Journal of Data Science 18 (5), 860–874 DOI: 10.6339/JDS.202012_18(5).0003
December 2020

On Public Sentiment and Topic Mining during the COVID-19 Pandemic Based on Sina Weibo Comment Data

Xiaomeng Du*[1], Wei Huang[1], Yijing Liu[1], and Haibo Su[1]

[1]*Data Science Lab, Percent Co., Beijing, China*

Abstract

In the wake of the COVID-19 outbreak, the public resorted to Sina Weibo as a major platform for the trend of the pandemic. Research on public sentiment and topic mining of major public sentiment events based on Sina Weibo's comment data is important for understanding the trend of public opinions during major epidemic outbreaks. Based on classification of the Chinese language into emotion categories in psychology, we use open source tools to build naive Bayesian models to classify Weibo comments. Visualization of comment topics is achieved with word co-occurrence network methods. Commented topics are mined with the help of the latent Dirichlet distribution model. The results show that the psychological sentiment classification combined with the naive Bayesian model can reflect the evolvement of public sentiment during the epidemic, and that the latent Dirichlet distribution model and word co-occurrence network can effectively mine the topics of public concerns.

Keywords *Dirichlet distribution; emotional classification; naive Bayes; visualization; word co-occurrence*

*Corresponding author, Email: xiaomeng.du@percent.cn

 Received March, 2020; Accepted May, 2020

Journal of Data Science 18 (5), 860–874
2020 年 12 月
DOI: 10.6339/JDS.202012_18(5).0003

新型冠状病毒疫情期间公众情绪演化以及话题挖掘研究 ---基于新浪微博评论数据

杜晓梦*[1]，黄伟[1]，刘译璟[1]，苏海波†[1]

[1] 北京百分点信息科技有限公司数据科学实验室

摘要

新型冠状病毒疫情期间，新浪微博成为公众了解疫情动向的重要平台。基于微博评论数据研究疫情期间公众情绪演化过程以及重大舆情事件的话题挖掘，对于把握重大疫情期间的舆论走向，做好舆论引导工作具有重大意义。本文基于心理学界对于中文的情感划分，通过开源编程软件构建朴素贝叶斯模型对微博评论进行情感分类；基于词共现网络方法实现评论话题可视化；基于潜在狄利克雷分布模型实现对评话题的挖掘。研究结果表明，结合心理学的中文情感与朴素贝叶斯模型可以体现公众在疫情期间的情感变化过程，结合潜在狄利克雷分布模型与词共现网络方法可以有效挖掘舆论场话题。

关键词 词共现；狄利克雷分布；可视化；朴素贝叶斯；情绪分类

1 引言

近年来，随着社交媒体的广泛应用，论坛、博客、QQ、微博、微信等社交网络平台成为公众发表意见和表达情绪的重要途径。社交媒体平台具有海量网络信息资源，并以其方便快捷的交互方式对公众在工作、学习以及日常生活等各方面发挥着一定的作用，同时公众可以在网上发表自己的言论。本文通过分析疫情期间公众在新浪微博中的评论，分析公众的情绪变化，通过理论研究对实践层面的监管起到一定的指导作用。

1.1 背景介绍

2020 年春节期间，2019 新型冠状病毒 (World Health Organization, 2020)（2019 novel coronavirus, 以下简称新型冠状病毒）爆发并迅速蔓延全国乃至全世界，病毒潜伏期较长，同时传染性较强，疫情的爆发严重影响了社会经济的发展和人民的正常生活。为控制疫情，中国疫情大暴发中心武汉于 1 月 23 日开始实施出行禁令，全国各地陆续启动了紧急应对措施重大突发公共卫生事件一级响应。此时正逢春节，突如其来的疫情让人们心态骤变，加之权威防疫信息不足，公众心态触发了网络舆论，而网络舆论像一面镜子，照出了疫情中的公众心态。

根据国发院百分点数据智能与国家发展实验室发布的《新型冠状病毒疫情公众调研》报告 (中央广电总台国际在线, 2020)，该报告采用大小数据融合的方式，通过线上调研问卷收集民众的直接情绪反馈，结合微博采集的数据对直接采集的情绪信息进行补充和辅助研判，力争深入分

*通讯作者。电子信箱：xiaomeng.du@percent.cn。
†作者信息按姓氏拼音字母顺序排名

投稿于 2020 年 3 月; 接收于 2020 年 5 月

析人民群众在本次疫情期间的心理状态、行为轨迹、认知变化和疫后看法等，以及民众对国家疫情期间政策的态度。报告显示，随着疫情的发展，官方数据的不断公布，各方渠道的集中报道，受访者接收了大量的疫情信息，对于疫情的发展状况非常担忧，“比较担心”与“非常担心”整体占比达到 90%（如图1所示），焦虑、难过、心烦、恐惧、愤怒等负面情绪整体占比为 80% 以上（如图2所示）。

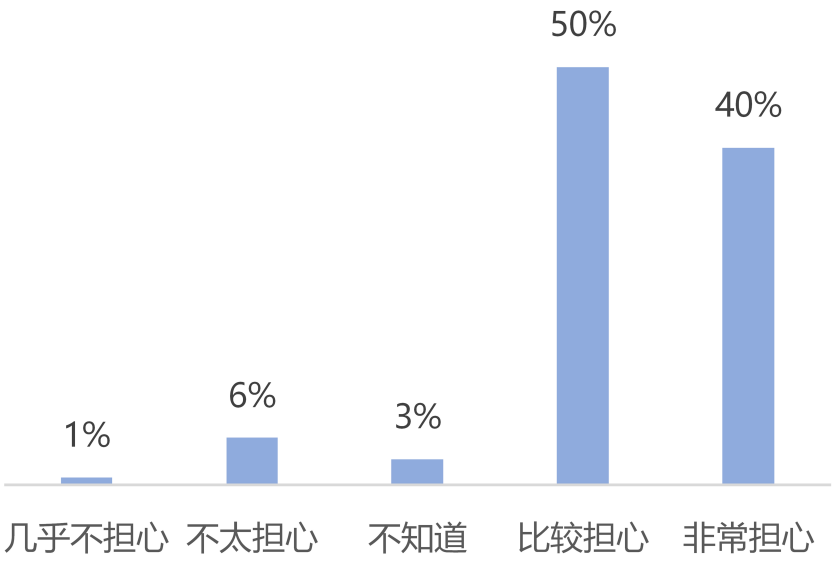

图 1: 对疫情的担心程度

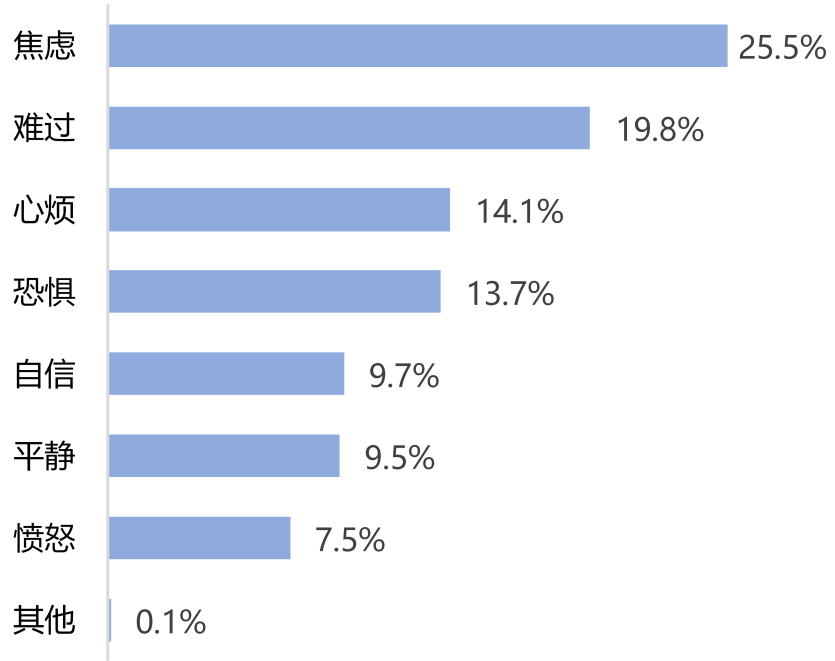

图 2: 对疫情的态度

随着疫情的发展，公众情绪不断变化，本文基于博文文本的情感分析，分析此次疫情从开始到武汉解封、全民逐渐复产复工的过程中，分析微博评论情感演变态势，发现舆情演变规律和潜在风险，为舆情引导提供决策支持。

1.2 研究问题

新型冠状病毒属于 β 冠状病毒属，是一种潜伏期长、传染性强且可致死的急性呼吸道传染病，主要通过飞沫和接触传播，自 2019 年底于武汉集中爆发后呈向全国乃至全球蔓延的发展趋势 (赵文明等, 2020)。由于疫情早期群众安全防范意识薄弱、受感染个体无法依靠客观知识正确判断自己是否感染，加之春节将近流动人口数量激增等因素使得感染人员数量激增，为了遏制疫情蔓延势头、阻断疫情外溢，2020 年 1 月 23 日上午武汉封城，与此同时全国也进入了疫情高发期。

为防止过度恐慌情绪的产生和蔓延并及时政府及媒体及时提出应急决策，对公众情绪进行实时监测及分析发生重大突发性公共卫生事件时公众的情绪衍变也显得尤为重要。但仅针对小数据或大数据的单一数据来源的研究很难实现主客观数据的融合，因此为更好的进行新型冠状病毒疫情期间公众情绪衍变趋势研究，本文收集了主流媒体数据和疫情期间确诊人数数据，基于朴素贝叶斯、词云、潜在狄利克雷分布和词共现社群挖掘方法的情感分类组合模型，分析了大众在本次疫情期间在不同时间段的情绪变化、关注热点话题和两者之间的关联性。

1.3 文献回顾

新型冠状病毒的爆发属于重大突发性公共卫生事件，重大突发性公共卫生事件的发生容易催生舆情主体的各种负面情绪和非理性行为。高文斌等 (2003) 通过心理咨询人数探索了 SARS 流行期间社会公众心理行为，证实重大突发性公共卫生事件会对公众造成了较大程度的心理压力。朱霞等 (2005) 分析了 SARS 期间调查问卷数据得出了突发性事件中不同角色公众的情绪状态也是不同的结论。

然而在通讯和网络发达的如今，仅通过小样本调研的方式已经无法完全获得公众全面的意见和看法，而社交媒体不仅是可以分享个人生活的平台，还可以用来监测公众对社会事件的看法及情绪变化的数据 (Han et al., 2019; Liu et al., 2010)，由此，广大学者也开始针对如何从互联网大数据中获取公众的情感指标进行分析研究。词频分析是定量语言学分析中一种广为人知、深入研究使用的统计方法 (Rajput et al., 2019, 2020; Moreno-Sánchez et al., 2016)。刘韩松 (2013)、孟雪井等 (2016)、陈茜等 (2015) 等学者从不同应用场景出发证实了通过文本挖掘、自然语言分析等方式可以将文字转化为对应的情绪，并进行量化分析和分类预测。王晰巍等 (2018) 基于贝叶斯模型对网络舆情用户情感演化进行研究，发现经济发达地区的微博用户相比于欠发达地区对同一话题的态度更加乐观。张海涛等 (2019) 基于复杂网络理论探索了微博热点事件评论与网民情感波动间的关系，并证实了网络舆论会影响对于舆情事件的发展方向。王一博等 (2014)、赵文清等 (2012) 等对中文期刊、微博新闻进行分析，证实了词共现方法的有效性。

与此同时学者们也基于互联网大数据，针对突发公共卫生事件等容易使公众及网民产生负面情绪事件进行了研究，邢云菲等 (2018) 基于情感极性及情感强度理论证实了负向情感很容易到达极端值，何高奇等 (2018) 也通过构建群体情绪感染模型证实了少数个体的恐慌情绪会蔓延

成为大规模的公众恐慌情绪。因此有学者将研究聚焦于缓解大众过度恐慌情绪的研究领域，杨阳等 (2020) 通过构建以网民和政府为代表的动态博弈模型证实了政府、媒体等权威官方组织的不同表态会诱发网民不同的情绪化行为和态度。安璐等 (2017) 利用社会网络情感网络图谱分析突发公共卫生事件中各类利益相关者的情感状态和分布，发现在事件爆发期和蔓延期，主流媒体和自媒体对普通群众的情感影响较大；在衰退期，政府人员和医护人员参与的增加对群众情感影响较大。由此可见发生重大突发性公共卫生事件时政府及媒体对公众情绪进行实时监测，做出合理的应急策略，对减少人员不必要伤亡、缓解公众恐慌情绪使非常必要的。

2 方法设计

2.1 数据采集

本次研究数据通过百分点互联网数据采集系统获取，两类数据源分别来自互联网公开确诊病例数据和微博数据。百分点互联网数据采集系统已覆盖 16000 余家资讯站点，近 4000 家论坛社区，以及主流微博、微信公众号，通过成熟的文本规则引擎，实现对信息内容的精准挖掘，提取与用户相关的价值信息。同时使用大量的人工标注的语料作为训练集，通过提取文本特征，构建分类器来实现情感的分类。

1. 公开数据
   基于百分点互联网数据采集系统爬取全国各地卫健委疫情数据，其中新增确诊、累计确诊的数据颗粒度精确到地级市与市改区。
2. 微博数据
   基于百分点互联网数据采集系统爬取疫情期间的新浪微博评论，同时经过文本去重（相同文本）、限制文本长度（最低 1 个中文字符以上）、剔除无意义（空白，纯符号）数据等预处理手段过滤，本次疫情期间共获取有效评论共计 647323 条，数据采集时间范围是 1 月 15 至 3 月 1 日。

2.2 疫情时间划分

目前全国共有地级市与市辖区共计 499 个，大约覆盖全国 76% 人口，通过分析疫情期间每日新增确诊病例的区域数量趋势，发现新增确诊病例的区域数量趋势与疫情发展趋势基本一致，从空间角度反映了全国的疫情发展状况，本文依据图3所示，将本次疫情分为 5 个阶段，分别为前期、发展期、平台期、回落期、稳定期，具体划分如下：

疫情前期：1 月 15 日-1 月 20 日，每日新增确诊的区域较少；

疫情发展期：1 月 21 日-1 月 27 日，每日新增确诊的区域急剧增加；

疫情平台期：1 月 28 日-2 月 8 日，每日新增确诊的区域保持在较高水平；

疫情回落期：2 月 9 日-2 月 25 日，每日新增确诊的区域下降；

疫情稳定期：2 月 26 日-3 月 1 日，每日新增确诊的区域保持在较低水平；

由图3所知，在疫情前期，1 月 15 日-1 月 20 日，每日新增确诊病例的区域较少，说明疫情此时正处于潜伏阶段，社会并未进入防疫阶段，1 月 21 日进入发展期后，随着检测力度的提高，新增确诊病例的区域数量快速攀升，并于 1 月 28 日进入平台期，每日新增确诊的地区数量均在

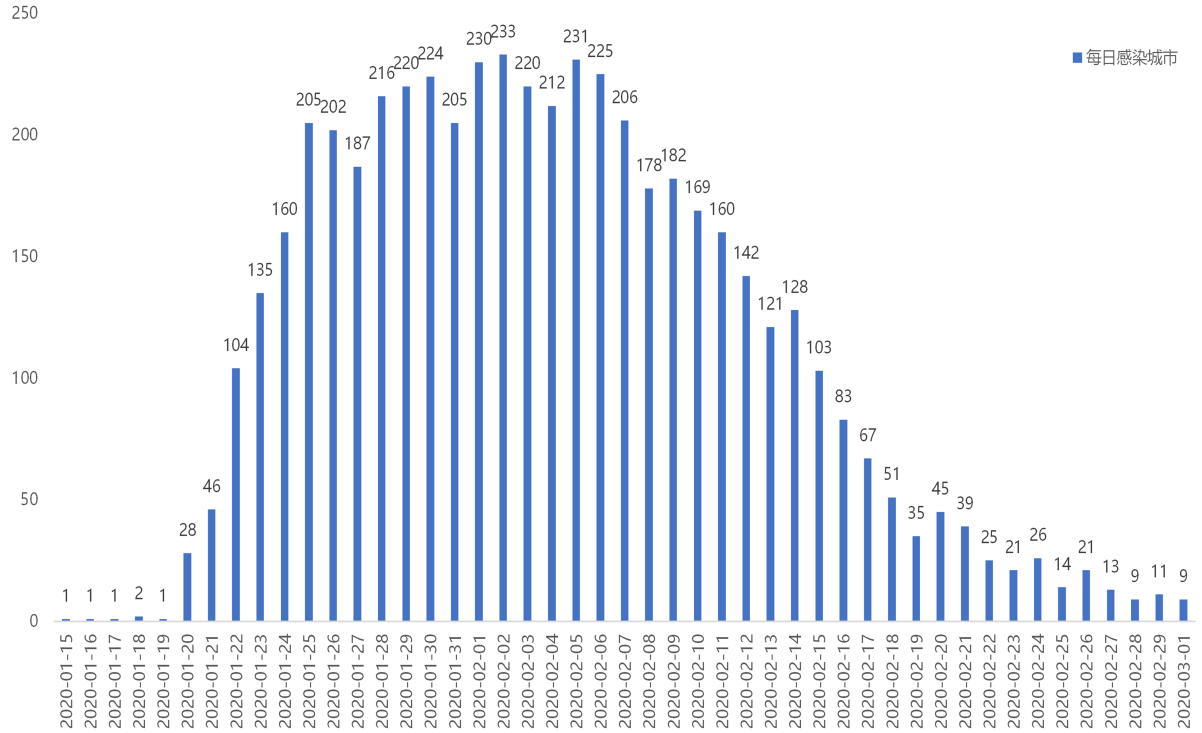

图 3: 疫情期间每日新增确诊病例的区域数量

200 以上，全国近半数的区域均出现了疫情，说明疫情进入战略对峙阶段，处于防疫最关键期，疫情于 2 月 08 日左右进入明显的回落期，随着前期防疫手段有效执行，在回落期全国各区域的疫情均得到初步控制，每日新增确诊的区域逐日减少，疫情于在 2 月 26 日进入稳定期，在这个阶段，全国每日新增确诊病例的区域数量均在较低水平，说明疫情已经得到控制，各方面处于恢复阶段。

2.3 情感的分类

本文基于表1对微博公众评论数据进行情感分类。目前心理学界对于中文情感的划分主要有 7 大类，分别是乐、好、怒、哀、惧、恶、惊。本文依据实际情况，添加 ‘‘中’’ 作为补充的情感大类，主要代表中立的言论，表1是关于更新后的情感大类解释：

2.4 朴素贝叶斯模型情感分类概述

朴素贝叶斯模型 (刘艳文等, 2020)（Naive Bayes Model），是一种基于贝叶斯定理与特征条件独立假设的分类方法，能处理多分类预测任务，在情感分类领域都有广泛的应用。

本次朴素贝叶斯模型的实现过程：

1. 随机选择已完成数据预处理的评论文本数据，基于人工标记情感类别，训练样本为 1 万条有效评论，标注比例约 1.5%；
2. 将训练文本数据进行分词处理，并且组成一个包含所有词语的词袋；

表 1: 情感分类表

编号	情感大类	情感分类	编号	情感大类	情感分类
1	乐	快乐	12	惧	慌张
2		安心	13		恐惧
3	好	尊敬	14		羞愧
4		赞扬	15	恶	烦闷
5		相信	16		憎恶
6		喜爱	17		贬责
7	怒	愤怒	18		嫉妒
8	哀	悲伤	19		怀疑
9		失望	20	惊	惊讶
10		内疚	21	中	建议
11		思念	22		设想

3. 将每条文本训练数据转换成词袋集合长度的向量；
4. 构建朴素贝叶斯模型，基于词向量计算每条评论属于某类情感的概率。

2.5 潜在狄利克雷分布模型话题聚类概述

潜在狄利克雷分布 (林江豪等, 2012) 是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。在本中潜在狄利克雷分布模型将用于对公众评论进行关键词抽取，完成话题聚类，洞察舆论焦点。

2.6 词共现网络模型可视化概述

共现词是在大量文本中经常搭配、共同出现的词汇，某词的共现词的集合在某种程度上描述了该词的语义环境，共现词之间的关联强度也在一定程度上反应了这些词所代表的语义之间的关联强度。因此，以大规模语料库为基础，构建词之间的共现网络，分析他们之间连接强度，是一种可行的分析词所代表的语义关联情况的定性定量化方法。本文将使用词共现模型对公众评论进行可视化分析，基于可视化挖掘热点话题事件。

2.7 分析过程

本次课题研究主要实现步骤如下：

1. 基于百分点大数据爬取系统收集疫情期间新浪微博公众评论，并输出结构化文本数据；
2. 基于原始数据进行预处理，主要处理手段为文本去重（相同文本）、限制文本长度（最低 10 个中文字符以上）、剔除无意义（空白，纯符号）数据，对评论时间戳进行年／月／日格式转换，最终剩余 647323 条有效评论；

表 2: 评论训练集人工标注情感标签示例

评论	日期	情绪
天呐！要注意安全啊！白衣天使！	1 月 20 日	惊
希望这些医护人员都平安。	1 月 20 日	好
请不要再伤害医护人员请不要再伤害医护人员请不要再伤害医护人员。	1 月 20 日	怒
钟南山访问要点： 1、确认可以人传人，基于武汉和广东案例。 2、有医务人员被传染。 3、源头目前不清楚，但可能是竹鼠、獾这种野生动物（尽量别去碰野味）。 4、有发热及时就医。 5、买不到 N95，普通口罩也可以起到阻止飞沫传播的作用（该戴口罩戴口罩）。	1 月 20 日	中
前有肆虐的病毒，后有病人的砍刀，为所有奋斗在前线的医护人员祈祷。	1 月 20 日	好
别扩大范围了，在武汉别出了，没在武汉别进了。	1 月 20 日	惧
希望试剂可以发到全国各个城市，不论发达还是落后。 越是落后的省份危机越大。大家防范意识弱或消息闭塞，医疗不发达无法确诊较多。 希望国家给每一个省份发相关试剂，拜托! 很多农民工都是从发达城市包括武汉返乡过年，希望重视!	1 月 20 日	惧
我是武汉人。我今天出门的时候大多数人都带了口罩，我的家人同学也都是 尽量不出门的。我们也不想发生这种事。所以那些地域黑的给我闭嘴!	1 月 20 日	怒
只有钟南山院士出来说话大家才相信才踏实！希望我们医护人员平安！	1 月 20 日	好
医护人员辛苦了尽量减少出入人多的场所吧.	1 月 20 日	好
这个时候我们的医护人员冒生命危险冲在第一线，向他们致敬！ 不希望再出现伤害医生的事件了，没有他们，我们只剩下绝望！ 17 年前的非典我们都战胜了，我们现在医学更强大了，要相信 国家和医护人员，按照专家的建议做好防御。	1 月 20 日	好
这个病是跨年前一天医院内部就通报了，现在快一个月了， 学生工人该回家的都回的差不多了，这一点才是最可怕的。	1 月 20 日	惧
经历过 SARS 的人现在听到钟南山这三个字心里都觉得踏实	1 月 20 日	好
病毒肆虐，医生被传染，然后医生被砍了！北京朝阳医院医生又被砍了， 被砍的是眼科某一细分领域的 top10，文章呢？安检进医院呢？ 我都开始担心我做医生的父母了!!! 他们是去治病救人的，他们有生命危险了！	1 月 20 日	惧
一开始就是很明显的人传人迹象，非得大面积扩散了，才承认， 现在就想问问这责任谁担。	1 月 20 日	恶
向白衣天使致敬。你们辛苦了。早日康复。	1 月 20 日	好
希望所有人都可以平安善有善报。	1 月 20 日	好
我真的希望武汉人最近不要外出，不是歧视。	1 月 20 日	惧
武汉市，湖北省，一直拖！到了今天春运高峰才引起中央重视， 责任很大。节后估计才是高峰期。	1 月 20 日	怒

3. 对随机抽取的 1 万条有效评论文本进行人工标注情绪标签，建立模型训练集；
4. 基于训练集搭建朴素贝叶斯情感预测模型，对剩余全量评论进行情感预测；
5. 基于潜在狄利克雷分布模型对公众评论数据进行关键词抽取，完成话题聚类；
6. 对公众评论进行中文分词，计算词频，筛选有效评论词；
7. 基于有效评论词通过计算机软件搭建词共现网络，通过可视化形式挖掘热点事件。

3 数据分析

3.1 疫情期间情绪演化

基于朴素贝叶斯情感预测模型获取全量文本情感预测标签，结合时间维度通过数据透视表得到表3分析结果，由表3情感占比表与图4情感趋势图可知，随着疫情的不断发展，公众的情绪发生了明显变化，疫情由前期发展到平台期，公众情绪中偏正面的“好”、“乐”的情感占比持续下降，较为中立的“中”情绪也持续走低，而在较偏负面情绪的“惧”、“哀”、“恶”、“惊”、“怒”整体占比则呈现上升趋势，其中公众“惧”的情绪占比在发展期达到高峰，占比为 38.2%，说明在疫情发展期内，公众对于疫情爆发感到焦虑与恐慌，而在进入平台期后，公众“怒”的情绪则达到 27.6%，说明在疫情进入平台期后，疫情衍生的相关事件易引起舆情，引发公众愤怒的情绪。

疫情由平台期发展至回落期，偏正面的“好”、“乐”情绪回升明显，分别达到 27.3% 和 16.5%，其余偏负面的情绪“惧”、“哀”、“恶”、“惊”、“怒”占比则显著降低，说明随着疫情的回落，疫情正面信息逐渐增多，疫情压力对于公众逐渐降低，通过情感分布可知公众的信心已有所提升，社会舆论偏正向发展。

疫情由回落期进入稳定期，国内疫情话题不再是公众首要关注，国外疫情与国内其他热点占据舆论场，因此各项情绪占比分布已趋于合理。

本次课题研究，通过对各时期的公众评论进行汇总并进行主题抽取，抽取原则为系数排名前 6 的话题词，如表4所示，通过话题结构挖掘出各时期的公众重点关注事件，实现对疫情各期间的话题演化的追踪，洞悉舆论走向。

在疫情前期，由话题抽取结果得知，公众情绪普遍对疫情的应对保持乐观，话题主要集中在对医护人员的祝福、相信钟南山等。

表 3: 疫情期间公众情感占比表

情绪/时期	前期	发展期	平台期	回落期	稳定期
好	37.8%	15.6%	8.5%	27.3%	16.6%
乐	2.5%	0.1%	0.0%	16.6%	12.7%
中	28.9%	11.9%	8.9%	13.5%	10.6%
惧	13.7%	38.2%	16.1%	8.4%	7.6%
哀	7.0%	10.5%	17.6%	10.2%	6.4%
恶	5.2%	1.3%	8.6%	1.2%	2.3%
惊	3.4%	7.8%	12.7%	7.2%	15.6%
怒	1.5%	14.7%	27.6%	15.6%	28.2%

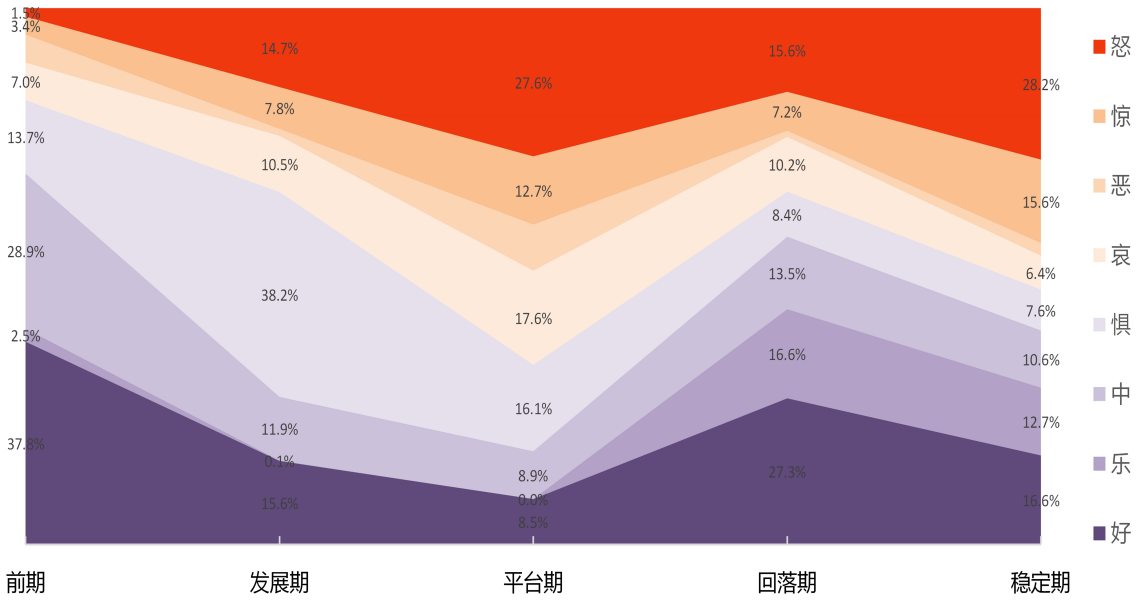

图 4: 疫情期间情感趋势

表 4: 疫情期间公众评论潜在狄利克雷分布模型话题抽取结果

时间段	话题结构组成（权重前 6）
前期	医护 × 0.096 + 平安 × 0.086 + 致敬 × 0.076 + 钟南山 × 0.072 + 相信 × 0.065 + 武汉 × 0.054
发展期	武汉 × 0.12 + 口罩 × 0.098 + 野生 × 0.086 + 物资 × 0.074 + 响应 × 0.68 + 加油 × 0.65
平台期	李医生 × 0.142 + 红十字 × 0.114 + 物资 × 0.098 + 公道 × 0.892 + 武汉 × 0.075 + 大理 × 0.64
回落期	复工 × 0.078 + 延期 × 0.072 + 口罩 × 0.065 + 加油 × 0.063 + 致敬 × 0.045 + 中医 × 0.053
稳定期	外国 × 0.075 + 条例 × 0.065 + 反对 × 0.064 + 调查 × 0.045 + 武汉 × 0.36 + 北京 × 0.34

在疫情发展期，公众的话题则主要集中在武汉的疫情、口罩物资、谴责乱吃野生动物等，说明疫情持续发展后，因新冠肺炎需要口罩进行个人防护，公众提高了对口罩物资的关注，同时因疫情对生活造成了影响从而产生了对疫情源头的谴责情绪。

在疫情平台期，由疫情衍生了多起舆情事件，由话题抽取得知主要因 ‘‘李文亮牺牲事件’’，‘‘红十字物资分配’’，‘‘大理口罩事件’’ 事件导致舆论爆发。说明在疫情防疫的关键期内，各类关于疫情的社会矛盾点会集中爆发，易引起影响力较大的舆情事件，需重点关注，引导舆论走向。

在疫情回落期，随着疫情的减缓，公众对于复工、开学延期等民生话题关注度提升，对于急需的口罩关注度也进一步提高，同时在回落期，关于疫情的利好消息逐渐增多，公众对于疫情的

信心提升，关于加油祝福的话题也占据一定的声量。

在疫情稳定期，公众对于国内疫情的关注逐步降低，相反疫情之外 ‘‘外国人永久居住条例’’ 占据了较高的话题量，说明疫情后期，公众的注意力已逐步从疫情转移至其他热门话题。

3.2 疫情期间重大舆情事件挖掘与情绪分析

由图5公众评论词共现网络图所示，公众评论主要集中在 5 大集合，由评论词集合的可视化，发现疫情期间公众主要的话题讨论集中在 ‘‘李医生’’、‘‘封城’’、‘‘北京’’、‘‘红十字’’、‘‘大理’’、‘‘口罩’’、‘‘物资’’、‘‘医院’’、‘‘武汉’’、‘‘湖北’’ 等关键词相关的热点事件，说明在疫情期间，以上关键词所对应的话题事件，是本次疫情期间主要的舆情事件。

主要的舆情事件：‘‘李文亮牺牲事件’’、‘‘红十字物资分配’’、‘‘大理截取口罩’’、‘‘武汉封城’’、

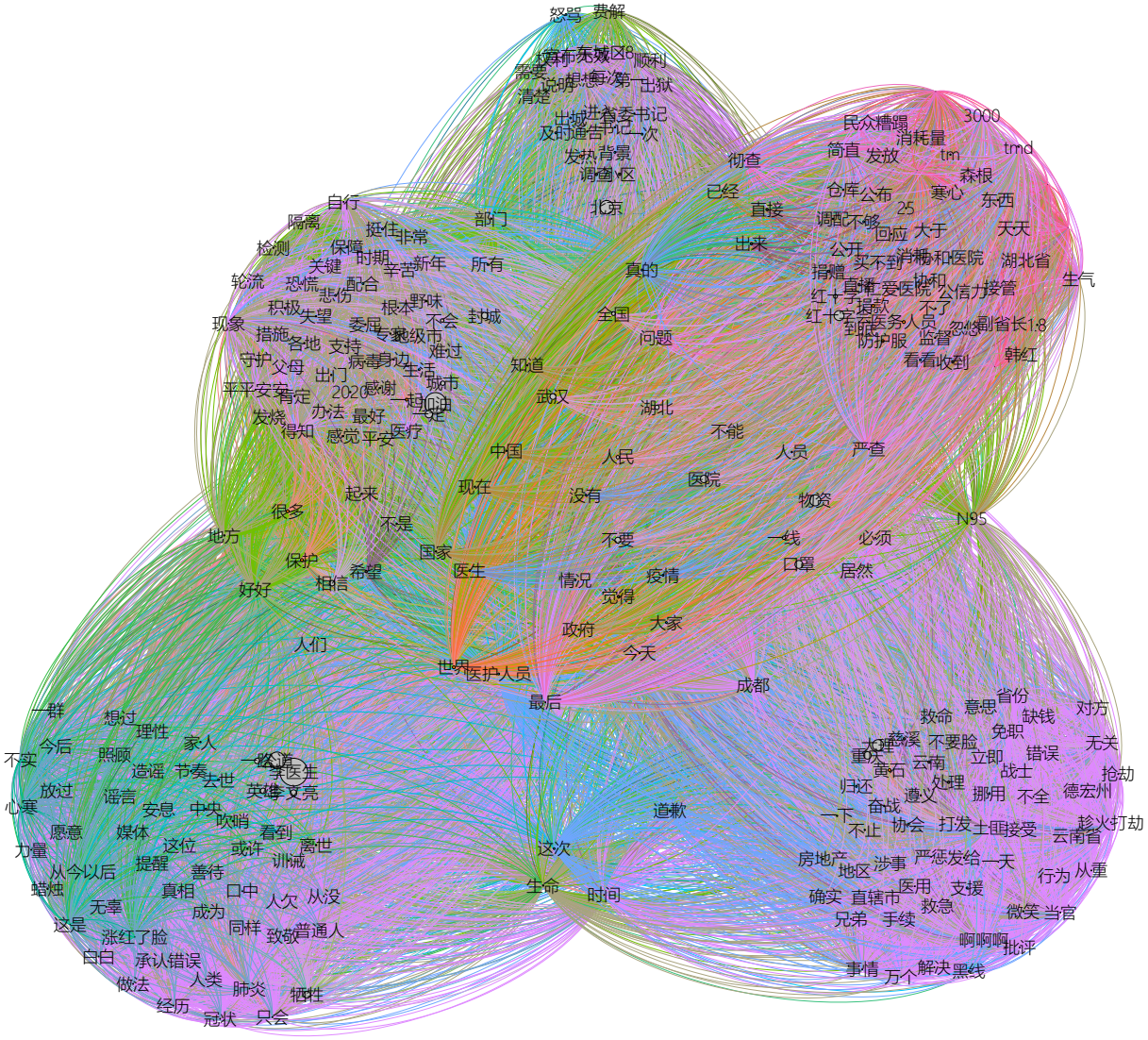

图 5: 疫情期间部分公众评论词共现网络

表 5: 重大舆情事件的事件跨度示例

事件/筛选条件	时间跨度	关键词
李文亮牺牲事件	2 月 6 日-2 月 8 日	李文亮、李医生、牺牲
红十字物资分配	1 月 3 日-2 月 1 日	红十字、物资、湖北
大理截取口罩	2 月 4 日-2 月 6 日	大理、口罩、云南
武汉封城	1 月 23 日-1 月 24 日	武汉、封城
离汉女子进京	2 月 26-2 月 27 日	女子、北京、湖北

表 6: 重大舆情事件的公众情绪分布

事件/情绪	好	乐	中	惧	哀	恶	惊	怒
武汉封城	43.2%	1.4%	12.1%	7.5%	1.4%	0.1%	17.5%	16.8%
红十字物资	3.5%	0.1%	2.6%	1.2%	9.0%	3.5%	25.6%	54.5%
李文亮牺牲	1.5%	0.0%	2.6%	1.2%	42.5%	3.1%	14.6%	34.5%
大理截扣口罩	1.3%	0.1%	2.6%	4.3%	6.5%	6.1%	30.6%	48.5%
离汉女子进京	2.4%	0.2%	11.5%	5.0%	19.5%	3.4%	23.5%	34.5%

‘‘离汉女子进京’’。

基于词共现图所挖掘的主要舆情事件，通过开源资料对舆情事件进行背景调查，基于重大舆情事件的起始时间与其重点的关键词等筛选条件获取针对化的评论数据集，课题研究通过筛选共筛选出 5 份评论数据集，如表5所示，分别代表 ‘‘李文亮牺牲事件’’、‘‘红十字物资分配’’、‘‘大理截取口罩’’、‘‘武汉封城’’、‘‘离汉女子进京’’ 等舆情事件，最后通过数据透视表对各舆情事件进行情绪分布的分析。

由表6重大舆情事件的公众情绪占比得知，除 ‘‘武汉封城’’ 事件偏正面外，其余在疫情期间发生的舆情事件均引起了较为严重的负面情绪，其中 ‘‘红十字物资事件’’ 的 ‘‘怒’’ 情绪最高，说明公众对于红十字的物资分配十分不满。在 ‘‘李文亮牺牲’’ 事件中公众 ‘‘哀’’ 的情绪最高，其 ‘‘怒’’ 的情绪也相对较高，说明公众对于 ‘‘李文亮牺牲’’ 主要表达的是哀思与不满。‘‘大理截扣口罩’’ 事件在 ‘‘惊’’、‘‘怒’’ 方面占比均较高，说明工作对于大理单方面截扣重庆口罩的行为表示不满与震惊。在 ‘‘离汉女子进京’’ 事件中，公众情绪集中在 ‘‘怒’’、‘‘惊’’，除此之外 ‘‘哀’’ 的情绪也偏高，说明公众情绪除了不满与震惊外，还包含对事件的失望心情。

4 结果分析与讨论

4.1 总结与讨论

本文基于朴素贝叶斯、词云、潜在狄利克雷分布和词共现社群挖掘方法的情感分类组合模型分析了主流媒体和疫情期间确诊人数数据，按照疫情期间确诊人数划分了 5 个阶段，分别分析了民众情绪在新型冠状病毒疫情期间的衍变趋势。

得到结论如下：

1. 随着疫情阶段的推移，偏正面的公众情绪持续下降，较偏负面情绪则呈现上升趋势，因此在

重大安全事件发生的中后期政府和有关部门更应关注公众情绪的演化方向。

2. 公众负面情绪大多来源于物资短缺、人员伤亡和生活不便，因此政府和相关部门可根据地区的实际情况进行实时监控，以尽可能减少不公平事件和哄抬物价等会造成公众负面情绪的情况产生。
3. 疫情期间主要的舆情事件中除 ‘‘武汉封城’’ 事件引起公众的情绪偏正面外，其余在疫情期间发生的舆情事件均引起了较为严重的负面情绪。

由此可见对新型冠状病毒疫情期间公众情绪演变趋势与话题挖掘进行研究，不仅为进一步的疫情防控工作奠定了基础，也为政府及相关部门在类似重大突发性公共卫生事件发生后进行应急决策提供数据和理论支持。

4.2 研究意义

2019 新型冠状病毒因具有潜伏期长、初期症状识别性较低以及致死率较高等显著特点，加重了疫情地区民众对疫情的治愈性的担忧，这种负面情绪经过网络传播大范围蔓延，甚至衍生了恐慌以及愤怒、憎恶、指责的情绪，给疫情防控也带来了一定程度的困难。尽早进行心理干预与支持，通过政府、媒体等具有影响力的官方渠道提供可靠信息均可提高公众安全感，从而减少因群体恐慌情绪而造成的不必要的人员患病和公共医疗资源的浪费。

4.3 实践和管理意义

近年来，国内外突发的疫情公共卫生事件多发，国外的如埃博拉病毒疫情、美洲的塞卡病毒，国内的如非典型肺炎、H7N9 禽流感病毒、甲型 H1N1 流感等，突发的疫情对国家经济发展、人民的生活造成了巨大的影响。这一次突发的新型冠状病毒疫情，再次挑动着全世界人民的神经，严重威胁人民的生命健康。随着互联网新媒体技术的不断进步，公众借助微信、微博、论坛等随时随地收到最新消息同时对信息进行传播、表达公众情感，但是微博中信息的来源、内容、形式纷繁多样，而且公众越来越希望信息透明化，但是当人们面临着大量的信息，人们很容易在信息中迷失了自己，所以很容易引起社会过度恐慌，不利于政府处理危机 [21]。通过本文分析判断整体群体的情感发展趋势，为建立突发的公共卫生事件预警和管理策略提供理论依据，有助于政府有效的引导网民情绪，维护社会秩序，针对于疫情发生的不同阶段，制定更有针对性的有效应急方案。

4.4 未来研究方向

针对新型冠状病毒疫情，本文主要以国内的微博进行情感分析，探索突发公共卫生事件中公众的情感状态与情感传播规律，本文结合疫情演化的生命周期分析各时期公众的情感类型和情感强度，并进一步分析随疫情话题的演化公众的情感演化过程。本研究为进一步研究互联网情感信息传播规律及特征提供参考依据，另外，有益于相关部门实时了解公众情感走势，及时控制情感传播源，采取有效的情绪引导措施，为突发公共卫生事件应急管理提供了支持，从而避免群体情绪极化现象的发生，以维持社会秩序的稳定。当然，本文还存在一些局限，仅通过微博无法全面获取公众的立场和情感态度。因此，在后续研究中针对具体事件还应搜集并整合来自多渠道的

相关数据，以更全面反映舆情发展与民众情绪。因本文主要探讨新型冠状病毒疫情期间公众情绪衍变趋势的研究，因此在未来的研究中，还需深入分析情感图谱的网络特征，探索影响用户情感表达与传播的因素，并建立社会网络情感传播模型，以期更好地服务于突发公共卫生事件的预警和应急管理。

致谢

本研究受光华思想力课题“基于 AI 自然语言处理和行业知识图谱的智能化市场聆听”支持。

参考文献

中央广电总台国际在线, 2020. 北大国发院-百分点联合发布疫情数据报告：民众对于战胜疫情非常有信心. http://www.peopleweekly.cn/html/2020/qynews_0218/4922.html.

何高奇, 边晓晖, 孙菲, 卢兴见, 2018. 基于传染病机制的突发事件下群体情绪感染模型. 华东理工大学学报 (自然科学版), 44(6): 909-917+949.

刘艳文, 魏赟, 2020. 基于 LDA 主题模型的情感分析研究. 电子科技(7): 1-6.

刘韩松, 2013. 基于文本挖掘及情感分析的社区负面舆论传播预测模型. 计算机安全(12): 7-11.

孟雪井, 孟祥兰, 胡杨洋, 2016. 基于文本挖掘和百度指数的投资者情绪指数研究. 宏观经济研究(1): 144-153.

安璐, 欧孟花, 2017. 突发公共卫生事件利益相关者的社会网络情感图谱研究. 图书情报工作, 61(20): 120-130.

张海涛, 刘雅姝, 张枭慧, 宋拓, 2019. 基于模块度的话题发现及网民情感波动研究——以新浪微博“中美间贸易摩擦”话题为例. 图书情报工作, 63(4): 6-14.

朱霞, 苗丹民, 罗正学, 2005. 突发性事件中不同角色公众情绪变化的研究. 中国行为医学科学, 14(4): 351-352.

杨阳, 王杰, 2020. 情绪因素影响下的突发事件网络舆情演化研究. 情报科学, 38(3): 35-41+69.

林江豪, 阳爱民, 周咏梅, 陈锦, 蔡泽键, 2012. 一种基于朴素贝叶斯的微博情感分类. 计算机工程与科学, 34(9): 160-165.

王一博, 郭鑫, 王继民, 2014. 基于词共现的大数据研究主题分析. 图书馆论坛, 34(8): 96-102.

王晰巍, 张柳, 文晴, 王楠阿雪, 2018. 基于贝叶斯模型的移动环境下网络舆情用户情感演化研究——以新浪微博“里约奥运会中国女排夺冠”话题为例. 情报学报, 37(12): 1241-1248.

赵文明, 宋述慧, 陈梅丽, 邹东, 马利娜, 马英克, 李茹姣, 郝丽丽, 李翠萍, 田东梅, 唐碧霞, 王彦青, 朱军伟, 陈焕新, 章张, 薛勇彪, 鲍一明, 2020. 2019 新型冠状病毒信息库. 遗传, 42(2): 212-221.

赵文清, 侯小可, 2012. 基于词共现图的中文微博新闻话题识别. 智能系统学报, 7(5): 444-449.

邢云菲, 王晰巍, 韦雅楠, 王铎, 2018. 新媒体环境下网络舆情用户情感演化模型研究——基于情感极性及情感强度理论. 情报科学, 36(8): 142-148.

陈茜, 连婉琳, 2015. 基于文本挖掘技术的互联网股票新闻的情感分类. 中国市场(24): 234-235.

高文斌, 陈祉妍, 王一牛, 史占彪, 杨小冬, 张建新, 2003. SARS 疫情期间公众心态影响及变化趋势分析. 中国心理卫生杂志, 17(9): 594-596.

Han X, Wang J, 2019. Using social media to mine and analyze public sentiment during a disaster: A case study of the 2018 Shouguang city flood in China. ISPRS International Journal of Geo-Information, 8(4): 185.

Liu IL, Cheung CM, Lee MK, 2010. Understanding Twitter usage: What drive people continue to Tweet//Pacific Asian Conference on Information System 2010 Proceedings: volume 92. 928-939.

Moreno-Sánchez I, Font-Clos F, Corral Á, 2016. Large-scale analysis of Zipf's law in English texts. PloS One, 11(1): e0147073.

Rajput NK, Ahuja B, Riyal MK, 2019. A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita. Pramana, 92: 60.

Rajput NK, Grover BA, Rathi VK, 2020. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv preprint https://arxiv.org/abs/2004.03925.

World Health Organization, 2020. Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.