Hans Journal of Data Mining 数据挖掘, 2022, 12(2), 114-122
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.122013

Hans汉斯

基于机器学习的新型冠状肺炎的舆情分析

谢　婷，罗　清

广西科技大学理学院，广西　柳州

收稿日期：2022年3月1日；录用日期：2022年3月30日；发布日期：2022年4月8日

摘　要

新冠疫情的爆发和肆虐引起群众关注，互联网上的相关话题不断攀升。如何利用计算机方法和数据分析算法准确地识别热点新闻和疫情主题，挖掘民众关注的话题，分析舆论走势，显得至关重要。本文提出一种基于GSDMM主题挖掘的“新冠肺炎疫情”舆情分析方法，利用数据预处理、特征提取、词云可视化技术挖掘目标数据的热点主题，再采用GSDMM主题模型、聚类分析对目标数据进行分析挖掘。通过深入进行了面向人民网的GSDMM短文本聚类算法研究，得到大家都一直十分关心中国和世界的疫情形势和经济形势的信息。此次肺炎疫情热点主题包括疫情、防控、工作、肺炎、患者等。

关键词

GSDMM，主题模型，新冠疫情，聚类算法

Public Opinion Analysis of Novel Coronavirus Pneumonia Based on Machine Learning

Ting Xie, Qing Luo

School of Sciences, Guangxi University of Science and Technology, Liuzhou Guangxi

Received: Mar. 1st, 2022; accepted: Mar. 30th, 2022; published: Apr. 8th, 2022

Abstract

The outbreak and ravages of the new crown epidemic have aroused the attention of the masses, and related topics on the Internet have continued to rise. How to use computer methods and data analysis algorithms to accurately identify hot news and epidemic topics, dig out topics of public concern, and analyze the trend of public opinion is of great importance. In this paper, a “new crown pneumonia epidemic” public opinion analysis method based on GSDMM theme mining is

文章引用: 谢婷, 罗清. 基于机器学习的新型冠状肺炎的舆情分析[J]. 数据挖掘, 2022, 12(2): 114-122.
DOI: 10.12677/hjdm.2022.122013

proposed, which uses data preprocessing, feature extraction, and word cloud visualization technology to mine the hot topics of target data, and then uses GSDMM theme model and cluster analysis to analyze and mine target data. Through the in-depth research of the GSDMM short text clustering algorithm for the People's Network, we have obtained information that everyone has always been very concerned about the epidemic situation and economic situation in China and the world. The hot topics of the pneumonia epidemic include epidemic situation, prevention and control, work, pneumonia, patients, etc.

Keywords

GSDMM, Topic Model, New Crown Epidemic, Clustering Algorithm

1. 引言

2019 年 12 月以来，新型冠状病毒肺炎的快速暴发引发了广泛的社会关注，新冠疫情传播之迅速，让人猝不及防防。为了阻止疫情蔓延，政府不得不严厉令停工停产，在给经济造成重大打击的同时，互联网上关于疫情铺天盖地的新闻也紧紧的揪着了所有居家的群众的心，引发了群众的日益激烈的讨论和关注，也给网络舆情分析带来了极大的挑战。如何利用计算机技术挖掘群众的关注的潜在主题和舆情走向变得至关重要，面对人民网发布的疫情新闻，本文提出一种基于 GSDMM 主题挖掘的“新冠肺炎疫情”舆情分析方法，有别于传统的 LDA 主题模型，GSDMM 在其输出的主题的连贯性和一致性方面有更大的优势。GSDMM 旨在众多文本中挖掘潜在的核心话题，了解群众对于疫情“舆情事件”的关注关心点，为实现正确引导舆论回应社会关切提供参考。

近些年，国内学者致力于舆情分析研究，并提出了相关分析方法。杨秀璋等[1]通过构建 LDA 主题模型分析挖掘新闻文本主题以及利用 snowNLP 库和贝叶斯模型进行情感分析，以此判断舆情事件的主题以及民众的情感倾向。基于刘惠等[2]通过构建 LDA 模型对豆瓣网站上《少年的你》的短评文本电影进行舆情分析，识别出热点话题。杨雪寒等[3]基于情感特征提出一种基于附加特征方法的文本挖掘来实现对医院的舆情监测以及情感文本挖掘，结果表明准确性得到提高并减少了实现时间。基于彭浩等[4]针对舆情文本特征的情况，结合主题及趋向性针对微博网络舆情进行分析，能够有效挖掘其中潜在的主题。岳宗朴等[5]基于微博数据挖掘的“新冠疫情”评论进行文本分析、文本语句中心词分析，绘制词云图，得到微博舆情良好、网民对于“新冠肺炎疫情”稳定在以积极情绪为主导的态势。黄渤等[6]通过 TF-IDF 关键特征词提取、OLDA 模型主题词演化分析，构建评论集词向量模型，最后使用 K-means 对主题进行聚类，对聚类结果通过词性标注进行分析，实验表明，该方法能够实现有效检测到文本主题。

2. 人民网新闻文本的主题分析

本文旨在分析“新冠肺炎疫情”的热点主题。其算法总体流程如下：

1) 通过 Python 技术爬取人民网“新冠肺炎疫情”相关的新闻，包括新闻标题、新闻内容、发布时间、新闻来源等信息。

2) 对所爬取的信息数据进行数据预处理，包括重复值处理、缺失值处理、分词、去除停用词、词性

标注，再将处理后的数据存入文本文件中。

3) 分析包括三个核心模块：数据预处理、GSDMM 模型主题分析、主题词层次聚类分析，最终得出实验结论。

2.1. 数据准备及预处理

本文主要选取新型冠状病毒新闻数据作为研究对象，本次爬虫目标网站是人民新闻网。通过使用 Python 编程可以实现新型冠状病毒新闻数据的爬取。本文共收集了 142 篇发布在人民网上有关疫情的新闻，发布时间如表 1 所示，从 2020 年 2 月 1 日武汉封城之后直到年末 2020 年 12 月 30 日，几乎贯穿了整个 2020 年。

Table 1. People's news network novel coronavirus news statistics table
表 1. 人民新闻网新型冠状病毒新闻统计表

发布时间	新闻数量
2020-02-01 至 2020-02-29	87
2020-03-01 至 2020-03-31	32
2020-04-01 至 2020-04-30	10
2020-05-01 至 2020-05-31	1
2020-06-01 至 2020-06-30	3
2020-07-01 至 2020-07-31	2
2020-08-01 至 2020-08-31	1
2020-09-01 至 2020-09-30	1
2020-10-01 至 2020-10-31	2
2020-11-01 至 2020-11-30	2
2020-12-01 至 2020-12-30	1
总计	142

这些海量数据进行挖掘前首先要对它们进行处理，首先对原始数据进行缺失值处理和重复值删除，然后利 Python 调用 Jieba 库进行新闻正文分词，并导入关键词和停用词词典完成停用词过滤和数据清洗，最后要为分词结果进行词性标注。经过数据预处理能够得到进行建模分析所需要的 481114 个分词。

1. 去除停用词后的部分分词结果

'概述', '疫情', '防控', '中', '权威', '信息', '大众', '变化', '丁香', '园', '医生', '团队', '天', '时间', '制作', '凌晨', '正式', '上线', '新冠', '肺炎', '地图', '产品', '国家', '卫生', '健康', '委员会', '中国', '疾病', '预防', '控制中心', '全国', '大部分', '省', '自治区', '直辖市', '医疗卫生', '机构', '媒体', '发布', '数据', '地区', '确诊', '疑似', '重症', '死亡', '治愈', '病例', '作出', '梳理', '汇总', '流行病学', '原理', '发展趋势', '分布', '情况', '可视化', '呈现', '扩展', '全球', '蔓延', '做法', '动态数据', '收集', '技术', '人工', '分析', '同步', '整理', '官方', '公布', '保证', '真实性', '可靠性', '基础', '第一', '建立', '新型', '冠状病毒', '实时', '动态', '更新', '页面', '新增', '趋势', '图', '累计', '病死率'

2. 词性标注部分结果

('冠状病毒', 'l'), ('冠状病毒', 'l'), ('新型', 'b'), ('冠状病毒', 'l'), ('病毒', 'n'), ('参与', 'v'), ('研究', 'vn'), ('华南农业大学', 'nt'), ('教授', 'n'), ('沈永义', 'nr'), ('媒体', 'n'), ('这', 'r'), ('批', 'q'), ('穿山甲', 'n'), ('广东', 'ns'), ('特定', 'd'), ('种群', 'n'), ('团队', 'n'), ('特定', 'd'), ('机构', 'n'), ('获取', 'v'), ('代表', 'n'), ('自然界', 'n'), ('中', 'f'), ('绝大多数', 'm'), ('穿山甲', 'n'), ('携带', 'v'), ('冠状病毒', 'l'), ('学者', 'n'), ('穿山甲', 'n'), ('参与', 'v'), ('新型', 'b'), ('冠状病毒', 'l'), ('传播', 'vn'), ('潜力', 'n'), ('英国', 'ns'), ('自然', 'd'), ('杂志', 'n'), ('网站', 'n'), ('援引', 'vn'), ('格拉斯哥', 'ns'), ('大学', 'n'), ('计算', 'v'), ('病毒学家', 'n'), ('戴维', 'nr'), ('罗伯逊', 'nr'), ('说', 'v'), ('穿山甲', 'n'), ('宿主', 'n')

2.2. 基于词云图可视化主题分析

词云图作为一种分析工具，旨在实现文本可视化，能够直观呈现关键词。本文采用 WordClound 模块中的 WordCloud 对“新冠肺炎疫情”新闻进行词云主题分析，选择前 100 个词，绘制图 1 所示的“疫情”新闻主题词云图，结果如下所示。

Figure 1. “Epidemic” news theme word cloud map
图 1. “疫情”新闻主题词云图

由图 1 可以看出，对新闻正文文本信息进行数据预处理后，分词效果较为符合预期。其中，“疫情”，“防控”，“工作”，“肺炎”，“患者”等词出现频率较高，由此可以初步判断这几个方面是新闻正文的重点，说明在这一阶段，新冠患者的确诊数量一直是群众关注的重点，以及政府部门对疫情防控工作的重视。

2.3. DMM 模型概述

DMM 模型基于的假设是每个文档由单个主题生成，而在文档的主题确认之后，文档内的词语的生成是相互独立的，而且词语生成概率与所在位置无关。DMM 首先先生成一个主题全局分布，然后生成文档主题，每篇文档对应一个主题，再去根据文档的主题抽取文档对应的所有单词，类似于 DMM 模型，GSDMM 模型同样羅每一篇文本通过 Gibbs 采样出一个主题，文本中的每个单词都是由采样得到的这个主题控制，每一个文档只赋予一个主题。

DMM 描述的文档生成过程是：

1) 采样一个 topic 的分布的超参数 theta

2) 对于 K 个 topic 中的每个 topic 生成一个 topic-word 分布的超参数 phi

3) 对于每一个文档

a) 根据 theta 采样得到一个编号为 z 的 topic

b) 根据编号为 z 的 topic-word 分布和对应的 phi 生成每个单词

相比于 LDA 的生成过程，在每次循环中，每个文档的 topic 只会生成一次。

DMM 方法概率图模型如图 2 所示。

关于 DMM 的概率图模型涉及的变量如表 2 所示。

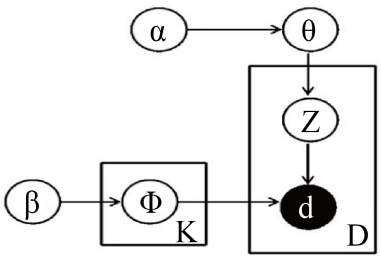

Figure 2. A graphical representation of the probability model of the DMM
图 2. DMM 的概率模型图表示

Table 2. DMM probabilistic model plots the meaning of each symbol
表 2. DMM 概率模型图各符号含义

主特征	子评论数量
β	Φ 的超参数
α	Θ的超参数
K	主题个数
D	语料库中文档个数
z	主题
d	文档
Φ	词分布
Θ	主题分布

2.4. GSDMM 模型主题聚类

主题模型是文本挖掘的重要工具，在文本挖掘领域，大量的数据都是非结构化的，很难从信息中直接获取相关和期望的信息，一种文本挖掘的方法：主题模型(Topic Model)能够识别在文档里的主题，并且挖掘语料里隐藏信息，并且在主题聚合、从非结构化文本中提取信息、特征选择等场景有广泛的用途。进而能够进行主题挖掘和舆情分析。传统的主题模型例如 pLSA、LDA，并不能很好地处理文本数据的高维和稀疏问题，不同于传统的主题模型 pLSA，LDA，由于 GSDMM 主题模型不需要文本数据的高维和稀疏问题，而直接对文档和词进行概率估计，故而可以有效解决文本数据的高维和稀疏问题。类似于 PLSA 和 LDA，GSDMM 同样可以获得每一个簇的代表词。GSDMM 是一种无监督非参数主题模型，能够自动

推测主题聚类个数，实现聚类结果的完备性和一致性。

用吉布斯采样(Gibbs Sampling)算法近似求解模型。GSDMM 模型假设文档是根据混合多项式模型产生的，并且主题和文档之间是一一对应的关系[7]。GSDMM 在聚类过程中遵循两个原则：第一，聚类后同一个主题下包含属于该主题的文本尽量多，使聚类的完备性更强；第二，聚类后同一个主题下尽可能只包含属于同一主题的文档，使聚类的一致性更强在基于 GSDMM 模型的主题聚类实验中，采用吉布斯采样(Gibbs Sampling)算法对模型进行近似求解。Gibbs 采样的物理过程，就是一个词在不同的主题上不断采样，最终得到这个词的主题分布矩阵，从而得到文档的主题分布和主题的词分布。利用 Gibbs 采样法对模型进行求解，在训练过程中采样的一篇文档属于某个主题的概率如下：

$$P\left(Z_d = Z \middle| \vec{Z}\neg d, \vec{d}\right) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_w^d} \left(n_{z,\neg d}^w + \beta + j - 1\right)}{\prod_{i=1}^{N_w^d} \left(n_{z,\neg d} + V\beta + i - 1\right)}$$

其中 $\neg d$ 表示去除当前文档 d 的信息。

将构建好的语料库输入到利用 Python 代码搭建 Gibbs Sampling DMM 模型。发现指定主题数为 3 训练效果最为有效，调用 GSDMM 主题模型对新闻分词进行训练指定主题数，指定值为 3，指定超参数 alpha，默认值为 0.1，指定超参数 beta，默认值为 0.01，指定吉布斯采样迭代次数，默认值为 2000，得到每个主题下对应的最有可能出现 10 个主题词数量，GSDMM 主题分析生成的主题和特征词如表 3 所示。

Table 3. GSDMM topic analysis topic identifiers and core phrases
表 3. GSDMM 主题分析主题标识和核心词组

主题编码	主题标识	核心主题词组(10 个)
Topic 0	民众关注政府如何统筹抓好疫情防控和对企业、单位的复工复产	企业 单位 医院 重点 情况 指导 会议 社会 传播 风险
Topic 1	各地区民众在关注新冠肺炎患者的病例变化情况的同时也在加强防护措施保护自我安全	疫情 工作 肺炎 患者 检测 病例 措施 地区 疫苗 习近平
Topic 2	民众关注国家对各省市加强疫情期间人员流动管理，层层压实责任，做好防疫工作的举措	防控 人员 病毒 新冠 国家 研究 生产 管理 责任 感染者

Topic 0 主要体现了民众关注政府如何统筹抓好疫情防控和对企业、单位的复工复产。经济是社会运作的基础。2020 年是决胜全面建成小康社会、决战脱贫攻坚之年。新冠肺炎疫情的突袭蔓延，不仅严重冲击经济社会的正常、持续运作，给各行业带来致命打击，也给深入进行广大贫困地区的扶贫工作带来许多新情况新问题新挑战。在这个背景下，政府如何做好常态化疫情防控，推动非疫情防控重点地区企事业单位复工复产，自然而然成为整个社会所有民众都关注的热点，民众甚至会产生对经济发展走“下坡路”的担忧情绪。基于此，各地政府作为舆情回应第一责任人，要快速反应、及时发声，根据处置进展动态发布信息。同时，也要做好行动，政府必须要在确保疫情防控到位的前提下，推动复工复产，恢复生产生活秩序，保障民生和促进社会和谐稳定，尽可能降低疫情冲击影响，完成经济社会发展目标任务。

Topic 1 主要体现各地区民众在关注新冠肺炎患者的病例变化情况的同时也在加强防护措施保护自我安全。新冠疫情席卷全球，也充分暴露出对卫生应急健康素养教育不够深入，公众缺乏基本的防范意识。而随着疫情的冲击和健康中国战略的实施，民众的防控意识前所未有的觉醒，对健康防护的了解程度再度成为关注热点。因而政府要聚焦公众疫情防控，关注民众防控意识，关注公众意识改变，切实回

应大众舆情。

Topic 2 主要体现了民众关注国家对各省市加强疫情期间人员流动管理，层层压实责任，做好防疫工作的举措。抗击疫情，中国在行动，政府在作为。政府的疫情防控举措“一举一措”都关系这民众的日常生活和工作，尤其我国又是一个人口大国，人员流动频繁。疫情防控工作繁杂多乱，层层压实责任势在必行。但要避免出现“层层加码”，“一刀切”的情况，这样反而会引起舆论的口诛笔伐。所以各地政府要准确理解防疫政策，主动关注舆情走向。

3. 主题词层次聚类分析

3.1. 系统层次聚类

距离分层的典型方法是层次聚类算法。层次的聚类方法(Hierarchical Clustering)是对文本挖掘得到的主题词进行系统层次聚类，其是层次化的聚类，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。最终得出来的是树形结构。

层次聚类法也称为系统层次聚类法，其想法是，首先将所有的样品都单独作为一类，然后计算任意两个类之间的距离，将其中距离最近的两个类合并为一类，同时聚类的数量减一。不断重复这个过程，直到最后只剩下一个最大的类别。层次聚类算法的步骤可以概括如下：

1) 根据适当的距离定义准则，计算现有的 N 个类别两两之间的距离，找到其中最近的两个类(不妨记为 P 和 Q)；

2) 将 P，Q 合并，作为一个新类 PQ，加上剩下的 $N-2$ 个类，此时共有 $N-1$ 个类；

3) 重复步骤 1)，2)，直到聚类数缩减为 1 停止。

系统聚类的算法复杂度是 $O(n^2)$，上述聚类的结果可以用一个树状图展示，如图 3 所示，采用了身高体重数据集为样本绘制的层次聚类示例图，其中树的最底端表示所有的样品单独成类，最顶端表示所有的样品归为一类，而在此之间，聚类数从 $N-1$ 变动到 2。在任何一个给定的高度上，都可以判断哪些样品被分在树的同一枝，而聚类数的确定，需要通过实际的情况进行判断。

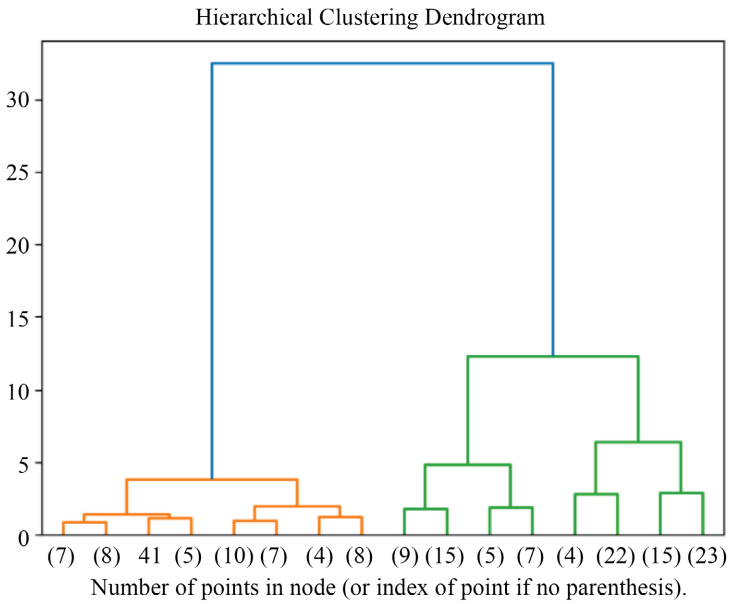

Figure 3. Hierarchical clustering treemap with a weight dataset example
图 3. 以身高体重数据集示例的层次聚类法树状图

3.2. 疫情新闻主题词分析

由于层次聚类绘制的树状图主题词太多，所以这里采用新闻正文分词中，指定出现频率较低的词，进行层次聚类分析，结果如下图 4 所示。最终生成图像如下所示：

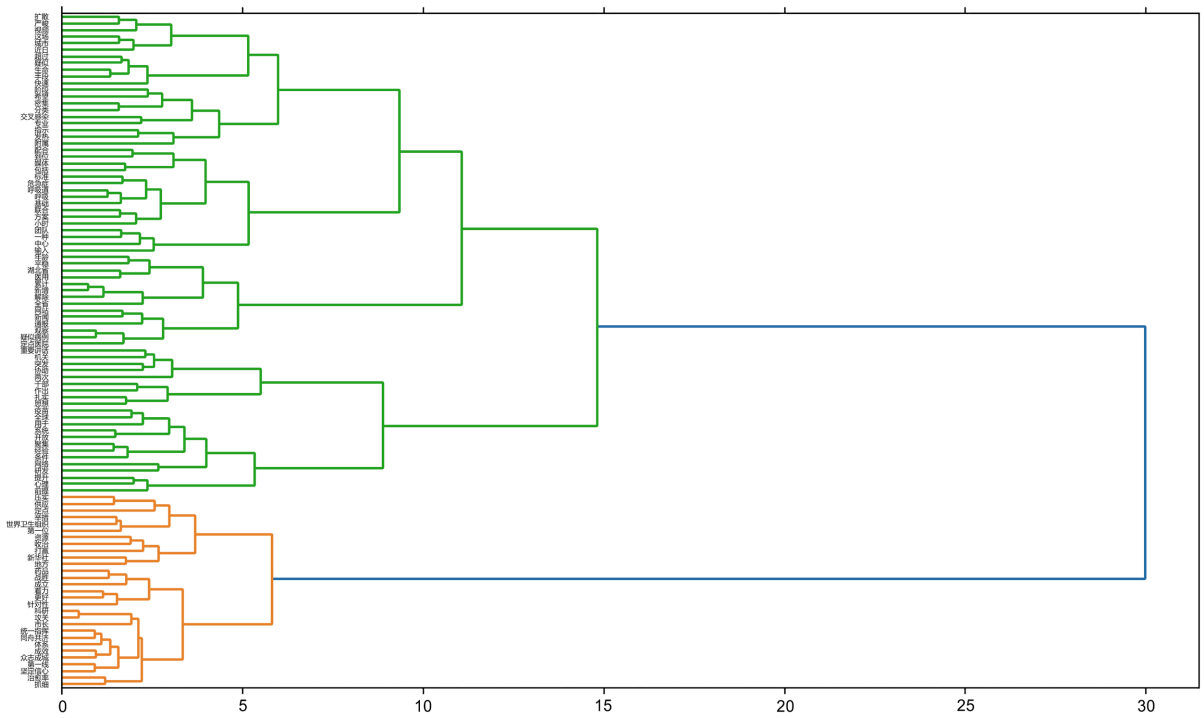

Figure 4. Hierarchical clustering dendrogram
图 4. 层次聚类树状图

根据图形检验初步分类结果的合理性并进行局部的矫正、相似合并等处理，得到的划分结果可以大致分为两类：前面 72 个对象分为一类，后面 29 个对象分为一类，第一类对象主要是疫情时期人们在网络上提及的疫情措施的实施情况，第二类对象是提及到的预防、防治等问题

4. 结论

首先，我们通过爬取疫情相关新闻 142 篇。对新闻正文文本信息进行数据预处理后，我们发现分词效果较为符合预期。经过特征提取、词云图进行可视化分析，得到“疫情”，“防控”，“工作”，“肺炎”，“患者”等词出现频率较高、使用 GSDMM 主题模型挖掘得到民众关注的企事业单位复工复产、自我安全防疫、疫情政策管理三大类，最后使用聚类分析将所有新闻文本分成疫情措施实施和防疫、防治两大类。

本文通过对疫情始发期新闻文本的具体分析，我们为政府“防控疫情，回应舆情”提供以下几点建议：一、坚持正确的防疫舆论导向。要加强防疫政策宣传，筑牢防疫底线思想，及时解疑答惑、回应网络舆情，畅通官方媒体传播渠道。避免造成公众恐慌，否则不仅不利于疫情防控工作的开展，还会动摇民心，进而影响社会的稳定。

二、提升主流媒体担当精神。新型主流媒体要增强发布的及时性、针对性和专业性。瞄准焦点问题，及时发布权威信息，正确传达中央政令，引导群众保持积极心态，传递社会正能量。展现主流媒体的流

量担当。

三、加强疫情防控卫生科普传播。有关单位要完善疫情防控卫生工作的常态化、长效化机制，通过多种措施入户、入村，做好多渠道、多层次、全方位、立体化防疫卫生宣传工作，提高群众的科学防疫意识。

参考文献

[1] 杨秀璋, 武帅, 夏换, 于小民. 基于主题挖掘和情感分析的“新冠肺炎疫情”舆情分析研究[J]. 计算机时代, 2020(8): 31-36. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2020.08.008

[2] 刘惠, 赵海清. 基于 TF-IDF 和 LDA 主题模型的电影短评文本情感分析——以《少年的你》为例[J]. 现代电影技术, 2020(3): 42-46.

[3] 杨雪寒, 焦玮, 张倩, 孟洁. 面向医院网络舆情分析的情感文本挖掘[J]. 微型电脑应用, 2020, 36(12): 31-34.

[4] 彭浩, 周杰, 周豪, 赵丹丹. 微博网络中基于主题发现的舆情分析[J]. 电讯技术, 2015, 55(6): 611-617.

[5] 岳宗朴, 刘彩, 李莹, 陆文静. 基于微博数据挖掘的“新冠疫情”评论文本分析[J]. 品位经典, 2020(12): 48-50.

[6] 黄勃, 陈欢, 方志军, 王明胜, 刘文竹. 基于微博的 COVID-19 热点话题分析[J]. 武汉大学学报(理学版), 2020, 66(5): 425-432. https://doi.org/10.14188/j.1671-8836.2020.0045

[7] 李丽. 基于改进 GSDMM 聚类模型的机器人文献主题划分研究[D]. 武汉: 华中科技大学, 2017.