

疫情下的谣言传播可视化分析系统

xwzeng

1 导论

1.1 研究背景与意义

自 2019 年以来，新冠肺炎疫情对全球各个国家和地区的公共卫生系统、社会经济发展和人们的生活方式造成了巨大的冲击和改变。在疫情的蔓延过程中，大量谣言与虚假信息产生，为我国的疫情防控工作带来了严重的挑战。谣言的出现可能导致公众产生恐慌、焦虑和不良行为，从而干扰疫情防控措施的有效实施，甚至对社会秩序和稳定造成不利影响。因此，对疫情谣言传播现象的深入研究具有重要的科学意义和实际价值。

综上所述，本文将聚焦于疫情背景下谣言的主题内容与时间空间分布，使用 2021 年 11 月 7 日至 2022 年 2 月 18 日今日头条收集的 366 条疫情谣言数据进行探索与分析，并根据可视化结果总结得到谣言的主要特征，帮助公共卫生部门、媒体机构和社会管理者更有效地进行辟谣行为，保证社会稳定。

1.2 研究内容

基于上述背景，本项目完成的主要任务如下：

1. 对原始数据进行预处理，如缺失值处理、数据质量改善、特征提取等。
2. 根据文本数据信息进行谣言的大小主题挖掘与情感分析。
3. 根据地理数据信息进行谣言区域分布挖掘，分别从谣言数量、谣言热度、谣言情感三个角度进行分析。
4. 根据时间维度信息，分析随着时间变化，大小主题的趋势与相对重要性，以及谣言区域分布的变化。
5. 根据用户浏览记录与辟谣来源分别构建网络，对用户偏好、媒体对于扼制谣言传播的作用进行分析。

1.3 数据集简介

本项目使用的谣言数据来自于今日头条，包含 2021 年 11 月 7 日至 2022 年 2 月 18 日的 366 条谣言，变量名、含义与缺失值数量见表1。需要注意的是，项目指导文件中认为 `source` 变量代表谣言的来源，然而我在观察数据时发现 `source` 变量的含义实为辟谣来源。`user_0` 至 `user_17` 不是 1 个变量，而是 18 个用户浏览谣言的记录，其中 1 表示用户浏览了相应谣言，0 则相反。数据显示，用户对谣言的浏览行为是非常稀疏的。

表 1: 数据集格式

变量名	date	source	content	province	user_0 至 user_17	like
含义	谣言发布日期	辟谣来源	文本内容	涉及的境内省份	用户浏览行为记录	点赞量
缺失值	0	6	0	89	0	0

2 数据预处理

数据预处理的目标是提高数据的质量和可用性，以减少后续分析和建模过程中的误差和偏差。文本数据是非结构化数据，常常存在各种噪声和冗余信息。要想从中精准地提取出有效信息，数据预处理的重要性不言而喻。

2.1 source 变量清洗

`source` 变量为文本数据，每条谣言可能有 1 个及以上的辟谣来源，通常以“、”和“@”符号隔开。含有‘|’符号的表示谣言散播者为个人，而非平台或媒体组织。然而，并不是所有 `source` 数据都符合这个格式，数据中还存在一些错误和缺失值，我将分别对此进行处理。

1. 我将使用“@”分隔的数据修改为以“、”分隔的格式，便于后续数据的统一提取。
2. 有些 `source` 中包含“#”字符，可能影响到字符串匹配。如第 326（本项目中所有索引都从 0 开始）条 `source` 数据为“三联生活周刊 # 护肤 #”，“护肤”是一个小话题，与辟谣来源无关。我提取出含有“#”字符的 2 条数据，将其手动修改。
3. 第 305 条 `source` 数据是以“capd.gov.cn”开头的长句，末尾写了“消息来源：国家乡村振兴局”，因此我将辟谣来源更改为国家乡村振兴局。
4. 有些 `source` 中包含多余字符，如第 251 条的末尾多了一个右括号 “)”，应该删除。
5. 对于 6 条缺失值，我通过搜索谣言内容分别寻找到了真实的 `source`，将其添加到数据中。

完成上述操作后，`source` 变量没有缺失值且数据格式相同。我使用 `pd.get_dummies()` 将 `source` 转化为哑变量并保存至本地 `data/rumor_source_data.csv`，用于后文的网络建模分析。辟谣数量最多的前 5 家媒体为中国互联网联合辟谣平台、光明网、中国新闻网、福建辟谣和澎湃新闻，感谢他们为辟谣做出的卓越贡献。

2.2 province 变量清洗

项目指导文件中介绍到，`province` 变量的缺失值代表谣言不涉及境内的地理信息。但是我发现对于某些包含境内地理信息的谣言，`province` 仍然是缺失的，这说明原始数据并不完整。为了让数据包含的信息更丰富，我将根据谣言的文本信息手动为 `province` 添加数据。如第 53 条谣言内容包含“滦州市榛子镇麻湾坨村”，滦州市位于中国河北省，我就将这条谣言的 `province` 赋值为“河北”。

检查并补充完数据后，`province` 变量的缺失值从 89 条降低至 53 条，说明仍然有 53 条数据是不涉及地理信息的。这些无地理信息的数据多发生在境外，或是着眼于健康、国家制度方面，符合我们的常识。

最后，为了能在后续的地理信息可视化任务中使用 `pyecharts.charts.Map` 画出地图，我根据 `province` 的行政区划单位（省、市、自治区）在数据后方添加相应后缀。

2.3 content 变量清洗

`content` 数据中，有些谣言的文本信息过长，含有不少冗余信息，因此需要以某种规则从中提取出重要信息。一般而言，重要信息会存储在双引号 “” 之间，于是我首先尝试提取了 `content` 数据的双引号内部的文本，其中有 14 条谣言未能提取出相应信息。我查看了这 14 条谣言的情况，发现无法提取的原因有：1) 谣言为问句，且不在双引号内；2) 谣言的双引号写反了。我在这些条目的谣言信息上添加双引号便于文本提取。

接下来检查文本是否都提取正确。正确提取的文本中应该不会再包含双引号，然而第 269, 280, 351 条却依然包含了后双引号，说明这几条数据的提取存在问题。我查看后发现错误原因在于：1) 遗漏或多余了一个双引号；2) 单双引号使用不规范，双引号内部再进行引用应该使用单引号而非双引号。我将这 3 条文本修改规范。

最后，当双引号的内容较短时，很有可能双引号只起到引用文本或表示特殊含义的作用（而非强调重点），提取出的文本信息并不有效，我需要针对此问题进行一些处理。当某条谣言所有双引号中提取出的内容中，没有一条在 10 个字符以上时，我将其原始 `content` 文本提取出查看情况，发现：

1. 有些谣言本身长度就很短，信息依然有效，如第 20 条“叶庄全村转移隔离”，第 208 条“华侨大学停办”。
2. 另一些文本则确实存在问题，这是因为引号内都为：1) 地点名称；2) 阳性、阴性等核酸检测词汇；3) 当事人的简短话语或配图文案，如第 226 条“再见了兄弟们”；4) 专有名词，如第 320 条“触电”。

我对于其中有必要修改的文本进行手动处理，再删除 `content` 中的所有空格，`content` 变量就清洗完成了。最后，我提取了 `content` 文本中所有双引号内的内容，并用“,”连接为一个字符串作为数据的新变量 `core`。当然，即使这样做也不能保证提取的信息完全准确有效，不过这确实能减少许多冗余和噪音。

2.4 增加新特征

从已有特征中提取有用的新特征，分别为：

1. 增加用户浏览量信息：将所有 `user_id` 列对行求和作为新变量 `views`。
2. 对点赞量 `like` 做变换：`like` 的原始分布为明显右偏分布（图1），由于点赞量最小值为 0，故我对 `like` 采用稳健的对数加一变换，变换后的数据基本呈正态分布。

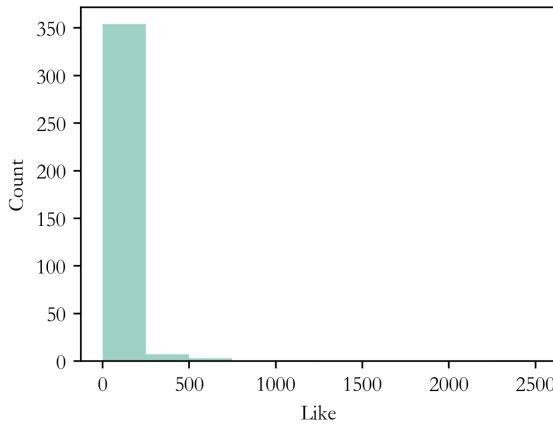


图 1: `like` 直方图

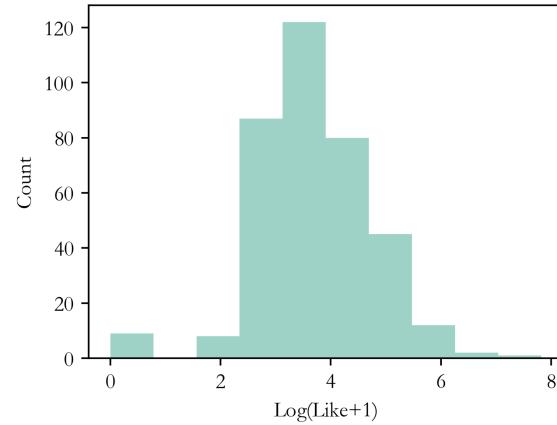


图 2: $\log(\text{Like}+1)$ 直方图

最后我将数据中所有 `user_id` 列保存至本地 `data/rumor_user_data.csv`，将除去所有 `user_id` 列的预处理后数据保存至本地 `data/rumor_text_data.csv`。至此，数据预处理部分圆满完成。

3 文本主题挖掘与情感分析

数据预处理过后，数据中表示谣言信息的变为两列 `content` 和 `core`，分别为谣言的详细信息和谣言的核心信息。前者虽然全面，但是冗余信息较多，可以用于挖掘谣言的大主题；后者则更加精准简练，适合用于细分主题（小主题）的挖掘。我将分别利用这两种信息挖掘文本主题，分析谣言特征。

3.1 谣言大主题挖掘

在这一节，我使用 `content` 列进行谣言大主题的挖掘与分析。

3.1.1 LDA 主题模型

Latent Dirichlet Allocation (LDA) 是一种用于主题建模的概率模型，可以用于发现文本集合中隐藏的主题结构。LDA 模型的基本思想是将文档看作是多个主题的混合，每个主题由多个单词的概率分布组成。通过 LDA 模型，我们可以推断出文档中每个单词属于哪个主题，并且得到每个主题的词语分布。

本项目利用 python 中的 `jieba` 和 `gensim` 库实现 LDA 模型，停用词文件见 `LDA/stopwords.txt`。我们首先利用中文分词工具 `jieba` 获取去除停用词后的 tokens 列表，然后利用 `gensim` 库创建 tokens 字典和 bag of words 列表，最后令主题数量为 6 拟合 LDA 主题模型。为了结果的可重复性，我将拟合好的模型保存至本地 `LDA/LdaModel6` 中，只需要使用 `LdaModel.load()` 即可载入模型。利用 `pyLDAvis` 可视化的结果如图3。

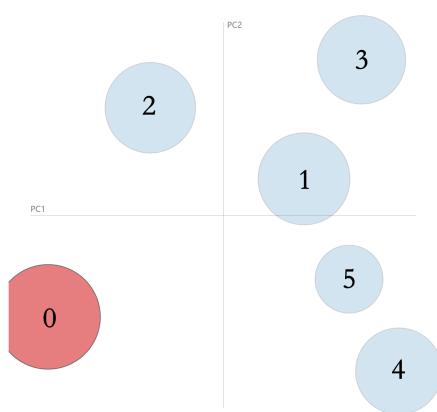


图 3: LDA 模型降维可视化效果

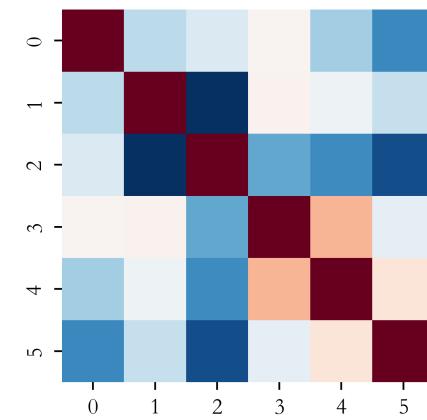


图 4: 各主题相似度热力图

图3中，6个大主题在降维后分隔较开，完全没有重合；图4中显示，主题3与主题4之间的相似程度稍高，其余主题两两之间的差别都较大，说明大主题的区分度较高，模型的拟合效果很不错。

6个主题的概率分布图见图5、图6，可以看出主题0, 2, 5的word distribution差异较大：主题1, 3, 4, 5的单词概率都集中在[0.001, 0.002]中间，说明这些主题包含的词汇在所有谣言中出现的频率比较低，是冷门主题；主题0, 2的word distribution明显峰度较低，尾部较厚，说明这些主题包含的词汇在所有谣言中出现的频率较高，是更为大众的主题。图7中显示，主题0, 1, 2的占比最多，超过20%，其次是主题3和5，占比最少的是主题4。这些主题各自的含义将在下一节进行介绍。

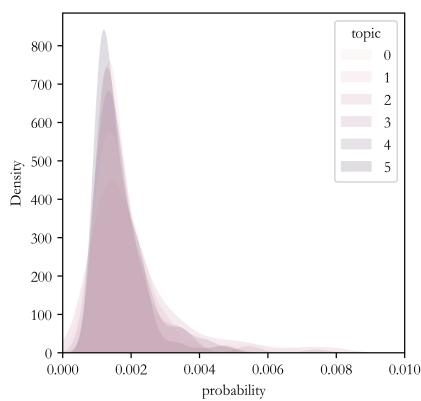


图 5: word distribution

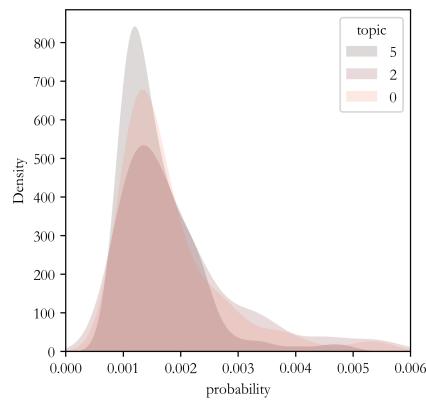


图 6: word distribution of 0, 2, 5

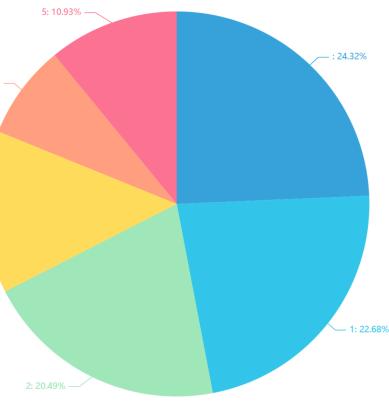


图 7: 6个大主题数量占比

3.1.2 词云图展示

上一节我们已经计算出了每个大主题的 word distribution，我据此画出相应的词云图（图8）。

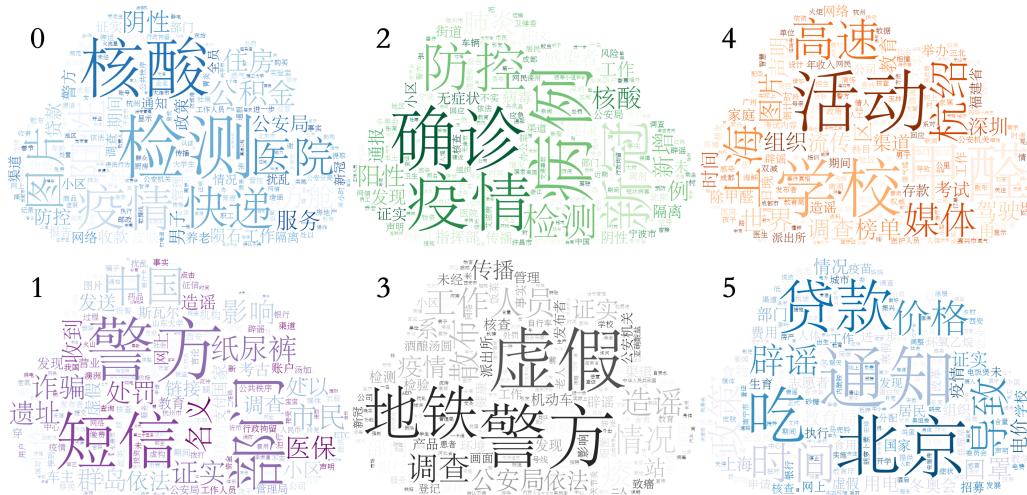


图 8: 6 个大主题词云图

根据 LDA 模型的结果和词云图，我们可以为这 6 个主题命名（表2）。第 0, 2 个主题都为疫情方向，但侧重点不同，前者偏向于服务，后者偏向于防控，二者的谣言数量占比之和超过 45%，可见人们对于疫情的关注度是很高的。第 1, 3 个主题都涉及到警方，但前者偏向于诈骗事件，如骗老人纸尿裤、医保等，后者偏向于对虚假事件的官方调查。第 4 个主题有许多生活相关词汇，而第 5 个主题主要与费用相关。

表 2: 6 个大主题描述

大主题	名称	占比	关键词
0	疫情服务类	24.32%	核酸检测、快递、医院、服务
1	虚假诈骗类	22.68%	短信、诈骗、纸尿裤、医保、账户、造谣
2	疫情防控类	20.49%	确诊病例、防控、阳性、通报
3	官方查证类	13.66%	警方、调查、公安局、工作人员、虚假
4	正常生活类	7.92%	活动、学校、甲醛、考试、驾驶
5	价格费用类	10.93%	贷款、价格、电价、用电、吃

3.1.3 情感分析

对每条谣言计算出情感分数，绘制情感分数直方图（图9），发现越靠近两端0和1分布越密集，这说明谣言中包含的情感大多比较强烈。然后分组统计每个主题的情感得分均值（代表情感倾向性，积极或消极）和情感得分绝对值均值（代表情感强烈程度），绘制条形图（图10、图11）。我发现6个大主题的情感倾向性都是消极的，其中大主题4, 5消极得不明显，可能是因为它们与负面事件相关性不大；大主题1, 3消极程度最高，诈骗事件会造成人们永久的物质和精神损失，非常合理；疫情相关的2个主题情感倾向性相近，受到新冠、封城等影响，消极程度也比较高。

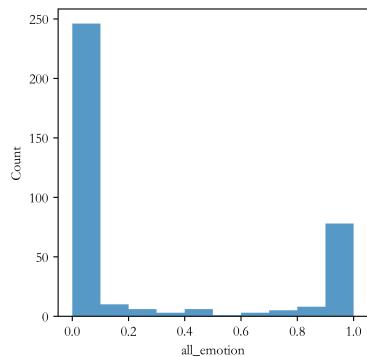


图 9: 情感得分分布

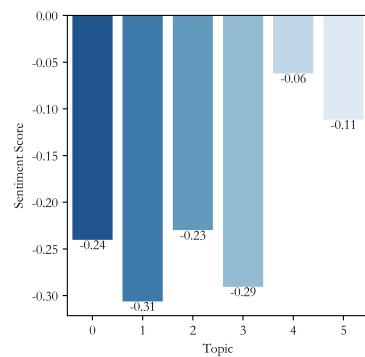


图 10: 情感倾向性

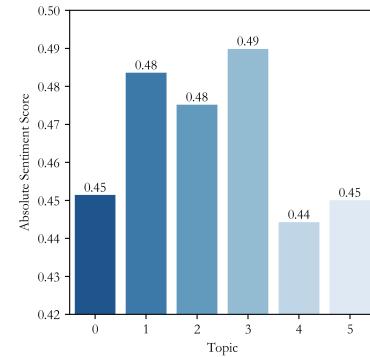


图 11: 情感强烈程度

3.2 谣言小主题挖掘

在这一节，我使用 `core` 列进行谣言小主题的挖掘与分析，主要的实验步骤与上一节是完全相同的，因此这里我不再赘述建模流程，主要展示可视化结果，并根据图表得出结论。

图12中，12个小主题降维后都分隔得比较开，说明小主题的 LDA 模型也拟合得很好。图13中，小主题 5, 6, 7 之间的相似度较高，主题 11 与其它所有主题的相似度都比较高，这可能是因为主题 11 是 background topic，其中的词汇会以一定概率出现在其它主题的文字中；且图14也告诉我们主题 11 的谣言数量占比最少，毕竟大部分文字都以有信息的主题为主，以无信息的 background topic（这里相当于地点的词典）为辅。

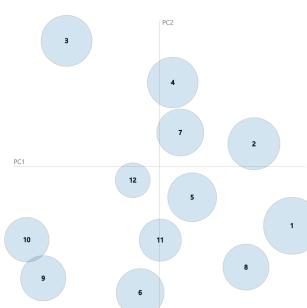


图 12: LDA 降维

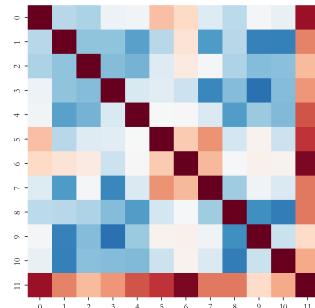


图 13: 各主题相似度

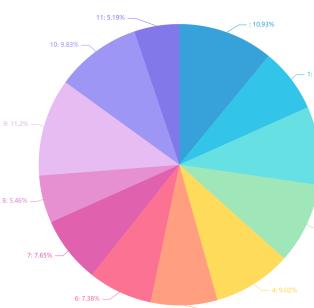


图 14: 小主题数量占比

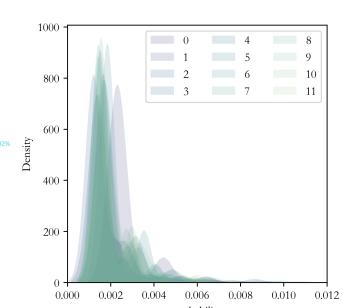


图 15: word dist

同样观察 12 个小主题的 word distribution（图15）和相应的词云图（图16），总结得到表3。

表 3: 12 个小主题描述

小主题	名称	占比	关键词	小主题	名称	占比	关键词
0	医疗类	10.93%	疫苗、治疗、医保	6	食物类	7.38%	吃、鸡蛋、超市
1	家庭类	7.38%	家庭、存款、收入	7	物资类	7.65%	物资、检测、小区
2	交通类	9.02%	隔离、抢票、驾驶	8	疾病类	5.46%	出血热、致癌
3	生活类	9.29%	养老、健康、考研	9	建筑类	11.20%	医院、方舱、核酸
4	信用类	9.02%	征信、修复、ETC	10	营业类	9.83%	营业、春节、税
5	通知类	7.65%	通知、指挥部	11	地点类	5.19%	很多很多地名



图 16: 12 个小主题词云图

在情感分析中，发现情感得分分布与大标题的相似（图17），但整体均值要低于详细谣言的情感均值。这是因为详细谣言的文本更长，`snownlp`能寻找到的程度词更多，相应地就有些失真。因此在后续的情感得分分析中，我将主要根据核心谣言的情感进行可视化和分析（图18）。只有小主题 11 的情感倾向明显为积极，但该主题为 background topic，并没有什么意义；小主题 4, 7 的情感倾向非常消极，与大标题情感分析中的结论相符。小主题 1, 4, 5, 7, 8 的情感非常强烈；0, 9 的情感强度较低（图19），可能是因为于医疗有关。

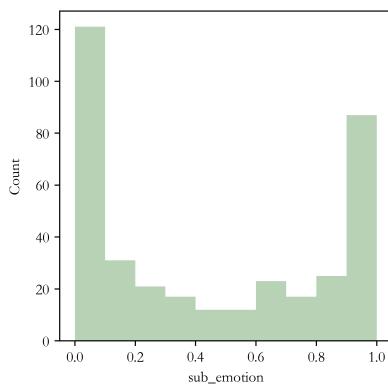


图 17: 情感得分分布

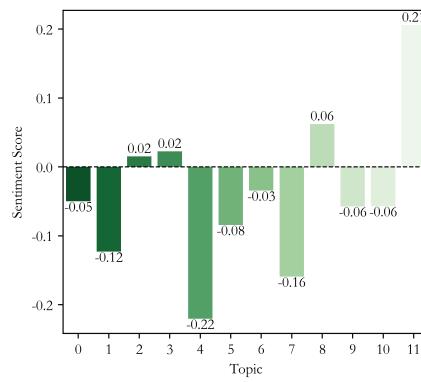


图 18: 情感倾向性

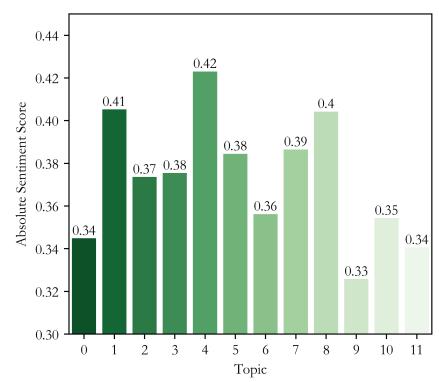


图 19: 情感强烈程度

实际上，大小标题之间存在包含关系。我分组统计了大标题下每个小标题的谣言条数，发现大小标题之间的关系可以总结为：大标题 0→ 小标题 0, 7, 11；大标题 1→ 小标题 3, 4, 6；大标题 2→5, 8；大标题 3> 小标题 2, 9；大标题 4→ 小标题 10；大标题 5→ 小标题 1。

在这一部分，我利用 LDA 模型挖掘出了双层主题，包括概括性的 6 个大标题，以及细分后的 12 个小标题，在后续建模中都有各自的应用价值。

4 地理信息挖掘

在这一部分，我对谣言的地理区域分布信息进行挖掘。我首先从整体角度查看谣言数量、热度（一般用点赞总量或点赞均值衡量）及情感（倾向与强烈程度）的分布情况，再从不同主题的视角分析谣言分布特征。

4.1 整体情况

按照省份，对所有谣言分组统计谣言的条数、点赞量总和与点赞量的均值，主要结论为：

1. 谣言数量：与浙江省、陕西省有关的谣言条数最多，分别为 63 条和 37 条；第二梯队的谣言来自于四川、广西、河北与河南，约为 20 条左右。我国西部、北部地区的谣言数量较少，中部和东部的谣言数量较多。
2. 谣言点赞总量：点赞总量包含了谣言数量和热度这两种信息。我国南部（广东、广西）、东部沿海（浙江、上海）、华北地区（辽宁、河北、北京、天津、河南、陕西）的点赞总量非常多，都超过 1000，说明我国人民对这三个地区的时事热点关注度很高。
3. 谣言点赞均值：如果要单纯比较每个省份的热度情况，应该以点赞量均值作为指标，此时只有华北地区的谣言热度远超平均，至少 100 赞/条；辽宁更是达到 281.90 赞/条。结合前 2 张图，辽宁的谣言数量并不多，但每条都能带来大量的热度，可谓“不鸣则已，一鸣惊人”。陕西就是辽宁的反面例子，谣言数量特别多，但点赞均值却相对较少，约 63.89 赞/条，说明陕西人民特别爱凑热闹，但是却翻不起大水花。

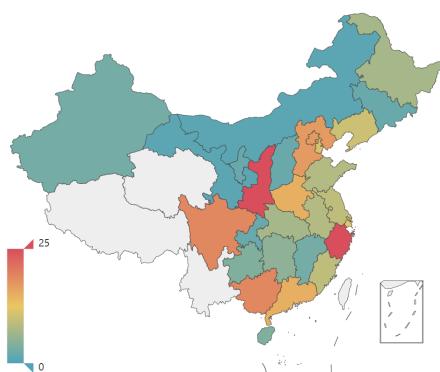


图 20: 谣言数量（整体）

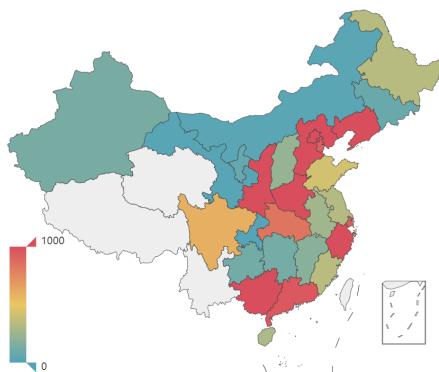


图 21: 点赞总量（整体）

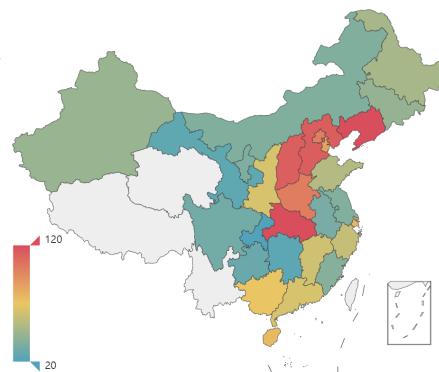


图 22: 点赞均值（整体）

4. 谣言情感倾向性：情感倾向性的数值位于 $[-0.5, 0.5]$ 之间，图中可以明显看出，大部分地区都呈现蓝绿色或黄绿色，即情感倾向为消极；只有广西地区谣言的情感倾向是明显正值，约为 0.22；贵州和宁夏其次，在 0.1 左右。这说明广西人民都是乐天派，在疫情这样艰苦的大背景下依然能保持积极向上的好心态，值得其他地区学习。重庆、甘肃、内蒙古和吉林的谣言数量本来就比较少，情感倾向又特别负面，可能一方面是因为这些地区受到疫情的负面影响较大，另一方面是全国人民对他们的关注度比较少。之后遇到这种情况时，我们应该团结各地人民，及时向这些困难地区提供鼓励和力所能及的帮助。
5. 谣言情感强烈程度：情感强度的数值位于 $[0, 0.5]$ 之间，从图中可以看出，我国西南地区（重庆、湖南）和北部地区的谣言情感最为强烈；华北地区其次。上一条我刚提到重庆、内蒙古、甘肃的吉林的谣言情感很负面，这里又显示他们的情感强度非常高，说明他们在疫情期间真的受苦了！与重庆、湖南相邻的湖北武汉本是疫情最初爆发的地方，情感强度却并不高，这可能是因为湖北在 2019 年末经历了许多艰难时刻，心态已经放平甚至麻木、习惯了；另一方面也是因为我们全国各地都为湖北提供了许多帮助。

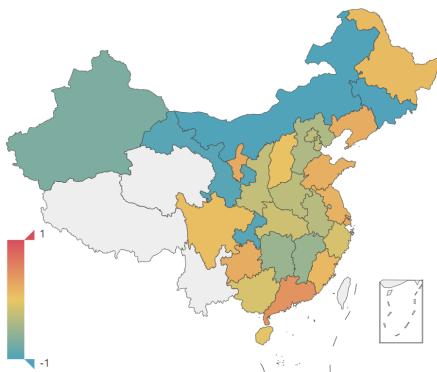


图 23: 核心谣言情感倾向 (整体)

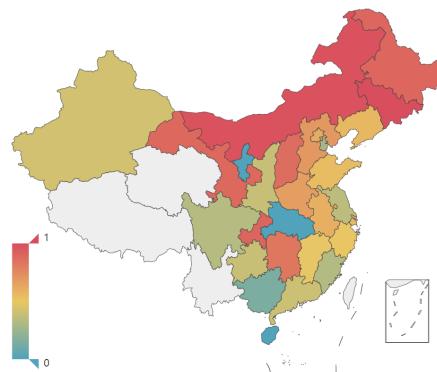


图 24: 核心谣言情感强度 (整体)

4.2 6个大主题的地理分布信息

首先看谣言数量分布。陕西和浙江这两个省份在第 0, 1, 2 个话题都有着很高的谣言数量，说明该地区的人们对于疫情的焦虑感最强，同时也比较天真单纯易被诈骗。四川省在第 0, 3 个主题谣言数量较多；京津冀地区在第 1, 2 个主题谣言数量较多。第 4 个主题（正常生活类）下，只有广西有比较多的谣言，说明广西在疫情背景下没有太多改变自己的生活，印证了前文说广西是乐观主义的结论。不过第 3, 4, 5 的谣言数量都不高，因为这 3 个主题的谣言数量占比本就不如前 3 个主题。

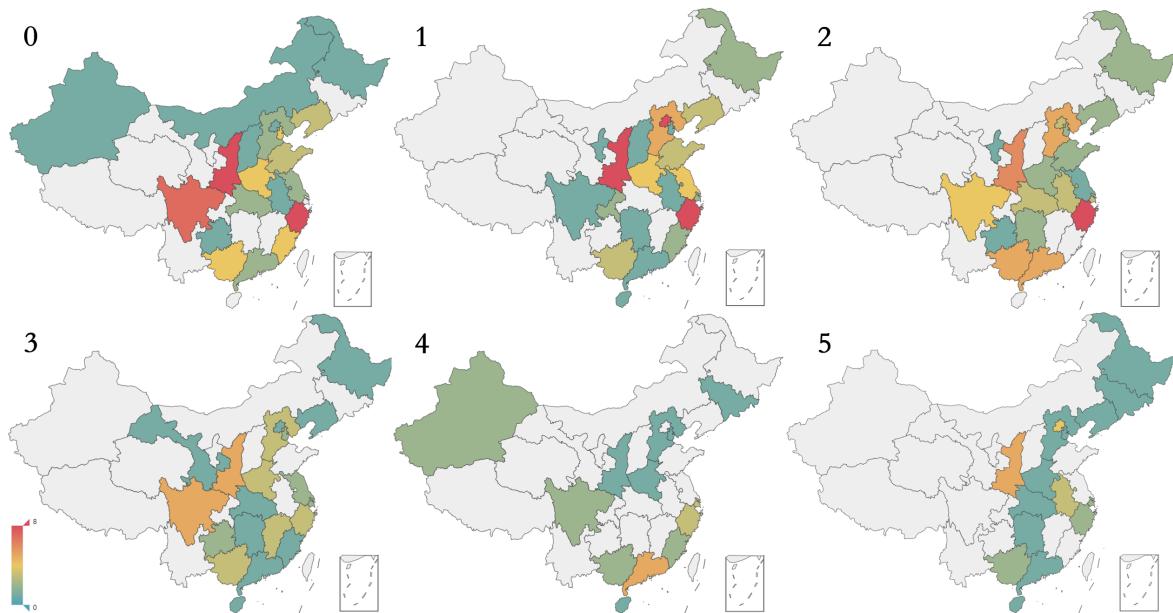


图 25: 6 大主题的谣言数量分布

接下来看谣言热度分布（这里使用点赞均值作为热度标准）。我国中部地区（湖北、陕西、山西、河北）在除了第 4 个主题下，都有着很高的热度，这大概是因为那段时间中部地区恰好疫情爆发。西藏在第 0 个主题的热度非常高，一方面是因为西藏的谣言条数很少，可能恰好其中的某一条就火了；另一方面是因为西藏的卫生条件和物质资源不是很丰富，在疫情爆发时很难做到医疗补给两手抓，从而引起全国人民广泛的讨论。我国华南地区（广西、广东）在前 4 个主题上都有比较高的热度。第 4 个主题只有天津市的热度是高的，而且很高，说明天津市在疫情期间有一则正常生活类的谣言引起了许多人的关注。

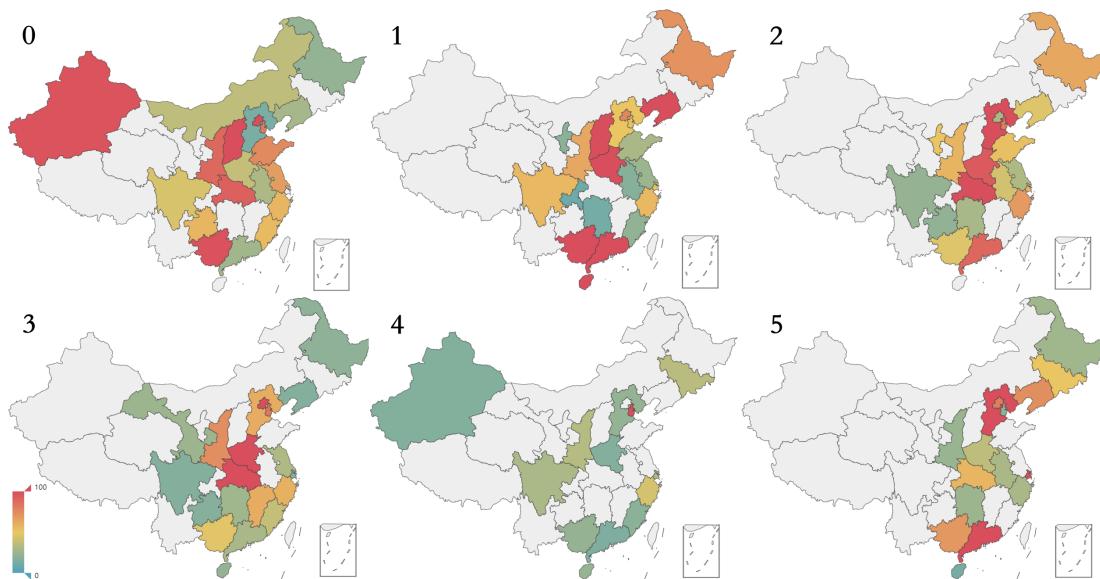


图 26: 6 个大主题的谣言热度分布

最后看谣言的情感倾向分布。第 4, 5 个与诈骗、疫情无关的主题的情感倾向明显更偏向积极。第 0, 2 个主题下，负面情感（深蓝色）占大部分区域，其中北京市的谣言情感都比较积极，体现了我国首都的在面对疫情时的信心；河南省的疫情服务类谣言情感比较积极，说明河南很期待受到医疗、食品物资等的援助；贵州、江苏、安徽则对于疫情防控类的谣言情感比较积极。

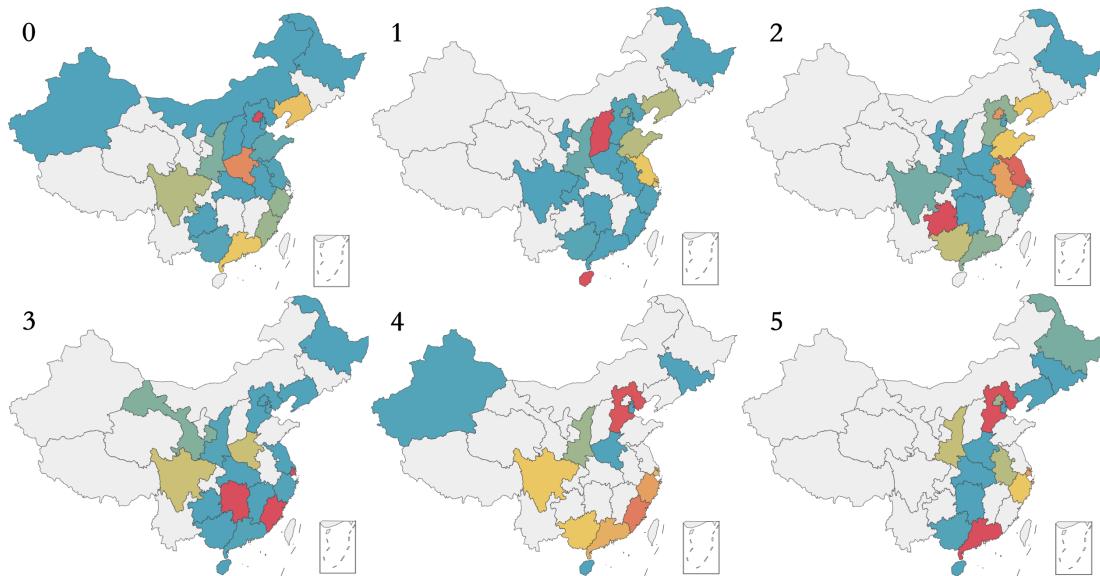


图 27: 6 个大主题的谣言情感倾向分布

本来还应该分析谣言的情感强烈程度分布，但是画出图后发现大部分区域的谣言情感都挺强的，极差在 0.1 以内，我就不再展示可视化结果了。

对于 12 个小主题的谣言区域分布，我将图画在 `3_ 地理信息挖掘.ipynb` 中了，结论与前面得到的差不多。而且含有境内地理信息的谣言一共就 313 条，再将其分摊到国内几十个省市和 12 个小主题，数据量不太够，可视化效果也不是特别好，因此我这里也不展示了。

4.3 不同省市的数据

分省统计谣言条数，降序排列后选择前 6 个省市绘制词云图（图28），发现不同省市的谣言确实存在一定区别。

1. 浙江省和陕西省的词云除了地名之外都比较相似，主要关注“核酸”、“阳性”、“确诊”、“检测”等内容。
 2. 四川的主要关注点虽然也在疫情上，但侧重于“方舱”，当初武汉的方舱医院建立完毕后，四川医疗队第一批入舱，对“方舱”重视是合理的；四川省的谣言还关于“考试”和“理工大学”。
 3. 广西除了关注疫情之外，也关注亲密关系，如“朋友”、“儿子”、“夫妻”等，说明广西人民非常重感情，传谣言都和友情、亲情、爱情有关（当然也有可能是因为八卦）。另外，还有一些相对奇怪的词，如“搂抱”、“亲嘴”等，看来广西人民擅长用亲密行为来表达心中感情。
 4. 河北省的谣言中，与疫情相关的重点词和其他省市都不同，是“死亡”，说明河北人民内心对于死亡是很恐惧的。其他高频词包括“公积金”（一种社会保障制度）、“出行”、“驾驶证”、“电动车”、“ETC”等，说明河北人民比较看重社会保障和出行方式。
 5. 北京市谣言对于疫情的关注度没有其他省市高，主要集中于2022年“冬奥会”的“志愿者”“招募”，以及“换发”、“养老”等看上去比较老年养生的词汇，独具特色。



图 28: 谣言数量 Top6 省市词云图

4.4 无地理信息的数据

无地理信息的数据存在着共同之处，我用 53 条不含境内地理信息的谣言绘制了词云图。图29中显示，谣言主要关于：1) 健康：与健康有关的谣言是适用于所有人的，如“致癌”、“酒后”等；2) 食物：我国常被网友称为“大吃货帝国”，与食物有关的谣言显然也不分境内地域，如“过午不食”、“酒后”、“鸡蛋”等；3) 境外：谣言有地理信息，但不在境内，如“澳洲”、“火山”等；4) 医疗：如“特效药”、“注射液”等。



图 29: 无地理信息的词云图

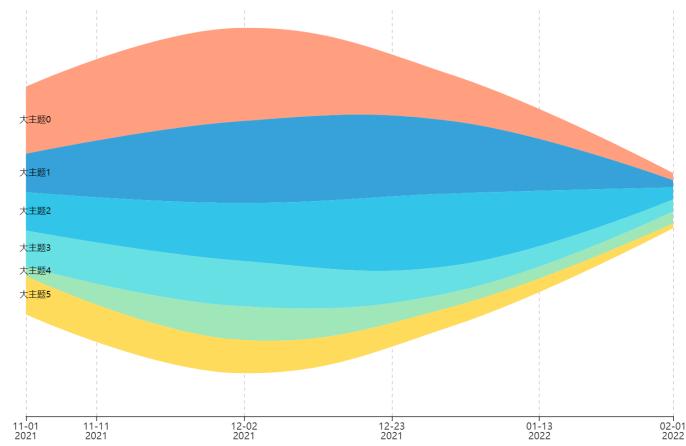


图 30: 大主题谣言数量河流图（频率：月）

5 时间信息挖掘

在这一部分，我将基于前文挖掘出的文本和地理信息，结合时间信息，分析谣言传播随时间变化的趋势。

5.1 主题河流图

主题河流图常用于展示随时间变化的多个主题或类别之间的关系和演变趋势，先前我们用 LDA 模型得到的结果正好可以使用主题河流图进行时间维度上的可视化。需要注意的是，原始数据中的采样频率为“天”，由于数据量较小，如果以“天”为单位绘制主题河流图，数据会非常稀疏，可视化效果也不好。因此，在作图之前，我会对数据按照“月”、“半月”和“周”进行重新取样，步骤如下：

1. 将数据中的 `date` 用 `pd.to_datetime()` 转化为时间数据，然后按照 `date` 从过去到现在进行排序。
2. 获取时期数据所在的日和月，如果按“月”采样，则令数据频段为当月的第一天；如果按“半月”采样，则根据所在日是否大于 15 判断数据所在频段为上半月还是下半月；若果按“周”采样，则以 [7, 15, 23] 作为每个月“周”的划分位点，将每个月的数据划分为 4 个频段。

用重新采样后的数据分别对 6 个大主题、12 个小主题进行可视化。

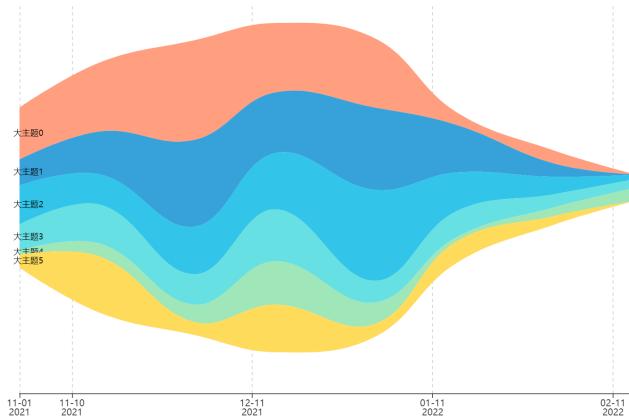


图 31: 大主题谣言数量河流图（频率：半月）

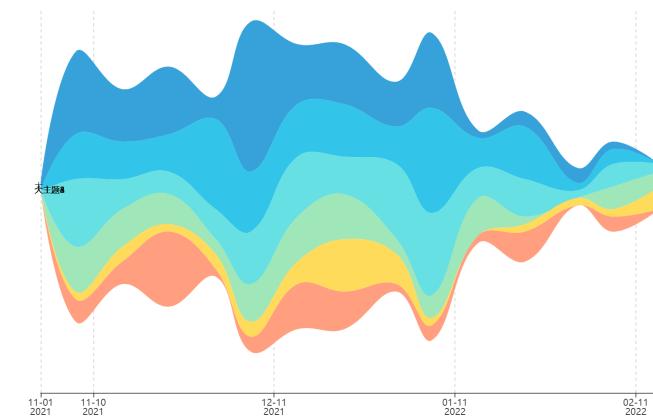


图 32: 大主题谣言数量河流图（频率：周）

5.1.1 大主题河流图

1. 谣言数量河流图

我按照月、半月、周三种频率绘制谣言数量的河流图，试图选择可视化效果最好的一种频率。首先排除的是以月为频率，因为通过比较图30和图31，就能发现前者过于粗糙，取样频率太小，一整月中间谣言数量的变化趋势难以看出。而究竟使用半月还是周作为频率就不太好选择，我决定只选择可视化效果最好的展示在报告中。

图32中显示，大主题0, 5 随时间变化的趋势比较相似，以一周为周期，周期性地降低或减少；而大主题3, 4 的变化趋势则恰好相反，在大主题0, 5 数量增加时，他们的数量就减小。这可能是因为这两类主题有相似之处，导致了此消彼长的情况。

2. 谣言热度河流图

这里以每个主题的平均对数点赞数作为谣言热度的指标。这里使用对数变换后的点赞数是为了让主题河流图变得较为平稳，更少受到离群值的影响（图33）。结果显示，除了 2022 年 2 月的下半月数据量减小导致的骤减的热度之外，所有主题的热度都比较平稳。

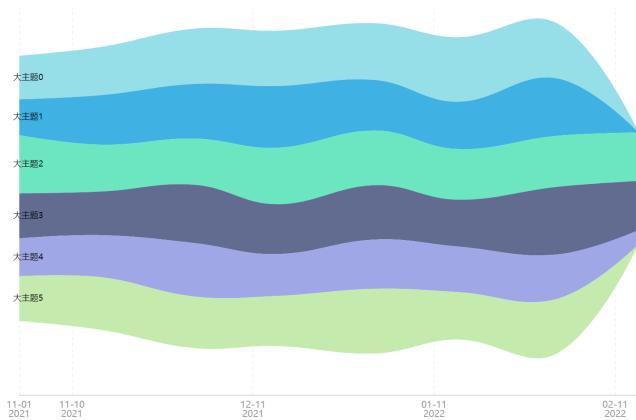


图 33: 大主题谣言热度河流图 (频率: 半月)

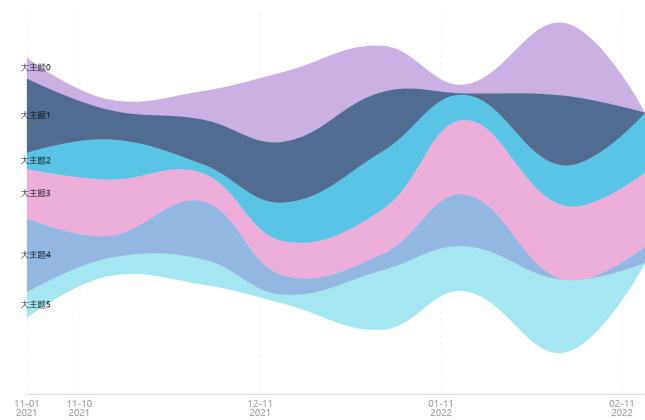


图 34: 大主题情感强度河流图 (频率: 半月)

3. 谣言情感强度河流图

由于主题河流对于负值不太好计算，所以就不对情感倾向性进行可视化了。这里以每个主题的平均情感得分为谣言情感强度的指标，以半月作为采样频率（图34）。大主题3, 5（官方查证和价格费用类）的情感强度随着时间变化持续变强，说明人们在疫情背景下，收入受到了持续的负面影响；大主题4（正常生活类）在 2022 年 2 月的情感强度骤降，而大主题0, 1 的情感强度却骤然加大，说明人们对于疫情的抗拒又到达了一波高峰；大主题2的情感强度基本保持着稳定。

5.1.2 小主题河流图

1. 谣言数量河流图

以周为频率绘制小主题的河流图，发现所有小主题谣言数量的趋势都一样，2021 年 11 月第 2 周相比于第 1 周而言，所有主题的谣言数量都增大；2021 年 12 月的第 1 周，相比于 11 月最后一周，所有主题的谣言数量又都减小；其他情况我就不一一列举了。说明当谣言被细分为小主题后，每个主题的数量在趋势上没有特别大

的区别，都是跟随着整体走：整体谣言数量上涨，每个主题的谣言数量也随之上涨。

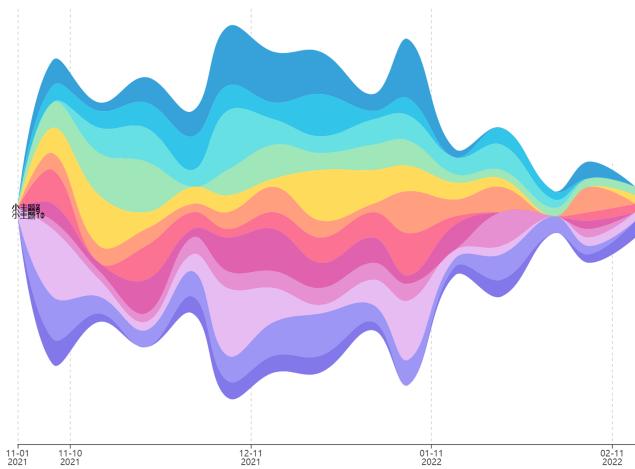


图 35: 小主题谣言数量河流图 (频率: 周)

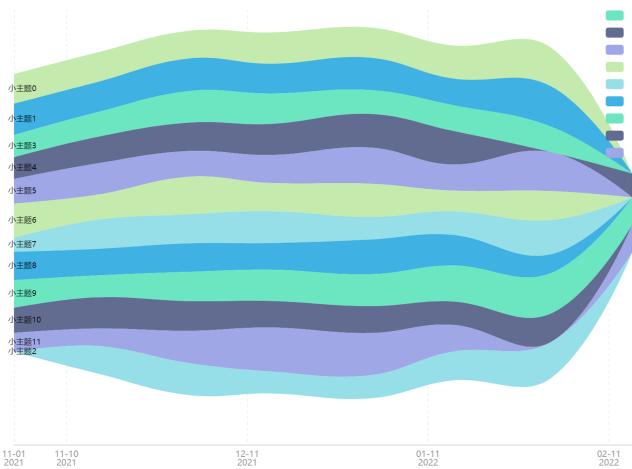


图 36: 小主题谣言热度河流图 (频率: 半月)

2. 谣言热度河流图

同样以对数点赞量的均值作为谣言热度的评价指标，以半个月为频率绘制主题河流图（图36），发现每个小主题谣言热度随时间变化的趋势比起谣言数量更加具有一致性，且波动性更小。这说明如果考虑进时间维度的信息，每个小主题的热度可能没有多大的差别。这说明人们爱八卦的天性是一视同仁的，不会因为谣言主题的变化就减少关注度。

3. 谣言情感强度河流图

以周为频率绘制情感强度的河流图（图37），发现这与河流真是非常相似了，在每个时点小主题的情感强度都在发生变化。我发现小主题 5, 6, 8 的情感强度随时间变化的趋势基本相同，小主题 1, 2, 3, 7, 9 的情感强度随时间变化的趋势比较相似。更多的主题河流图请见 [4_ 时间信息挖掘.ipynb](#) 文件。

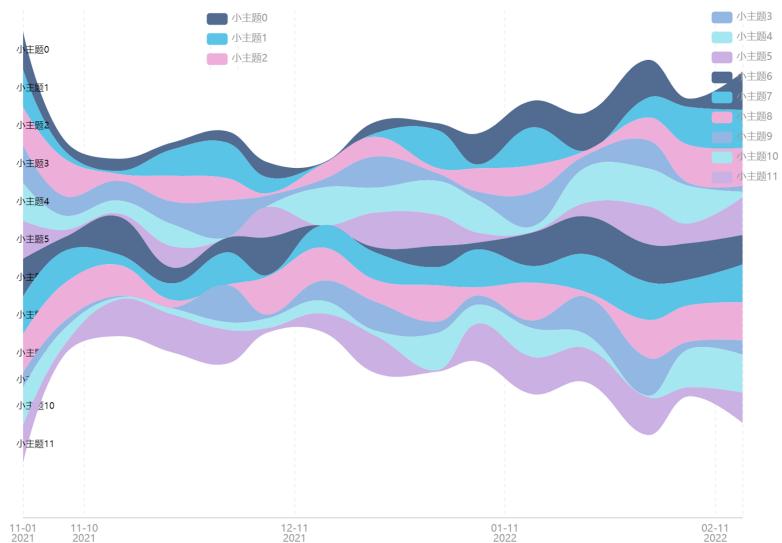


图 37: 小主题谣言情感强度河流图 (频率: 周)

5.2 词云图

不论是谣言数量、谣言热度还是谣言情感强度的主题河流图，我们都可以从中看出，谣言的变化趋势是先增大后减小。实际上，2021年11月，2日四川、河南疫情爆发，4日大连，10日北京，25日上海，27日内蒙古，且不断向外地扩散，疫情局势越来越紧张；2021年12月，2日云南，11日陕西，12日黑龙江，21日广西；而后的事件不再枚举。由此，我们知道2021年11月至2022年2月大致可以分为3个阶段：疫情初期、疫情爆发期和疫情中期。需要注意的是，这里的疫情初期不是整条时间线上2019年的疫情初期，而是在数据收集时间段内的时期中疫情严重程度较轻的时段。

结合数据本身和主题河流图，我将2021年11月至2022年2月从['2021-12-01', '2022-01-15']这两个时间点切开，分为3个时期（疫情初期、疫情爆发期和疫情中期），这3个时期的谣言内容应该有所不同。因此我尝试分别绘制出这3个时期的谣言词云图，看看谣言主题随时间的变化情况。



图 38: 疫情初期



图 39: 疫情爆发期



图 40: 疫情中期

这3张词云图中其实存在不少重复的词汇，如最主要的“检测”、“核酸”、“疫情”、“防控”等。下面对于3段时期的分析将尽量避开这些文本频率高的词汇，挑选出这些主题中最为特殊的词汇进行主题挖掘。

1. 疫情初期：2021年底的疫情还不算严重，只有几地有少量的阳性病例，许多人没有察觉到危机感，为了在社交平台上博眼球将阴性的核酸照片P成阳性，哗众取宠散播恐慌。因此，这段时期的谣言主题在于“检测”、“核酸”、“防控”、“平台”，并让“警方”、“公安局”等官方机构查证。
2. 疫情爆发期：此时各地疫情陆陆续续开始爆发，谣言中心转移至“疫情”、“确诊”、“病例”。
3. 疫情中期：此时各地疫情已经基本受控了，官方机构开始“辟谣”、“核查”，同时谣言中与日常生活相关的一些词汇开始频繁出现，说明人心惶惶的疫情爆发期已经过去。

5.3 随时间变化的谣言分布地图

5.3.1 谣言数量分布

在疫情初期，谣言主要来自于四川省，因为2021年11月四川正好疫情爆发；在疫情爆发期，谣言数量明显增加，华北、华南和东部沿海地区的谣言很多，这与疫情发展的轨迹在一定程度上是相符的；在疫情中期，由于数据量较少，有些地区没有数据，但是依然可以看出广西和浙江的谣言数量较多。

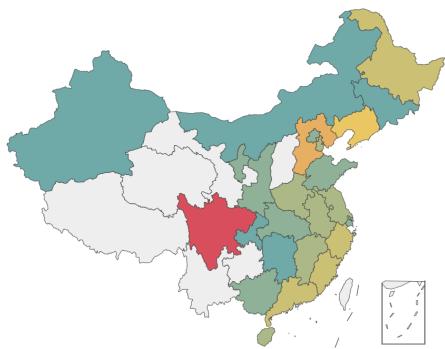


图 41: 疫情初期

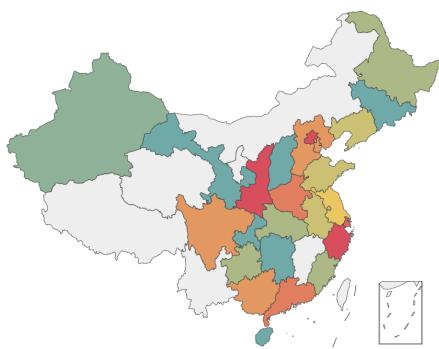


图 42: 疫情爆发期



图 43: 疫情中期

谣言热度与谣言数量在不同时期的分布区域比较相似，这里就不再详细阐述。

5.3.2 谣言情感倾向分布

在疫情初期，大部分地区的谣言情感还都挺积极，暖色调居多，只有除了黑龙江省的北部地区、中部地区明显消极。到疫情爆发期，可以明显看出地图色调向冷色靠近，即谣言情感变得消极了许多，这说明疫情为人们不但带来了身体上的病痛，还带来了心理上的折磨。疫情中期，疫情得到初步控制，人们又看见了希望的曙光，谣言的情感倾向也随之变得积极。

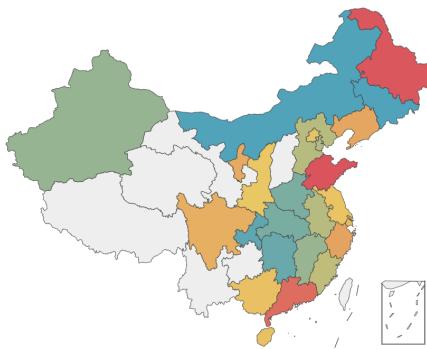


图 44: 疫情初期

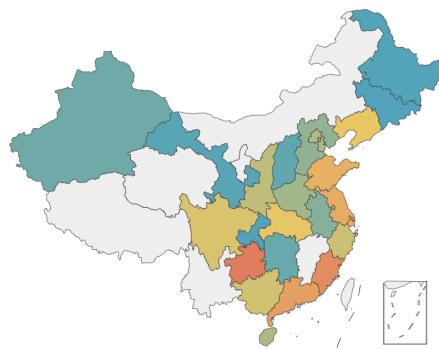


图 45: 疫情爆发期



图 46: 疫情中期

6 网络构建与分析

我们终于要用到数据预处理后得到的网络数据了！在这一部分，我将构建 2 个不同的网络：一是用户 → 谣言网络，二是辟谣媒体 → 谣言网络。需要注意的是，由于用户浏览谣言的行为和媒体进行辟谣的行为都是用户或媒体对谣言的单向操作，所以构建出来的网络是不对称的有向网络，而且还是二分网络。这对于建模非常关键，因为这意味着我们不能使用传统的、优秀的网络社群发现算法，而需要采取别的方法挖掘网络中的信息。

6.1 聚类

为了更清晰地挖掘数据中的特征，我将对用户、媒体通过谱聚类的方法降维。传入谱聚类的参数为相似度矩阵（这也是我们在这个项目中能够利用已有数据计算得到的）。要计算相似度矩阵，我们需要首先利用 0-1 矩阵计算出用户或媒体的余弦距离矩阵，然后利用公式 $\text{similarity} = 1/(1 + \text{distance})$ 的到相似度矩阵。

辟谣媒体的数量实在比较多，有 335 家，且矩阵非常稀疏，直接进行聚类效果一定不好。因此我首先根据辟谣行为发生数对所有媒体进行降序排列，然后选择所有辟谣行为发生数大于 4 的 18 家媒体进行研究。

我以用户浏览行为、媒体辟谣行为作为依据，利用计算得到的距离矩阵画出层次聚类的热力图。我们要使用的聚类方法是谱聚类，这里只是为了通过图像大致观察出应该聚为几类。

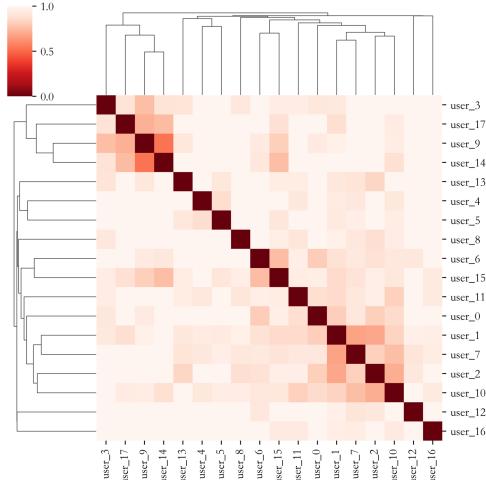


图 47: 用户聚类

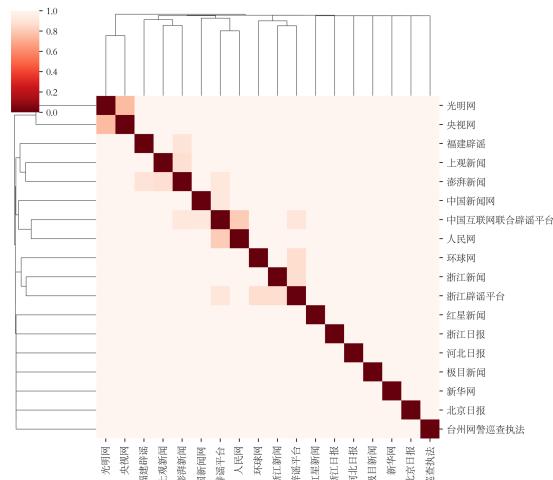


图 48: 辟谣媒体聚类

结合上图，我决定将用户和媒体都使用谱聚类聚成 5 类。聚类结果中各类数量很平均，说明聚类效果较好。

6.2 用户 to 谣言网络

我只保留了入度大于 0 的谣言作为节点加入有向图。我使用 `networkx` 库对网络进行可视化。为了让图片更美观，我将用户节点置于最内层，将入度为 1, 2, ≥ 3 的谣言从外到内放置。图49中，左图将谣言节点根据大主题重新着色，右图则将用户节点根据聚类结果重新着色。

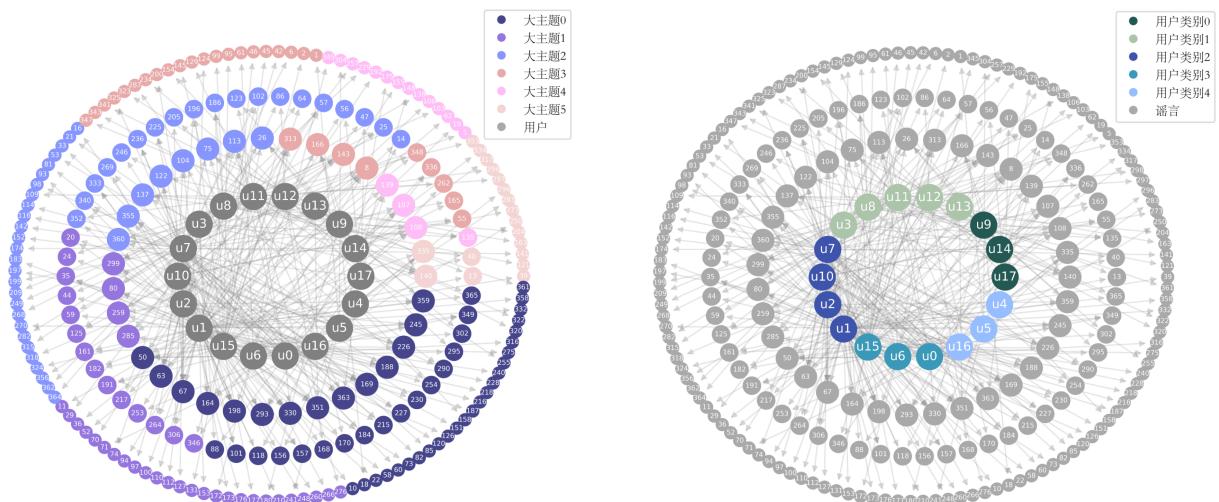


图 49: 用户 → 谣言网络

接下来，我利用 LDA 模型的主题降维结果，对谣言节点进行降维。我按照大主题分组统计的用户节点的出度（即主题节点的入度），以此作为有向边的权重，构建有向图。为了让图片更美观，我在对边进行可视化时，先

对边权重进行了对数变换操作，再以变换后的权重作为边的粗细和颜色深浅参数，如图50。

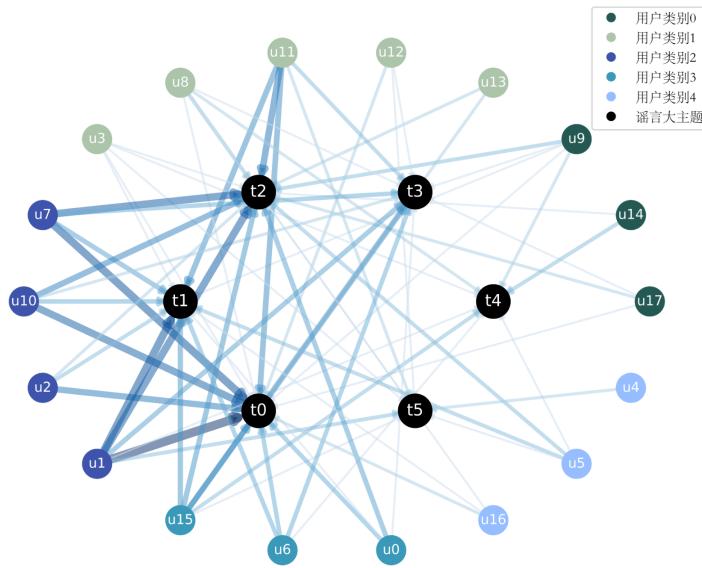


图 50: 用户 → 主题网络

可以从用户和大主题两个方面分析结果：

1. 用户：深蓝色的用户浏览量最大，蓝绿色的用户次之，再次是浅绿色的用户，说明这些类别的用户受到谣言的影响程度深；其他类别的用户浏览行为较少，说明他们不是人云亦云的人，非常理智。
2. 主题：大主题 0, 2 (都是疫情类主题) 更加受到用户关注，入度非常高；第二梯度是大主题 1, 3 (诈骗、官方辟谣类主题)，说明用户在疫情之外对这些主题也很感兴趣；最后是与疫情无关、生活消费方向的大主题 4, 5，它们的入度非常低，并不受用户关注和欢迎。

最后，我将用户也根据聚类结果进行降维，并将大小主题的包含关系也可视化在图51中。

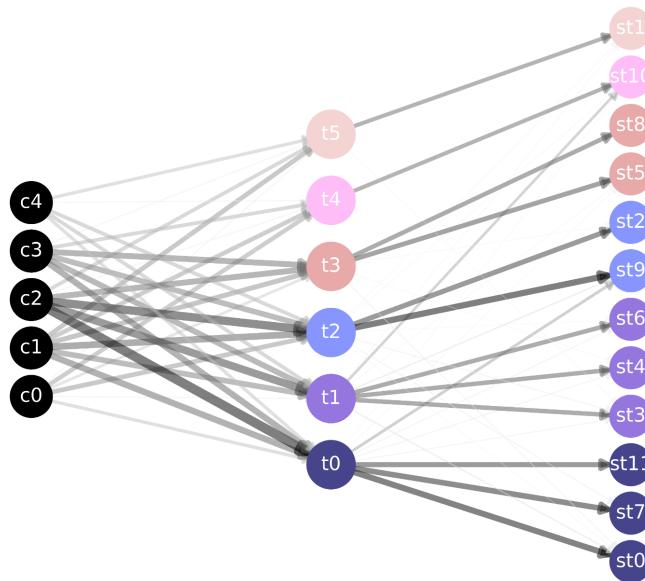


图 51: 用户群 → 主题网络

用户群体 0, 4 的出度较小，说明他们对于谣言的兴趣并不高；用户群体 1, 2, 3 的出度比较大，说明他们对于谣言有着较浓厚的兴趣。下面针对每个用户群体进行偏好分析。

1. 用户群体 0 (生活质量型)：浏览量不高，主要浏览大主题 2, 4 (疫情防控、正常生活类)，说明该用户群体对于生活的要求很高，不希望因为疫情封控、隔离等因素影响到他们的生活质量。而且该群体几乎完全不浏览大主题 1 (虚假诈骗类)，说明他们非常机智聪明，从不上当。
2. 用户群体 1 (江湖百晓生)：浏览量比较高，而且是对于所有类型的谣言主题广泛涉猎，雨露均沾，说明该用户群体是个来者不拒的吃瓜爱好者，非常热爱接收外界的信息。
3. 用户群体 2 (关注疫情型)：浏览量非常高，且着重关注大主题 0, 1, 2 (疫情类、诈骗类)，说明该用户群体受到谣言的影响程度高。该群体几乎从不浏览大主题 4 (生活类)，说明他们更关注负面的谣言信息。
4. 用户群体 3 (无忧无虑型)：浏览量比较高，主要关注前 5 个大主题，对大主题 5 (价格费用类) 几乎充耳不闻，说明该用户群体的经济条件很好，家庭比较富裕，才能从来不为消费感到困扰。
5. 用户群体 4 (随便看看型)：浏览量不高，除了大主题 4，其他主题都有涉猎，但兴趣度并不高。

6.3 辟谣媒体 to 谣言网络

用出度排行前 18 的媒体和所有入度大于 0 的谣言构建有向图，发现网络图比较零散，可视化效果并不好，因此我直接用将媒体降维至 5 个聚类类别绘制网络图，并将谣言节点根据大主题重新着色。图中显示媒体类别内部存在多个指向同一个谣言的媒体，说明之前的谱聚类确实将相似的媒体聚在了一起，聚类效果不错。

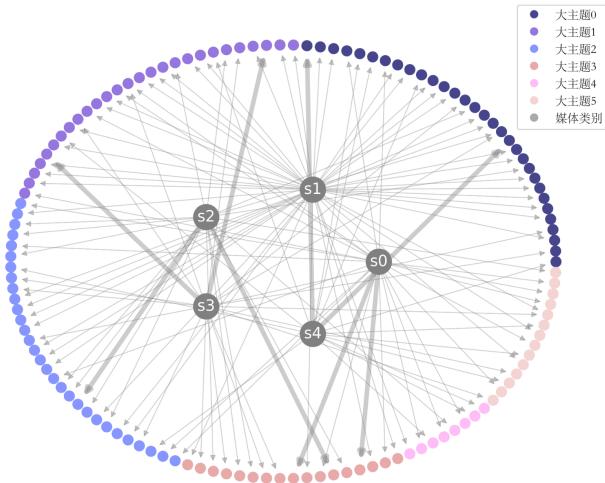


图 52: 媒体类别 → 谣言网络

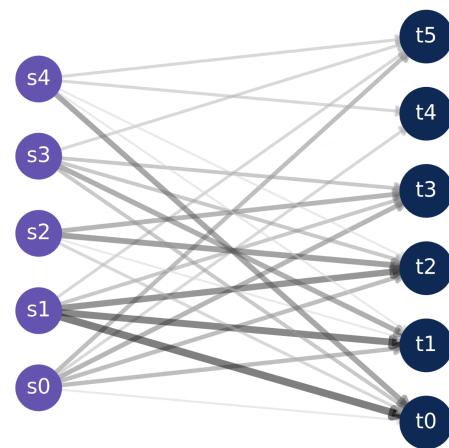


图 53: 媒体类别 → 主题网络

1. 媒体类别 0 (上观新闻、澎湃新闻、福建辟谣)：几乎从不对大主题 0 中的谣言进行辟谣。
2. 媒体类别 1 (浙江日报、河北日报、极目新闻、新华网、北京日报、中国新闻网、台州网警巡查执法、红星新闻)：主要对 0, 1, 2 这类谣言数量多的重点主题进行辟谣。
3. 媒体类别 2 (浙江新闻、浙江辟谣平台、环球网)：主要对大主题 2, 3 进行辟谣。
4. 媒体类别 3 (人民网、中国互联网联合辟谣平台)：辟谣的主题都比较严肃，且从未生活类谣言辟谣。
5. 媒体类别 4 (央视网、光明网)：专注于对大主题 0 (疫情服务类) 谣言辟谣。

7 总结与补充

在总结之前，我补充几张箱线图，用来展示谣言热度与辟谣媒体数量的关系，以及谣言在不同时期下的情况。图54说明谣言热度与辟谣媒体数量存在正相关关系，不过究竟是辟谣媒体数量导致的高谣言热度，还是高谣言热度导致的高媒体数量，我们并不知道。图55中显示，疫情爆发期时，谣言的热度最高，疫情初期与中期的热度相差不大。图56显示，疫情爆发期时，谣言的情感强度最高，疫情初期与中期的情感强度相差不大。

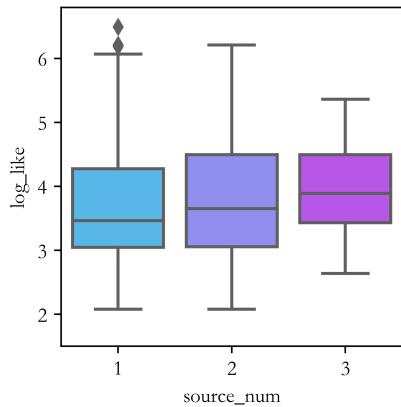


图 54: 谣言热度 vs 媒体数量

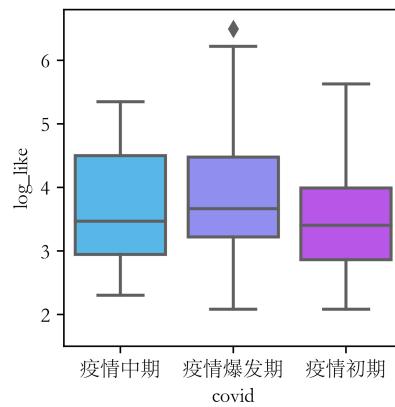


图 55: 谣言热度 vs 时期

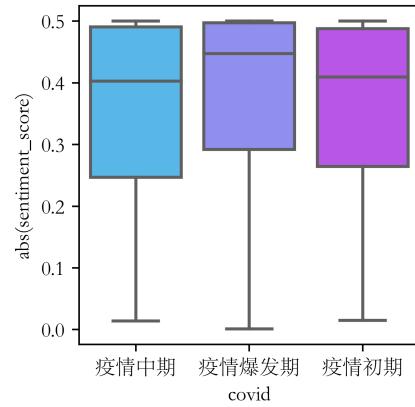


图 56: 谣言情感强度 vs 时期

最后，我会对这个项目中实现的任务与得到的结论进行简短的总结：

1. **数据预处理：**我对原始数据进行了缺失值处理、数据格式修正、特征提取与转换操作。
2. **文本主题挖掘与情感分析：**我对谣言的详细信息和提取出的核心信息分别拟合了 LDA 主题模型，得到了大小两个层次的主题。其中，大主题有 6 个，分别为：0-疫情服务类、1-虚假诈骗类、2-疫情防控类、3-官方查证类、4-正常生活类、5-价格费用类；小主题有 12 个，详见表3。我还对谣言文本进行了情感分析，统计比较了各个主题的情感倾向性与情感强烈程度。
3. **地理信息挖掘：**谣言的特性共有 4 个角度，分别为：谣言数量、谣言热度（通常用点赞总量或均值衡量）、情感倾向与情感强度。我首先对整体数据绘制区域地图并解释结果，然后将数据分割成 6 个大主题再次进行分析，接着绘制了谣言条数最多的 6 个省市的词云图，最后利用词云图展示了无地理信息的数据。
4. **时间信息挖掘：**我综合了第 2, 3 点中的文本与地理信息，再结合数据中的时间信息，分析谣言传播随时间变化的趋势。在主题河流图中，我从谣言数量、谣言热度（这里使用对数变换后的点赞均值）、谣言情感强度这 3 个角度入手，对大小主题的河流图都进行了观察和分析。然后，我根据主题河流图随时间变化的趋势，将 2021 年 11 月至 2022 年 2 月分割为了 3 个时期：疫情初期、疫情爆发期和疫情中期，分别绘制词云图展示文本主题。最后，我绘制了在 3 个时期下谣言各个特性的区域分布地图。
5. **网络构建与分析：**为了深入研究谣言内容、主题与用户、辟谣媒体之间的关系，我基于第 1 点数据预处理得到的用户与辟谣来源数据，构建了 2 个不同的二分有向网络，分别为用户 → 谣言网络和辟谣媒体 → 谣言网络。我使用谱聚类对用户和辟谣媒体进行降维，结合第 2 点 LDA 模型得到的结果，又在这 2 个网络下构建了新的子网络，分别为用户 → 主题网络、用户群 → 主题网络、媒体类别 → 谣言网络和媒体类别 → 主题网络。基于此，我总结了 5 个用户群（生活质量型、江湖百晓生、关注疫情型、无忧无虑型和随便看看型）与 5 个媒体类别的特点。