华医学图书情报杂志，2019, 28(08): 71-76+80. DOI: 10.3969/j.issn.1671-3982.2019.08.011. [Wang WK, Yu HT, Cao X, et al. Accession and use of health information with intelligent mobile phone in community elderly people[J]. Chinese Journal of Medical Library and Information Science, 2019, 28(08): 71-76+80.]

5 唐亚梅，张殷殷，李建国，等. 传染性非典型性肺炎患者出现焦虑、抑郁等神经精神损害症状的特征 (英文) [J]. 中国临床康复，2005, (24): 208-209. DOI: 10.3321/j.issn:1673-8225.2005.24.007. [Tang YM, Zhang YY, Li JG, et al. Characteristics of neuropsychiatric impairment symptoms in patients with severe acute respiratory syndrome[J]. Chinese Journal Of Clinical Rehabilitation, 2005(24): 208-209.]

收稿日期：2020 年 2 月 4 日 修回日期：2020 年 2 月 8 日
本文编辑：李 阳 杨智华