数据科学大作业 计算社会学篇

重大突发公共卫生事件下的
网络社会心态及公众情绪引导研究

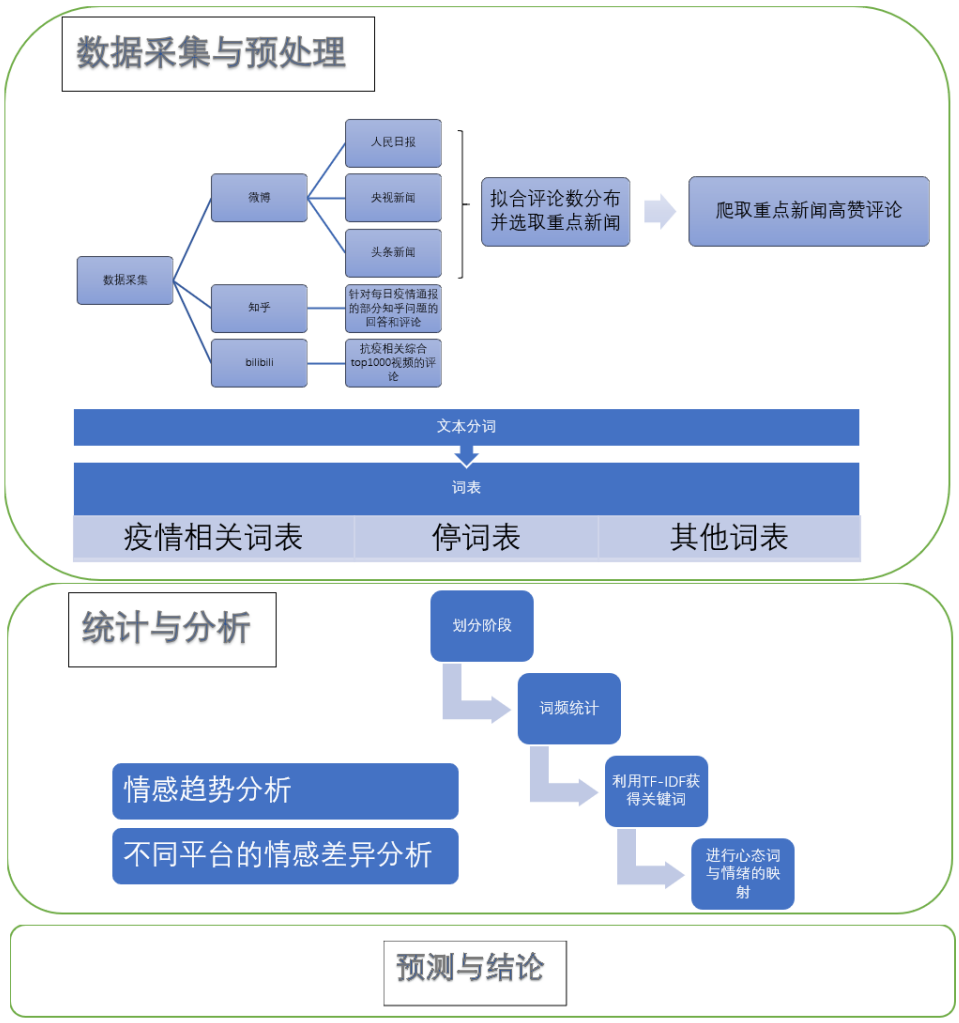
小组成员信息：

学号	姓名	邮箱	职责
191250215	周子杰	191250215@smail.nju.edu.cn	bilibili 爬虫，数据拟合筛选，词频提取，数据格式处理，结果数据处理与数据可视化
191250170	严佳欣	191250170@smail.nju.edu.cn	微博爬虫、数据可视化、微博关键词分析
191250082	梁言	191250082@smail.nju.edu.cn	知乎爬虫，TF-IDF，心态词映射/筛选

摘要：本文依托各类网络平台上疫情期间的相关信息，基于爬虫技术、分布拟合技术、TF-IDF 信息检索与数据挖掘技术等，获得了疫情期间（时间从 2019 年 12 月 8 日至 2020 年 6 月中旬）三大互联网平台不同阶段的各个情绪占比走势并依托各类图表对相关数据进行可视化展现。我们发现疫情期间网络社会心态宏观上呈现从消极到积极的转变，但也受到突发事件的影响而产生波动。该研究同时试图基于此做出重大公共卫生事件下网络社会心态的预测并对该类事件下的网络社会心态和公众情绪引导进行探究，结论包括如疫情中段的心理建设需重视，正面宣传效果显著，注意不同人群情绪变化差异等。

关键词：新冠肺炎疫情；情感分析；爬虫；社会心态与情绪

项目流程图：



目录

1. 绪论

1.1 研究问题

重大突发公共卫生事件下的网络社会心态及公众情绪引导的研究

1.2 研究原因

中国社会正处于深刻而快速的转型期，其中，在社会变迁层面，社会结构的快速分化，以“撕裂”的方式强化了社会团体、阶层之间的张力，使得整体社会结构出现紧张（李汉林、魏钦恭、张彦，2010），并投射在个体心理层面，进一步凸显出公众的社会认知、情绪、信念、意向、行动等对社会治理的重要影响（王俊秀，2014；杨宜音，2006）。同时，随着互联网应用的不断普及，日益多元复杂的公众情绪，借助网络的力量传播和放大，对社会心态的塑形力量进一步增强，赋予了群体心理及集体行为的极化可能（周晓虹， 2014）。当下新型冠状病毒（COVID-19） 肆虐全球，给人们的生产和生活产生了极大影响，也形成了疫情下独特的网络社会心态和公众情绪。

1.3 研究意义

立足此次新型冠状病毒（COVID-19）重大突发公共卫生事件情境，借助适宜的数据与计量手段，准确并客观地了解公众的网络社会心态与基于此呈现出的行为规律，期望能够实现对公众的情绪引导，让大众以积极的心态与政府一起应对和处理公共卫生事件及其衍生问题，从而达到维护国家与社会的长治久安的目的。

1.4 研究说明

本次研究由三人合作完成，研究过程中的相关代码及数据结论已存入：https://github.com/jiaxinr/ProjectOfDataScience

2. 数据采集

2.1 微博

新浪微博已经成为人们生活中非常重要的信息来源之一，它作为全球最大的中文社交网络平台，拥有超过 4 亿的活跃用户。本次研究之所以选择新浪微

博作为数据来源，主要考虑到其数据开放程度高、思想观点表达丰富、数据量大，故有较高统计学参考价值。

本次共采集了人民日报、央视新闻、头条新闻三位博主自 2019 年 12 月 8 日至 2020 年 6 月中旬的所有微博。其中，人民日报 6830 条、央视新闻 4674 条、头条新闻 4643 条。经过相关数据筛选（详见 3.2），又利用爬虫技术获取了一定数量的高赞评论，共计约一万二千余条。

在爬取数据的过程中，主要遇到了以下几个问题。首先，微博有手机客户端、网页端、移动端等版本，对于数据爬取时的选择来说，三种版本各有优劣。最终，因网页端的相关内容较为丰富，且较易达到翻页的目的，故选择网页端作为爬取对象。其次，微博反爬机制较为完善，故程序在爬取时需设置较长的休眠时间，从而使得微博数据的爬取变得较为缓慢。第三，除了爬取微博正文之外，还将评论数、点赞数、转发数、发表时间等信息一并抓取了下来，以便后续对数据进行处理及分析。

2.2 知乎

对知乎数据的爬取主要分为两部分。第一是爬取了 1.26-2.29 针对每日疫情通报的知乎问题或其他疫情相关热点问题（如：“1 月 27 日武汉疫情发布会称筹款统一归口，只通过省红十字会接受捐赠，会对物资援助带来什么影响”）的回答和评论，时间间隔为 2-3 天或更短，共爬取问题 16 个，其中回答数目 3829 条，评论数目 11802 条,；第二是爬取了与疫情相关，时间段不明但是比较能体现和反映网民心态的三个问题，分别是：“武汉乃至湖北各区市实际情况如何”，“新冠肺炎给中国带来的积极意义是什么？”，“钟南山院士称新型冠状病毒传染性比 SARS 最强时弱，两者对比如何，从非典我们学到了什么？”，其中回答数目 6843 条，评论 15934 条。

爬取知乎的难度在于，知乎不像人民日报，荔枝网等新闻网站那样是静态加载的，而是动态加载的。比如知乎一个页面内只能容纳一定数量的回答，要想获得全部回答需要不断下拉；需要点击每个回答下面的“xxx 条评论”才可以进入评论区，因此爬取知乎不能像静态加载的网站一样直接从 elements 中获取。

经过观察发现，网站源码中 network 一栏体现了动态加载的过程，随着评论区，回答的不断加载，network 中也会出现新加载的部分，而且可以通过 preview 栏来查看 answer 元素或者 comment 元素的内容，储存在 data 域里。同时我还注意到，answer 和 comment 元素有一个 paging 域，这个 paging 域是用来在动态加载中“翻页”的，也就是说，可以通过当前已加载的元素的 paging 来获得在他之前加载的（previous）和在他之后加载的（next）的起始 url。于是我决定利用 network 来获得每个问题下面第一个 answers 元素，然后利用 paging 的 next url 不断获取之后的 answers 元素，并根据 paging 中的 is_end 来终止循环。

对于每个回答的评论同理，只是评论的构成更为复杂，由一个 root_comments 元素和（如果有）的 child_comments 元素构成，其中 child_comments 对应着对某条 root_comments 的全部回复，尽管有些复杂，但是两者的规则和回答差不多，依然主要依靠 paging 的 next 不断获取之后元素的 url，并且根据 paging 中的 is_end 来终止。

2.3 bilibili

bilibili 近年来作为发展势头正猛的视频弹幕网站，受到越来越多年轻人，乃至各年龄段用户的喜爱。本次研究之所以选择 bilibili 作为数据来源，主要考虑到其与之前选择的两大平台，即知乎与新浪微博均具有较大性质及受众差异，因此能期待在 bilibili 获取到的数据中获得新信息。

bilibili 获取的数据是抗疫相关视频综合排序 top1000 共计 1000 个视频的弹幕，热门评论与播放量，并出于以下原因在综合考虑后仅选择评论作为分析对象而放弃分析所获取的弹幕。其一，弹幕存在刷屏与跟风的情况，也在社区文化的影响下导致有部分“玩梗”情况存在，弹幕内容有效信息较少。其二，不同的弹幕之间难以进行筛选，因为弹幕中点赞的情况较少出现并且点赞数隐藏，导致不存在“热门弹幕”等便于直接筛选的门类，或点赞数等便于拟合筛选的数据。综合考虑以上原因，最终投入使用的数据只有每个视频的热评部分，而放弃使用所爬取的弹幕。

爬取弹幕与评论的过程都需要首先获取每个网页上所有视频的 BV 号，再通

过 BV 号进一步反推 AV 号，视频 oid 等信息。其中 AV 号的反推借鉴了知乎用户 mcfx 提供的算法，其余信息则在特定由 BV 号生成的 url 中获取。获得相应信息后，即可组合获得显示相关视频弹幕，评论或播放量的 url，进一步通过正则表达式提取相关信息并存入文件。bilibili 也具有一定反爬机制，需要进行一定请求头伪装与定时休眠。

2.4 人民日报、百度新闻、天涯、荔枝网

除了上述三项信息来源，我们还对人民日报-人民网、百度新闻、天涯等网络新闻平台进行了数据抓取。但因其数据量重复率高、新闻受众面小从而导致极少的评论互动等原因，最终在统计时未用到这些数据。但在研究的分析及预测阶段，我们通过人工干预分析的方式，同样参考了这些数据。

3.数据处理

3.1 格式处理

由于本次作业中所涉及到的代码，例如 TFIDF，hanlp 分词与提取词频，评论数分布拟合等不同板块所需的数据形式与存储格式不同，在项目过程中，我们一直使用 python 读写文件进行数据格式转换操作与文件切分等，以便不同代码直接使用。所爬取的数据也由此分为多个 txt，单个 txt，xlsx 文件等多种形式存储。

3.2 拟合与筛选

对于爬取到的数据，我们们要在其中筛选出有效部分进行进一步数据分析。

首先对于 bilibili 平台，由于原本的爬取目标是抗疫相关视频的综合排序前 1000 视频的热门评论，数据本就经历了一次平台的筛选，原始就具有较强可靠度，初步认为不需要筛选。考虑到 bilibili 视频影响度评估里播放量一般被认为是最重要的凭证，我们爬取 top1000 视频的播放量进行分布拟合，发现大约如下图（其中横轴的单位为 10 万播放量，纵轴为视频数）：

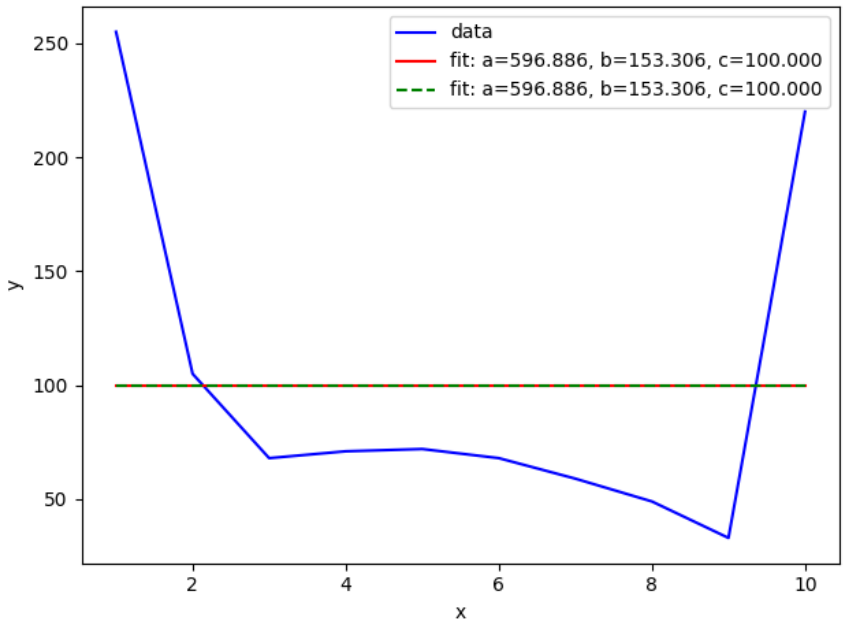


图 1 bilibili 播放量分布

由于数据量较少以及数据本就经过一轮筛选，播放量分布较难拟合成具体函数，但我们能够发现除了 10 万左右以及 100 万以上数量较多，所有视频的播放量都是在一个较高的水平上较均匀分布，直接投入使用应该偏差不大。

其次对于知乎平台，由于收集方式的关系（详情见 2.2），出于与不筛选 bilibili 数据类似的原因，加之对评论数等具体数字的观察，数据可信度也本身具有较大保障，因此决定对知乎数据也不进行第二轮筛选。

最后对于微博平台，由于爬取方式是爬取某账号某时间段的全部微博，原始数据未经过筛选，我们决定对所有微博评论数通过分布拟合和直接排序两种方式综合筛选，并选取两种筛选方式综合考量下约最有效的 10%微博爬取其热门评论并进行分析。

在进行数据拟合的时候，我们也选取了两种不同的横纵坐标生成方式。方式 1 是横坐标代表评论数（单位：千）而纵坐标代表具体微博数量。方案 2 是横坐标代表微博编号（从 1 开始的自然数）而纵坐标代表播放量。获取大概图形后发现并非服从正态分布而更类似于指数分布，故用 y=a * exp(-b * x) + c 或 $y=a*x^2+b*x+c$ 的公式尝试进行拟合，最终选取指数分布所得到的结果。获得的具体结果如下图所示。

其中在数据形态增长率过大时拟合偏差较大甚至出现参数溢出情况，其余

情况能获得较准确的拟合上下界，其具体参数值见图中。

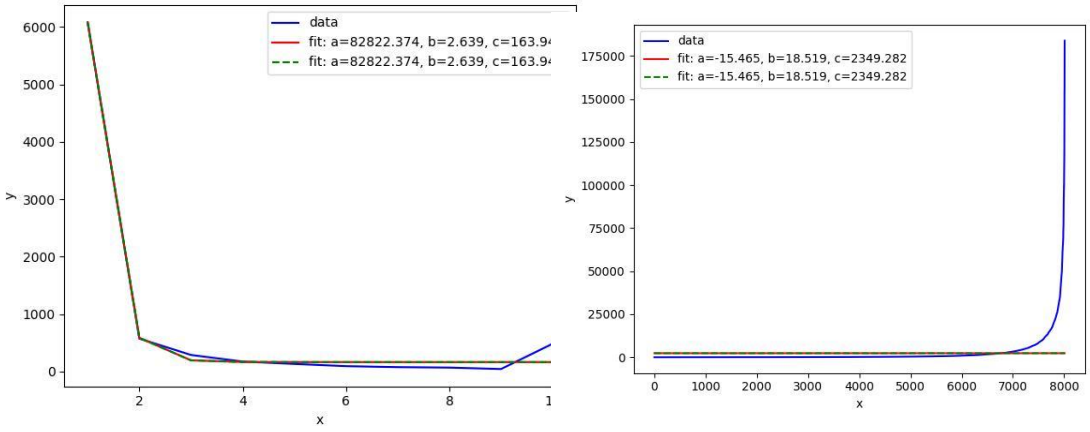


图 2.1，2.2 央视评论数拟合

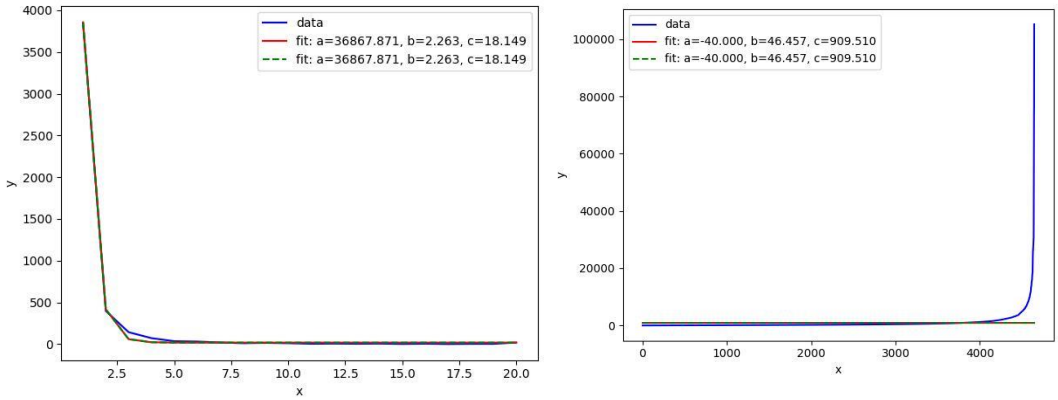


图 3.1，3.2 头条评论数拟合

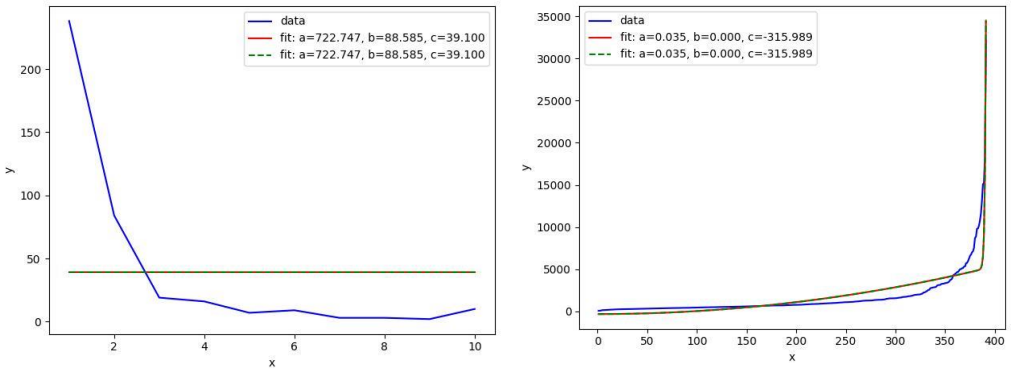


图 4.1，4.2 人民日报评论数拟合

从获取结果来看，微博评论量较多分散在 2000 以下的部分，而在 2000 以上的部分也有较大数量差异。通过排序发现，2000 左右评论正好接近微博评论数分布的 10%上分位点，故在综合考虑后，我们直接选取评论数排名最高的 10%的微博对应的热评为有效微博数据。

3.3 文本分词

采用 jieba 的全模式分词，这个分词方式主要是和 TF-IDF 结合在一起进行关键词提取和心态映射。

为了避免分词结果过于单一绝对化，我们同时采用了 pyhanlp 库进行文本分词，并用这个版本的分词程序获取了一个版本的词频，同时人工对停词表进行一定量补充。

4.数据统计

4.1 划分阶段

微博：微博数据研究时间段的划分按照作业指导上的建议分为以下时段：

2019.12.8-2020.1.22 不重视与无奈扩散阶段

2020.1.23-2020.2.7 武汉封城，资源缺乏阶段

2020.2.8-2020.2.13 严格管控与物资配给阶段

2020.2.14-2020.3.10 按部就班防控疫情阶段

2020.3.10-2020.6.18 有序复工阶段

知乎：鉴于知乎数据的特殊性，对于疫情比较严重的一个月（1.26-2.29）的数据，采取了更严格的划分。根据中国疾病预防控制中心的数据来看，这一个月左右的时间又可以划分为四个阶段：第一阶段 1.26-1.31，这一阶段是疫情刚刚爆发；第二阶段 2.1-2.10，这一阶段疫情疑似和确诊都快速增长；第三阶段 2.10-2.15，这一阶段疫情有被控制住的趋势，治愈率显著上升；检测更为准确科学，因此疑似显著下降，确诊显著上升，确诊增速明显增快；第四阶段 2.16-2.31，这一阶段疫情开始得到控制，疑似进一步下降至可忽略不计，确诊成功率大幅提高，确诊出现拐点，先是增速逐渐下降到 0，然后在 2.20 日前后开始出现减少，治愈出院率大幅度提高。

对于第二类数据，鉴于问题本身较为宏观，而且时间跨度比较大，因此不归纳到具体的某一个时间段内，而是单独作为“时间段不详”的一类。

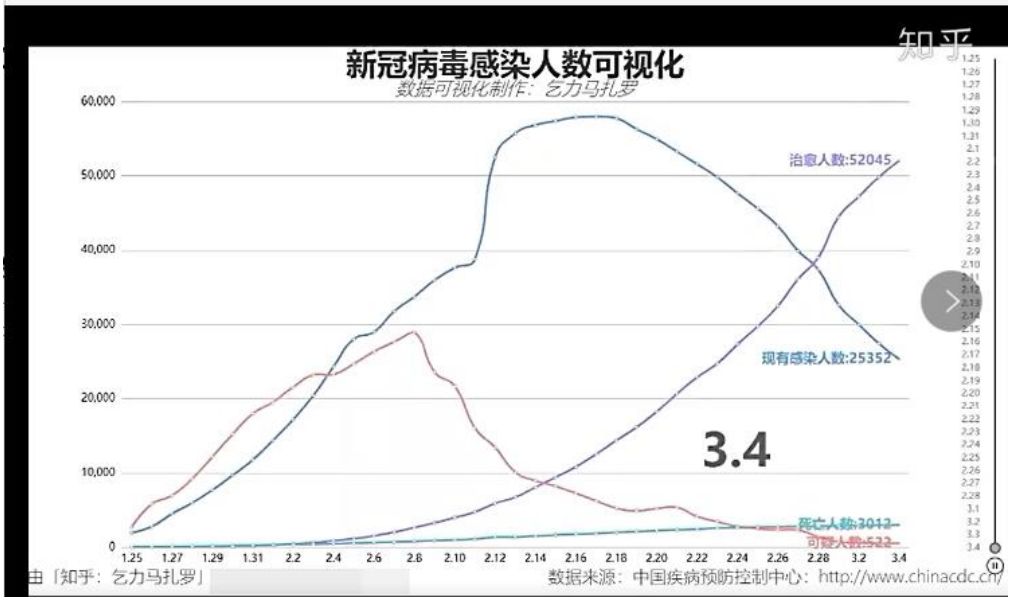


图 5 网友根据中国疾病预防控制中心数据绘制的疫情情势变化图

Bilibli：Bilibili 数据出于一些原因（详见 5.1）未进行阶段划分而是作为整体分析

4.2 词频统计

我们选择用高频词进行相关词云生成。由于词云的形式是为突出重点而非把握趋势，该版本的词频统计采用的是未划分阶段，而仅区分平台的数据集。统计高频词词频的时候，选取三个平台汇总的数据，分别统计其间高频词出现数量，并通过词语-数量的对应生成三幅词云。

高频词由于仅关注词语出现数目而会造成一定偏差，因此虽然可以用于生成词云把握大体趋势，但不能直接对应为文章关键词。因此我组采取了下述 TF-IDF 的方法进行关键词提取。

利用词云相关技术，我们通过将高频词可视化的方式作为辅助，从而更好地把握疫情期间公众的关注点走向趋势。

我们总共制作了四张词云图，分别为：微博正文、微博评论、知乎、bilibili。

词云图中，我们关注到：

微博正文高频词为病例、确诊、疫情；

图 6 微博正文词云

微博评论高频词为：武汉、中国、加油；

图 7 微博评论词云

知乎高频词为：病例、确诊、新增；

图 8 知乎词云

bilibili 高频词为：中国、美国以及其他相关表情。

图 9 bilibili 词云

之所以选择使用高频词而非关键词进行词云生成，是因为我们想先通过词云大致把握并直观体现疫情期间较常被提起的词语，而实现最粗略的把握与比对。

4.3 利用 TF-IDF 获得关键词

TF 是词频（Term Frequency），IDF 是逆文档频率（Inverse Document Frequency）,TFIDF 是 TF*IDF，用来刻画一类词语是否具有很好的区别能力，适合用来分类。

TF 指的是某一个给定的词语在该文件中出现的频率，这个数字是对词数（Term Count）的归一化，以防止它偏向较长的文件。由于我们的数据分析是针对整个语料库的，因此某一个词的 TF 是这个词在整个语料库（所有文件中）出现的总频数/整个语料库中所有词的总频数（这里面的所有词不包括在停词表中）。IDF 则通过总文件数目除以包含该文件的数目来计算。

进行关键词提取的第一步是进行分词，本组在 TF-IDF 中采用的分词是 jieba 全模式。关键词提取的第二步是使用停词表，本组采用的初始停词表综合了中文停词表(cn_stopwords.txt)和哈工大停词表(hit_stopwords.txt)，以达到对疫情相关词汇更准确的筛选；在关键词提取过程中，我们也利用不同语料库（不同平台不同时间段的数据）选取比较小的关键词个数（在代码中是 topK，含义为 TFIDF 值最大的 K 个元素）例如 topK=30 或 40 进行试验，不断将所得结果中情感不明确或者与疫情无关的词汇加入停词表。不断重复试验以保证最终获得的 TF-IDF 数值最大的关键词与疫情相关。

在具体实现方面，利用 os.walk 和_iter_方法实现对指定文件夹下文件的遍历，并且每次返回一个文件。然后将拿到的文件进行分词，返回一个列表，对这个列表进行遍历，如果当前元素不在停词表中就计算它的 TF 和 IDF。Term 的 TF 和 IDF 分别被保存在两个词典中，一个对应 value 为 x 出现的总频数，一个对应 value 为 x 出现的文件数。TF 的计算比较容易，只要 x 在文件中出现了，就把对应 TF 词典 key = x 的 value +1 即可；对于 IDF 则是首先定义一个记录 key 值是否在文件中出现的字典（记为 isInFile），在每进入一个新文件的时候将所有 key 值的 value 记为 False，表明所有 Term 在这个文件中目前还没有出现过，遍历分词后文件中每一个 Term（记为 x），如果 isInFile[x] = false，表明 x 还没出现过，这是第一次出现，那么将对应值改为 True，IDF 词典对应 value + 1。依照此思路以不同平台不同时间段的数据作为语料库，根据给定 topK 打印前 K 个关键词，即完成了对应语料库关键词的提取。

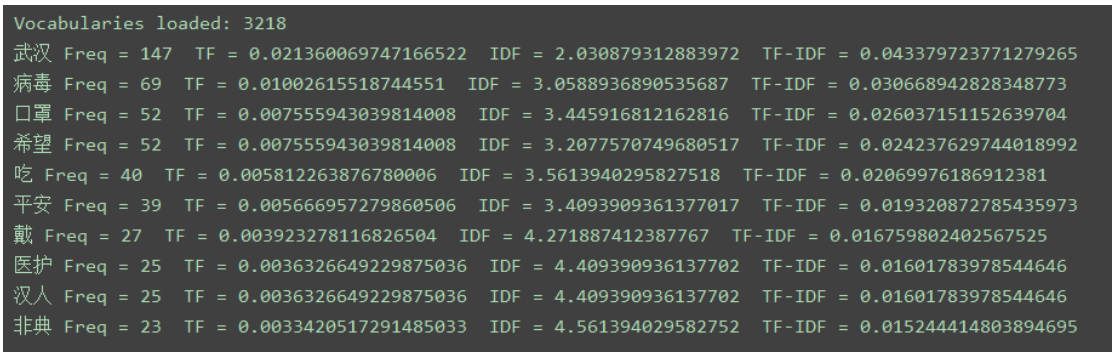


图 10 TF-IDF 实现关键词提取实例（此时 topK=10，语料库为微博 1.22 日以前的评论）

4.4 情绪映射

情绪映射准确的前提是最终选取的关键词集合应该是不同语料库重要关键词的合集。因此我们选取较大的 topK（topK>=50），每次选取不同的 topK 值，并对该值下不同语料库提取出来的关键词进行比对。我们发现：在 topK 比较小比如 50，60 的时候，每个语料库的关键词存在遗漏现象，不能基本涵盖疫情相关的重要（指 TF-IDF 值比较大）词汇，在大于 100 较多时，比如 120 或者 150 的时候，又会有一些疫情无关或者是不那么重要的词汇进行关键词的范围，因此我们最终选择 100 作为最终的 topK 值，以达到在无关词汇较少的情况下尽可能准确的映射。

我们首先人工对每一个词贴情绪标签，将关键词和其情绪一一对应起来。然后先对几个语料库的情绪标签和其包含的关键词进行汇总，修订一个比较完善的“情绪字典”，依照这个情绪字典对剩下的语料库进行映射。在汇总的过程中我们发现：

第一，有些词汇本身可能对应不同的情绪，比如“啊啊啊”既可以表示高兴，也可以表示害怕；有些词汇因为语境不同也可能代表不同的情绪，比如“医生”“医务人员”既可以在“医务人员们辛苦了！”这个句子中表示感动/感谢的情感，也可以在“医务人员们要注意安全啊！”这个句子中表示担心或者关切；

第二，人工贴标签的情感比较精确，因此最终情绪标签的种类比较多，达到了将近二十种，而标签太多太精细不便于进行情绪的趋势分析和预测；

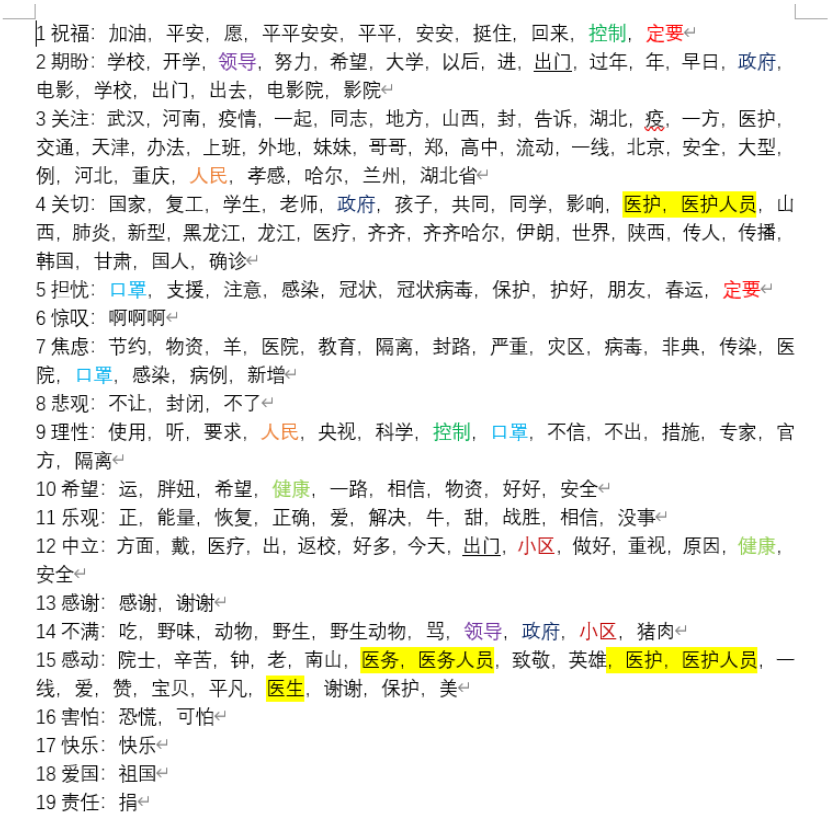

1 祝福：加油，平安，愿，平平安安，平平，安安，挺住，回来，控制，定要
2 期盼：学校，开学，领导，努力，希望，大学，以后，进，出门，过年，年，早日，政府，电影，学校，出门，出去，电影院，影院
3 关注：武汉，河南，疫情，一起，同志，地方，山西，封，告诉，湖北，疫，一方，医护，交通，天津，办法，上班，外地，妹妹，哥哥，郑，高中，流动，一线，北京，安全，大型，例，河北，重庆，人民，孝感，哈尔，兰州，湖北省
4 关切：国家，复工，学生，老师，政府，孩子，共同，同学，影响，医护，医护人员，山西，肺炎，新型，黑龙江，龙江，医疗，齐齐，齐齐哈尔，伊朗，世界，陕西，传人，传播，韩国，甘肃，国人，确诊
5 担忧：口罩，支援，注意，感染，冠状，冠状病毒，保护，护好，朋友，春运，定要
6 惊叹：啊啊啊
7 焦虑：节约，物资，羊，医院，教育，隔离，封路，严重，灾区，病毒，非典，传染，医院，口罩，感染，病例，新增
8 悲观：不让，封闭，不了
9 理性：使用，听，要求，人民，央视，科学，控制，口罩，不信，不出，措施，专家，官方，隔离
10 希望：运，胖妞，希望，健康，一路，相信，物资，好好，安全
11 乐观：正，能量，恢复，正确，爱，解决，牛，甜，战胜，相信，没事
12 中立：方面，戴，医疗，出，返校，好多，今天，出门，小区，做好，重视，原因，健康，安全
13 感谢：感谢，谢谢
14 不满：吃，野味，动物，野生，野生动物，骂，领导，政府，小区，猪肉
15 感动：院士，辛苦，钟，老，南山，医务，医务人员，致敬，英雄，医护，医护人员，一线，爱，赞，宝贝，平凡，医生，谢谢，保护，美
16 害怕：恐慌，可怕
17 快乐：快乐
18 爱国：祖国
19 责任：捐

图 11 第一版的“情绪字典”，可以看到这个里面情绪多达 19 种，而这仅仅包括了两个语料库的词汇。里面用不同颜色标注的词是映射有歧义的词汇。

对于这两个问题，我们提出了以下两个方案：

对于第一个问题，我们认为出现歧义的词语可以归到两个类当中。由于这个词语对两个歧义心态 TF 和 IDF 的贡献是相同的，所以这个词语在这两类情绪谁更占主导上并没有显著作用，两类情绪谁胜出还是更取决于歧义词之外词汇的贡献。这种方法既可以保证情绪映射的严谨性也可以保证情绪映射结果的可靠性。

对于第二个问题，我们采取了两层标签合并。所谓两层就是第一层标签合并比较细致，合并为 8 个标签；第二层标签合并比较粗略，主要是根据积极，消极和中立来把关键词映射到三类。这主要是考虑到细版本（第一层）的划分情绪数目比较多，可能出现各阶段情绪占比差异不大，情绪趋势分析和预测结果可靠性不高的情况；粗版本（第二层）的心态虽然便于预测，但是每一类内在含有的情绪又太丰富，严格说来划分为一类有些牵强，因此又不能单纯用粗版本进行预测。基于这两种考虑，我们的情绪趋势分析和预测采取机器+人工的

方式：也就是机器进行拟合，对于每一类情绪内部再借助细版本的情绪标签进行进一步划分，达到情绪预测和情绪分类都比较准确的效果。

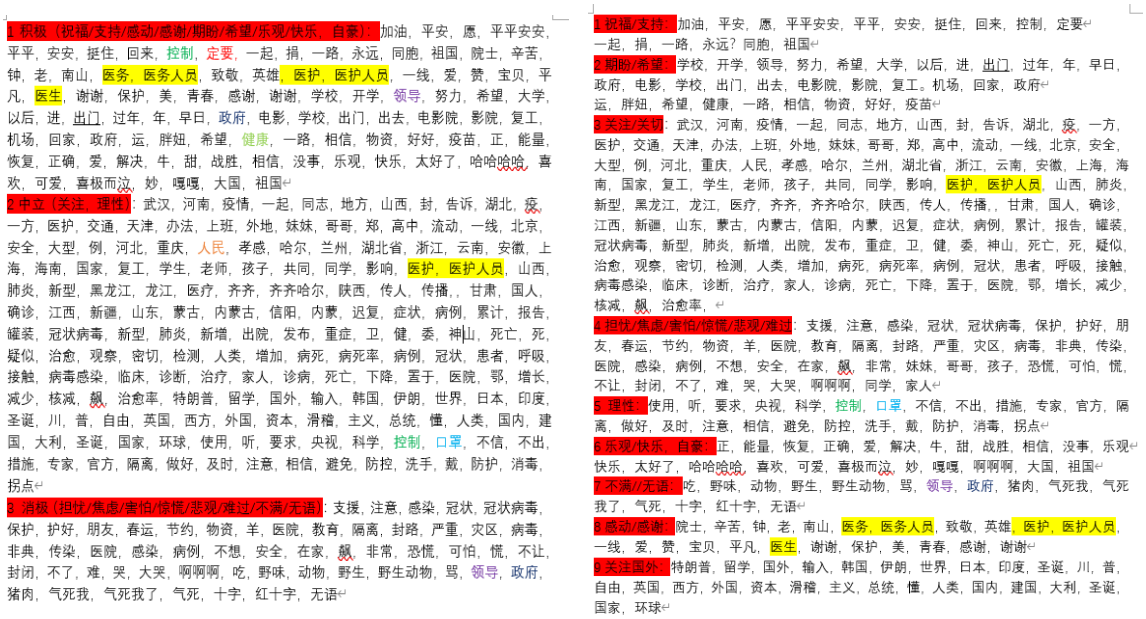


图 12 粗版本（左）和细版本（右）的“情绪字典”

类比关键词提取，我们觉得 TF-IDF 值也可以用来体现某种情绪在某一时间段内的比重，因此决定使用 TF-IDF 进行情绪分析。对于细版本和粗版本分别建立对应的 TF 和 IDF。此时的 TF 和 IDF 键值 key 分别对应不同的情绪，例如“关心关切”。我们分别创建了粗版本和细版本共 3+8 种情绪的“字典”，保存在 txt 文件中，其内容为可以映射到该情绪的关键词，词与词之间用逗号隔开，然后对于每一个不在停词表中的，也就是疫情相关的重要词汇，判断其是否在对应的“字典”里，然后类比关键词提取，对 TF 和 IDF 词典中该词对应情绪的 value 做处理。判断的方式是把文本读入，然后以“,”做切分（split）形成一个 list，判断词 x 是否在 list 中即可。

```
Vocabularies loaded: 9
关注国外 Freq = 64084  TF = 0.31333088864442  IDF = 0.07896946581822814  TF-IDF = 0.02474357290060057
理性 Freq = 12443  TF = 0.06083852829727417  IDF = 0.16413684908090528  TF-IDF = 0.009985844337434076
祝福支持 Freq = 6918  TF = 0.03382471580491382  IDF = 0.2755581317493283  TF-IDF = 0.00932067549415403
不满无语 Freq = 9143  TF = 0.044703581469258034  IDF = 0.16089118957725104  TF-IDF = 0.007192412400952481
感动感谢 Freq = 10213  TF = 0.0499352157437966  IDF = 0.12882998034495244  TF-IDF = 0.006433152862794275
期盼希望 Freq = 13917  TF = 0.06804547121378804  IDF = 0.09436657077673093  TF-IDF = 0.006421217775331936
关注关切 Freq = 44541  TF = 0.21777777777777776  IDF = 0.02929430679403805  TF-IDF = 0.006379649035146063
担忧焦虑 Freq = 25942  TF = 0.12684023957951351  IDF = 0.03374021182523033  TF-IDF = 0.00427961655137575
乐观快乐 Freq = 17324  TF = 0.08470358146925804  IDF = 0.03968947815641674  TF-IDF = 0.0033618409464943825
```

图 13 细版本情绪的 TF-IDF 结果（语料库为 bilibili 评论）

```
Vocabularies loaded: 3
中立 Freq = 83769  TF = 0.5096522982386761  IDF = 0.013108879053099785  TF-IDF = 0.006680970336745145
消极 Freq = 33230  TF = 0.20217199525446414  IDF = 0.024862060478777415  TF-IDF = 0.005026412373131588
积极 Freq = 47366  TF = 0.28817570650685975  IDF = 0.007268000464039107  TF-IDF = 0.002094461168616654
```

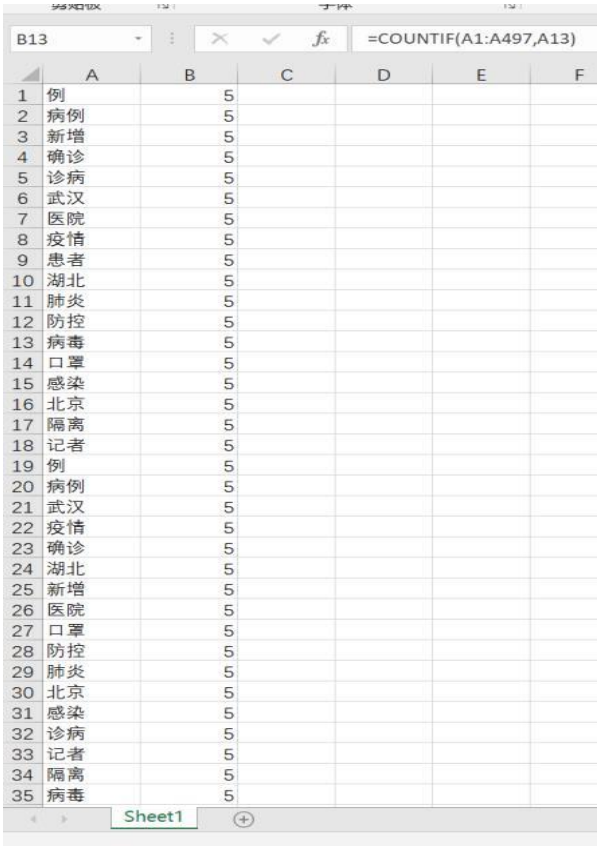
图 14 粗版本情绪的 TF-IDF 结果（语料库为 bilibili 评论）

5. 数据分析

5.1 关键词变化分析

当我们对所获取到的数据进行一定处理后，我们发现不同时期，人们所关注的重点不尽相同。通过观察微博正文中关键词的词频变化，我们可以更加清晰地感知到这一点。

首先，将所得到的分时段的高频词及其 freq、tf、idf、tf-idf 五项数据整理出来，利用 Excel 中的=COUNTIF()函数来选择出同时出现在这五个时间段的关键词（如图）



	A	B	C	D	E	F
1	例	5				
2	病例	5				
3	新增	5				
4	确诊	5				
5	诊病	5				
6	武汉	5				
7	医院	5				
8	疫情	5				
9	患者	5				
10	湖北	5				
11	肺炎	5				
12	防控	5				
13	病毒	5				
14	口罩	5				
15	感染	5				
16	北京	5				
17	隔离	5				
18	记者	5				
19	例	5				
20	病例	5				
21	武汉	5				
22	疫情	5				
23	确诊	5				
24	湖北	5				
25	新增	5				
26	医院	5				
27	口罩	5				
28	防控	5				
29	肺炎	5				
30	北京	5				
31	感染	5				
32	诊病	5				
33	记者	5				
34	隔离	5				
35	病毒	5				

图 15 筛选得到的关键词

紧接着，我们将其中重复的（如：病例与确诊）、无效的（如：例）词汇去除，从而得到了最终的共 11 个可用以分析公众关注点变化的高频词。它们分别

是：病例、新增、武汉、医院、湖北、防控、病毒、口罩、感染、北京与隔离。

当得到我们需要分析的词汇后，又整理出这些词汇在各个时间段中的 tf-idf 值，并将其以二维表格的形式整理出来。最后，将这些数据制成折线图，从而能够较好地表现出微博正文中所体现的，公众在疫情期间对相关话题的关注点变化趋势。

	12.23-1.22	1.23-2.7	2.8-2.13	2.14-3.10	3.10-6.18
病例	0.0164769	0.02456728	0.016694	0.024567	0.02453965
新增	0.00764958	0.02089297	0.010743	0.017833	0.01191371
武汉	0.01572775	0.01602411	0.013514	0.016453	0.00813673
医院	0.00671144	0.01403355	0.010337	0.012711	0.00982183
湖北	0.00686876	0.01226212	0.01167	0.014244	0.00638328
防控	0.00666272	0.01141426	0.009155	0.010497	0.01057746
病毒	0.01445784	0.01124876	0.005654	0.010724	0.01135426
口罩	0.00766031	0.01107746	0.009441	0.008875	0.0075296
感染	0.00949571	0.01099069	0.010991	0.008426	0.00903099
北京	0.00938034	0.01017415	0.008545	0.009443	0.01856269
隔离	0.00603387	0.00774994	0.007297	0.007492	0.00811552

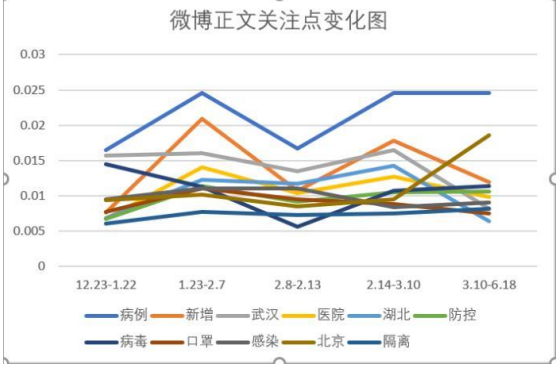


图 16 微博正文关注点变化折线图

从图中我们可以看出，“病例”这一词汇始终处在较高水平，说明人们对于患病者本身的关注度始终是最高的——无论是因为关注病例数是否有增长，还是关注病例数量所反映的疫情趋势，人们都自始至终不忘关注疫情本身。

此外，可以看出人们对“新增”这一词汇的关注度波动较大。在 2020 年 1 月下旬至 2020 年 2 月上旬，人们对新增关注度最高。通过数据分析，我们发现那段时间正是疫情新增数较高的时候，人们也逐渐意识到了疫情的严重性。那段时间正是网络上统计全国各地疫情新增数目网站出现率最高的时段，与我们的数据基本相符。

有关地区的词汇在本次高频词表中共有三个，它们分别是“武汉”、“湖北”与“北京”。尽管都是地名相关词汇，但出现频率趋势却截然不同。“武汉”与“湖北”在疫情刚发生时，作为起源地关注度颇高；而当疫情蔓延到全

国之后，北京作为首都，当有本土新增病例时，也受到了广泛关注与讨论。

5.2 情感差异分析

5.2.1 各平台情感分析

建立字典映射，获得关键词对应情绪后的 TFIDF 值后，我们利用 python 处理各阶段各情绪的 TF-IDF 值，计算其相对大小，从而获得每一类情绪在对应阶段所占的比重大小，并绘制折线图以关注其占比走势。其中折线图以每种情绪在所有情绪里所占的比例为纵坐标。通过读文件获取相应数据并计算后，利用 python 的 matplotlib 库，分别绘制了在两种划分方式与两个不同平台下各个情绪的走势。由于对情绪影响因素较多而不确定，阶段较少等原因，所获得的数据仅用于绘制折线图观察走势，而不做具体的函数拟合以防止过拟合。

此外，bilibili 数据由于形态较为特殊，因为从发生事件，收集素材，到剪辑上传视频需要一定时间而具有延迟性等特点，我们未进行阶段划分而是直接得到其总体占比，直接使用 excel 绘制饼状图。

5.2.1.1 bilibili

图 17 Bilibili 评论情感占比（粗版本）

在粗略划分三类的心态字典中，我们可以看到 bilibili 网友的态度里中立最多而消极最少。该平台的评论对“号召性”的重视程度较低于所选择的另外两个平台，因此网友也会在评论区更多展现自己内心较为消极与中立的一些想法。

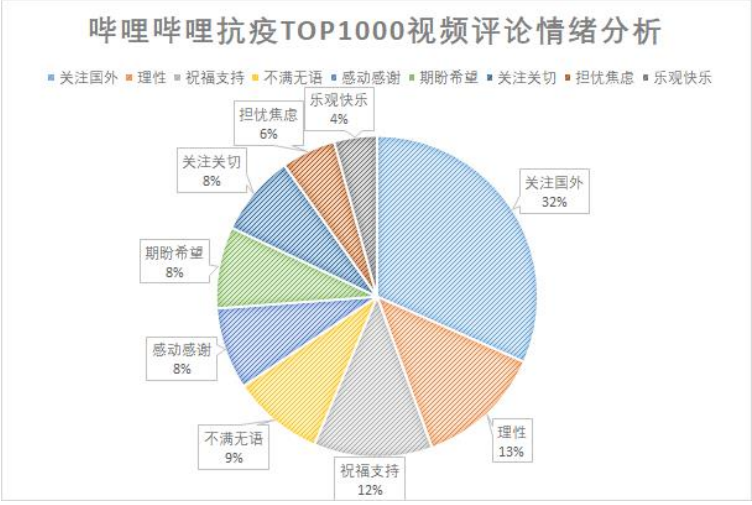
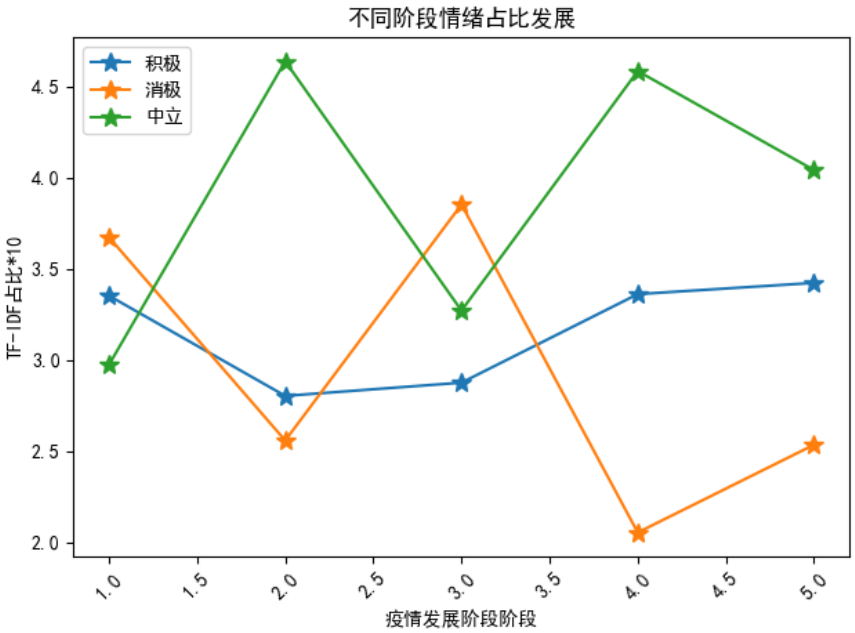


图 18 Bilibili 评论情感占比（细版本）

在更细化的九种情绪分布里，可以看到 bilibili 网友在评论里是很关心外国相关局势的，这也和年轻受众的猎奇心理与国际视野有关，同时也和视频发布的非即时性有一定关系，即许多视频剪辑完成发布时，会由于时间差，导致网友评论的重心不会完全放在事件本身上而会去和当下做一定对比。同时在 b 站发布的抗疫相关视频本身就有很大的体量涉及到国外，这也和 b 站的受众有一定关系。

5.2.1.2 微博



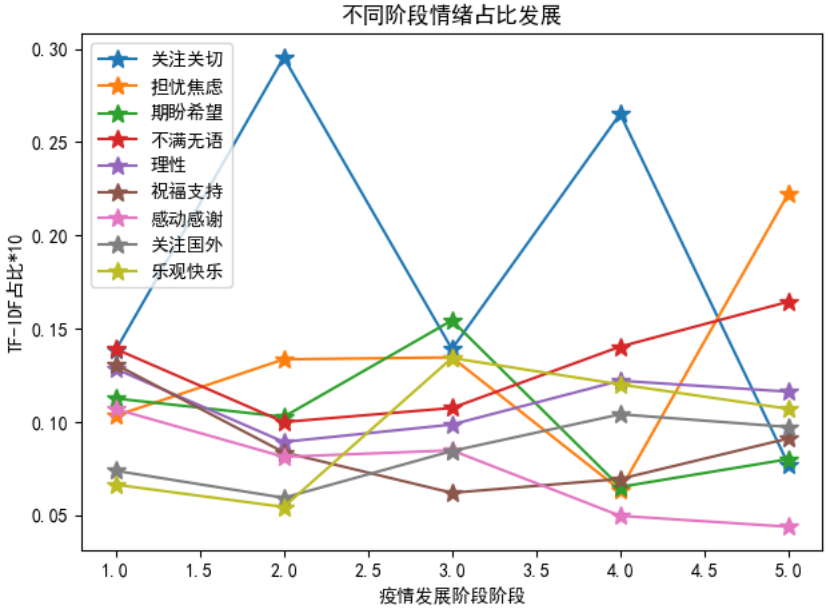
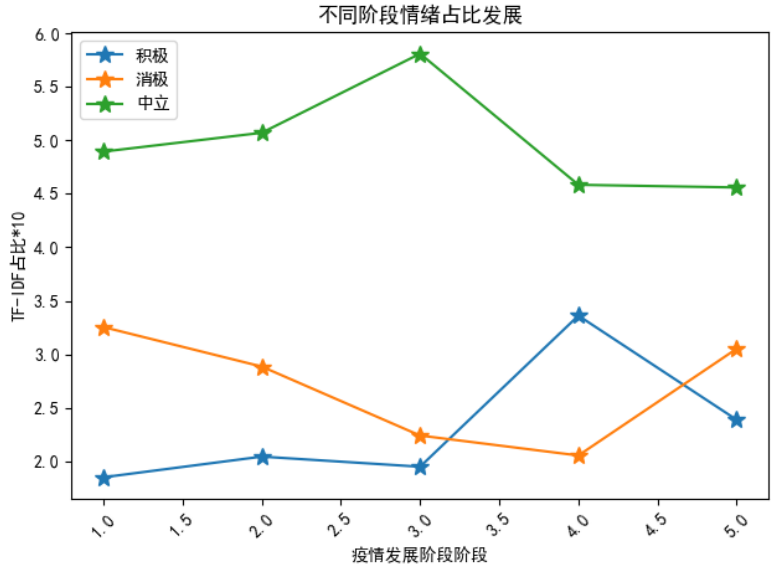


图 19 微博平台两种不同情绪划分方式的情绪占比各阶段走势图（阶段划分见 4.1）

从微博数据中可见，在疫情发展中段，形势未完全向好而网友已经逐渐陷入自我隔离的疲乏期时，情绪迎来一个偏向消极的小高峰点，此时的心理状态需要格外关注。此外如对国外的情况关注度渐渐提升，由于复工复产开学等系列举措造成不满略有上升，由于形势向好导致感动之情略微下降等走势都符合我们原本的预期。

5.2.1.3 知乎



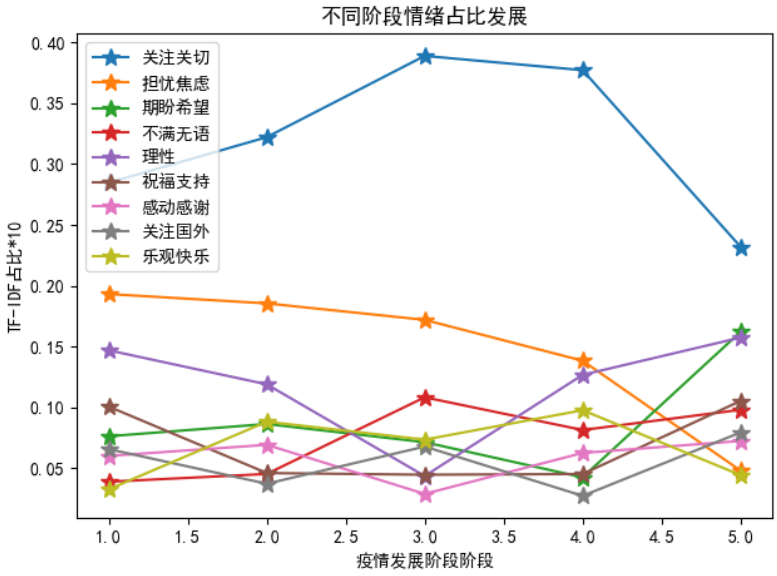


图 20 知乎平台两种不同情绪划分方式的情绪占比各阶段走势图（阶段划分见 4.1）

在绘制走势图时，我们遇到的问题是，从那些不便于划分阶段，较为“宏观”的问题里得出的数据，应该放入走势图还是应该单独列出。最终我们认为，从宏观的历史的角度看待疫情，所得出的数据会是比较接近最终平稳状态的，因此可以作为第五阶段画入走势图中。同时不将难归纳部分画入走势图的版本可以直接通过忽略图中第五阶段的方式获得，因此直接绘制拥有第五阶段的走势图是较好的选择。

知乎平台的言论相比微博平台，拥有更多思考与分析的成分。中立情绪在知乎评论中占据主体，排除第五阶段（即难以划分的阶段）来看，随着我国抗疫形式变好，知乎用户的消极情绪也有明显下降，伴随积极情绪的明显上升。

5.2.2 各平台对比分析：

尽管从诸多平台上获取的数据中，我们可以得到疫情期间不同时间段，当公众遇到重大突发公共卫生事件时的网络社会心态及其情绪变化，但不难发现，我们所抓取数据的不同平台，其主要受众面也有很大不同。

对于微博正文来说，因为我们抓取的来源主要是三个大型官方媒体，其措辞与关注点都较为官方，且鲜有主观心态的直接体现；对于微博评论来说，其受众面广泛，之中充斥着对疫情方方面面的各种心态；对于知乎而言，其关注点更是

扩散到了方方面面，从每个人的日常生活开始，再到当时公共机构、媒体等，其中不乏对这些方面的质疑、批判、赞赏角度；对于 bilibili 而言，其中表情的出现更为频繁，公众情绪也更为明显、激烈。

据此，我们认为这些不同平台所体现出来的情感差异，主要是由各个平台不同的受众面而引起的。举例来说，微博正文中的数据来源主要是由官方媒体人运营，如果带有太多煽动性情绪词，会对社会发展产生不利影响。而微博评论中的评论者受众面则更为广泛，其关注点更为普遍。知乎发展之初，是针对高学历、知识面广泛、逻辑性的人们，尽管现在也逐渐发展成为了“全民平台”，但其中回答与评论还是能看出其理性成分，也对日常生活的方方面面关注度更高。Bilibili 发展之初是针对那些对二次元文化较为感兴趣的人们，发展至今，其活跃用户也以年轻人为主。年轻人的情感更为强烈，对于网络用语的运用也更为频繁，我们的数据中已经明显地体现出了这一点。

6. 预测与引导

6.1 预测

通过对微博和知乎的分析我们发现，在本次新冠肺炎疫情中这两个平台的网络心态从宏观上可以归纳为以下规律：疫情最开始消极情绪占主导，随着疫情发展消极情绪逐渐减弱，积极情绪逐渐增强，随着疫情不断得到控制积极情绪逐渐占据主导，呈现出从消极到积极的转变趋势。而这个趋势也比较符合客观规律和普遍认知。

同时我们注意到，在疫情变化的过程中情绪也存在波动。例如在疫情发展中段，疫情情势尚不明朗，公共卫生部门尚未找到足够有效的解决方案和控制举措，突发疫情带来的恐慌的延续和居家隔离，停工停产居家隔离导致的焦虑和情绪上的疲乏导致消极情绪达到了疫情爆发初期之外的又一个小高潮。再比如疫情爆发初期以湖北省红十字会为首的几家红十字会和慈善机构归口口罩导致各地捐赠物资没办法及时送到需要的地方，将当时疫情初期各处蔓延的慌乱，恐慌，焦虑全部点燃，消极情绪汇聚成对红会的不满和愤怒，消极情绪迅速增长。由此我们看出，情绪的变化并非是一成不变的，遵循着从消极到积极的发展规律；而是会受到重大公共卫生事件爆发，处理，控制过程中各类突发

事件的影响而产生波动，因此情绪预测很难公式化和规律化，不同事件下的情绪总会因事件发展不同而呈现出不同的变化趋势。

同时由于平台不同，平台活跃的人群也不同，所以我们在进行预测的时候也不能忽略平台本身的影响。因此我们可以做出如下预测：

1.在宏观上呈现出从消极情绪向积极情绪的转变，但这个转变需要依赖于疫情得到控制的“信号”（权威专家对疫情的走向做出预测，诸如‘有望在 2 月底之前控制以前’）或者疫情得到控制的事实（确诊人数下降，治愈人数上升等）

2.对于不同平台：

知乎由于其受众相较微博和 b 站学历更高，更理性，见识更广，因此其情绪相对而言会更理性一些，但是当出现问题时，带来的批评，质疑和不满等负面情绪也会更强烈，扩散更广；正面情绪同理。

B 站由于受众年轻化和低龄化，因此无论是积极还是消极情感都会比较强烈，同时也容易缺乏理性，容易被“带节奏”，这也与 b 站本身网络舆论氛围紧密相关。

微博受众面更大，覆盖年龄层面更广，因此情绪会比较复杂。同时微博由于各大官方媒体公众号的存在，微博用户比较容易得到较为官方的信息；但是由于也有大 v 的存在，网络心态也容易受到大 v 言论的影响。微博的网络心态很大程度上取决于这些具有影响力的微博号内容所带来的舆论氛围。

6.2 引导

不管是对于哪个平台，在重大公共卫生事件面前，网络心态的混乱和消极主要是来源于由人们缺乏信息，或者说是获得的信息说服性，科学性低而带来的对事件本身了解和认识的缺乏。想要让网络心态平稳，同时不那么焦虑，慌乱，害怕，就要让大家认识到事件的本来面目，就需要有一个可信度，科学性高的消息来源给大家打一针“强心针”，而官方媒体和影响力大的媒体，自媒体在这之中毫无疑问起到了至关重要的作用。因此对于官方媒体来说，对于时间情况不能隐瞒不通报，而是要透明疫情发展的现况；同时要散布正向的舆论，比如请高级别的权威专家对疫情进行解读和预测，向大家透露出：“疫情可能来

势汹汹，但是是可控的”的信号；由于官方媒体性质的内在要求，对于官媒来说“讲透明，说真话”是底线，但是不具有官方性质的自媒体，大 v 同样也在网络心态发展中有至关重要的作用。这要求有影响力的自媒体和大 v 对自己的言论负责，要基于客观和理性的原则，必要时也可以传递一些积极的信号；但绝对不能带节奏，传谣言。事实上，网络心态受到每个人的影响，因为没人能知道自己说的一句无关紧要的话是否会在互联网的某个角落产生影响。这些对于大 v 和高关注度自媒体的要求本质上也是对我们每一个网民的要求。

然而不置可否的是，官方媒体在这个过程中发挥着最为重要的作用。但人们也不会一直官媒说什么，人民信什么。官媒是政府的喉舌，官媒的可信度既取决于人们对其作为一个媒体的信任，更本质的是对于政府的信任。人们对媒体的信任需要日积月累，想在重大事件面前说话有人听，就要在小事情面前说话有人听。现在微博上有些官媒有时发一些不符合自己身份，打破了自己底线的内容，在引起民愤的同时也丢掉了信任，那么在疫情来临的时候，说话就会没人听。因此官媒必须要认识到自己在心态和舆论引导中起到的重要作用，承担责任，坚守底线。

除此之外，政府本身也必须强力，值得信赖。例如上文提到的湖北省红十字会事件的发生，毫无疑问是对政府可信度的挑战。政府可信度归根结底体现在政府座位上，因此政府在重大公共卫生事件面前必须执行力强，举措迅速，公开透明，这既是博取信任的要求，更本质的说，这也是控制疫情的要求。

7. 结论与反思

面对重大突发公共卫生事件，心理健康是生理健康之外同样至关重要的一环。此次项目从某种意义上来讲，是抓住网友们几个可以用来表达自身心态与情绪的出口，并对其中挖掘出的信息进行分析。丰富的内心活动化为文字，高频词，关键词而最终变成图表上的数据，我们欣喜地看到大家渐渐减少的消极情绪，看到大家始终如一的对疫情的关切，对英雄的感激，对举国同心战胜疫情的信心。而对于如何在重大突发公共卫生事件中关注并调节广大人民群众的情绪，我们也得出了几点结论。

其一，正面的宣传是有效的。观察微博和知乎的走势图都能发现，疫情初

期的正面情绪里“感动，感谢”与“祝福支持”占了很大比重，而后续随着疫情渐渐趋于平稳，感人事迹密度下降，网友的不满情绪会出现小幅度上升。这说明大量正面的，积极的宣传引导，是能够有效调动网友情绪，促进其往正面发展的。热爱祖国与热爱家乡的网友还是占绝大多数，家国情怀层面的宣发能在个人情绪调节上也起到较好效果。

其二，疫情中段的心理健康建设不能掉以轻心。抗疫是持久的过程，疫情中段容易对疫情未完全放缓但已经持续了较长时间的居家隔离生活产生倦怠心理，同时因为还看不到形式立刻明朗的希望，网友情绪容易在此时产生一定波动。倦怠期的心理关注是必要的，也可以多加一些相关引导。

其三，不同人群的关注点与情绪走势会有较大差距。本次项目分析三类平台的不同受众，已经能见得其在关注点与情感走势上有较大差距。若要借助各个平台进行情绪引导，应好好把握其特定受众的特点与关注重点，实现情绪上的“精准扶贫”。