基于新冠疫情相关微博数据的主题挖掘和情感分析

1810055 刘汉青 1810057 刘焜鹏

南开大学数学科学学院

摘要

本文研究了自 2020 年 1 月 1 日至 2020 年 2 月 18 日期间，在新浪微博发表的与新冠疫情相关的微博数据。通过 jieba 分词、去停用词，计算词频并绘制词云。利用朴素贝叶斯方法，训练出基于新冠疫情话题的情感分类器，从而对所有文本进行 0-1 的情感倾向评分。对情感倾向数据进行聚类分析、核密度估计，并对社群情绪随时间的波动情况进行研究，发现网民对疫情消息的反应及时而敏感，同时在疫情初期受到了负面情绪的广泛影响。借助情感分析分类器，对微博话题进行探究，发掘出话题情感极性和用户关注度之间的强烈相关性。此外，借助 LDA 主题生成模型，挖掘出了 6 个微博内容当中的潜在议题。利用话题中提供的根词（root word），借助刘焕勇博士的研究，开发出与新冠疫情数据密切相关的领域情感词典，为他人在该领域内的研究提供支持。

关键词： 主题提取、情感分析、舆情分析

一、 研究背景

在新型冠状病毒引发的肺炎疫情爆发前后，各社交网络平台产生了大量的与疫情现实、个人情绪、知识科普有关的文本表达。除去新浪微博的 tag（用##包括的内容）之外，文本之间是否存在强烈的主题联系？用户的文本表达背后表达了怎样的个人情绪？而随着时间的推移、疫情信息的发布、疫情发展态势的改变，公众的集体心理和情绪又发生了怎样的改变？情感极性对微博用户的关注度有着怎样的影响？对于疫情相关的中文表达，能否生成一本针对本领域的情感词典？本研究借助自然语言处理领域的机器学习方法，对上述问题进行了评估，并对本领域的其他研究提供了支持。

二、 数据预处理与描述性分析

1. 数据来源

本题数据来源于 https://www.kaggle.com/liangqingyuan/chinese-text-multi-classification 提供的公开数据，摘录自北京市经济和信息化局、中国计算机学会大数据专家委员会联合主办的科技战疫·大数据公益挑战赛。该赛题也是第二十六届全国信息检索学术会议 （The 26th China Conference on Information Retrieval, CCIR 2020）评测大赛赛题。数据采集自 2020 年 1 月 1 日至 2020 年 2 月 18 日期间，通过疫情相关的关键词，提取从微博收集的微博文本和有关信息。数据中包括微博 id、微博发布时间、发布人账号、微博中文内容、微博图片（地址）、微博视频（地址）。其中 100k 条数据提供了先验的微博中文内容的情感标签，以 -1 表示消极情感，以 1 表示积极情感，以 0 表示无明显情感，用作监督学习的训练数据。同时提供了 10k 条无情感标签的数据用于模型测试。

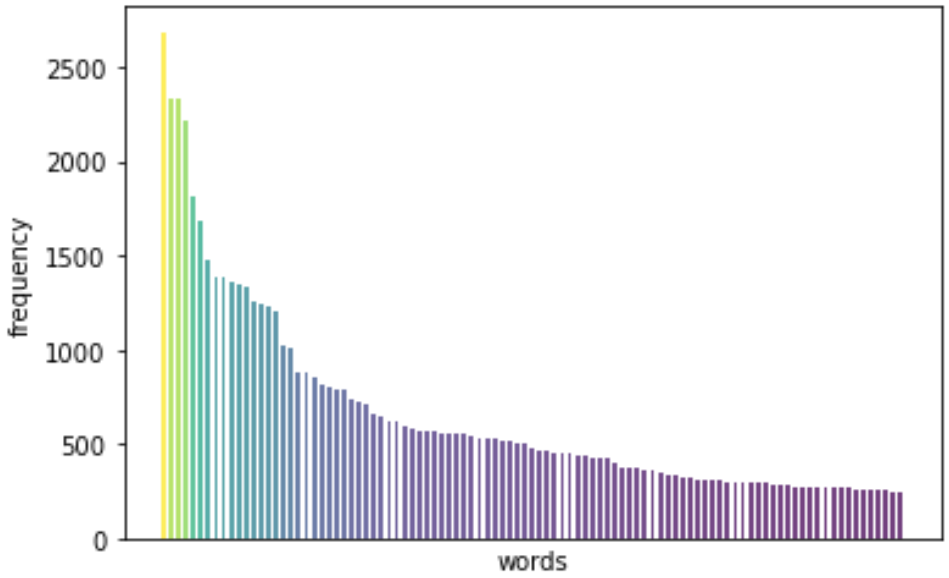
此外本题还采用了 www.kaggle.com 提供的数据集，包括疫情期间微博热搜话题的文本内容及其浏览量。本次研究通过分析这些话题的情感倾向和浏览量之间的关系，来研究极端情感与用户关注之间的潜在联系。

2. 数据预处理

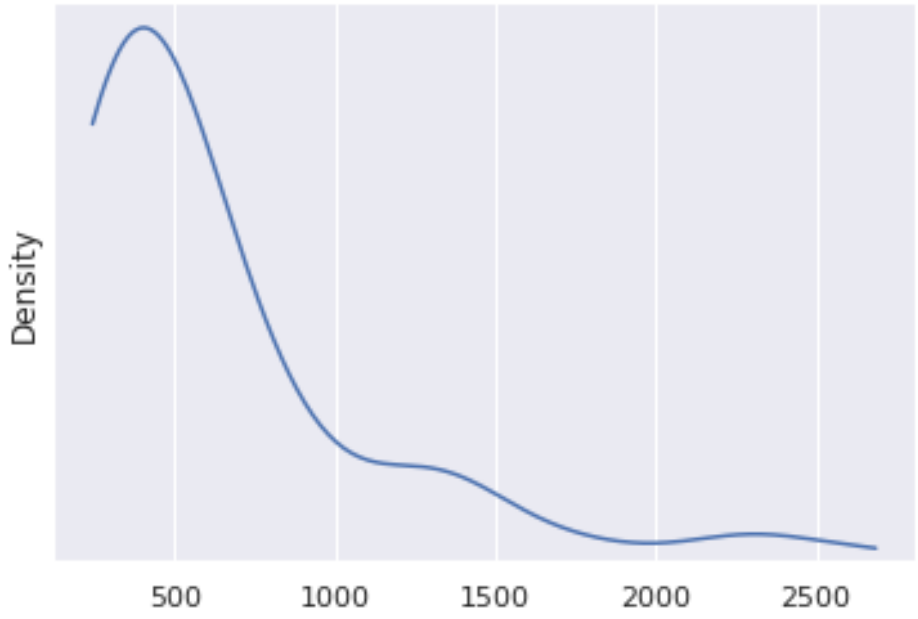
以 UTF-8 编码读取.csv 文件，借助 jieba 库提供的中文分词工具，基于哈工大停用词表、百度停用词表和 CSDN 论坛提供的常见中文停用词整理，对文本进行精细模式分词，并去除停用词。对获得的[词汇：词频]矩阵进行观察，将遗漏的停用词手动添加到停用词表中，并重复上述分词操作直至高频词中不再出现明显的停用词。

3. 描述性分析

获得[词汇:词频]矩阵，并绘制柱状图、核密度曲线和词云图，描述文本中不同词汇的出现频数和互相比较关系。



图一 词频前 100 位的词汇频数分布



图二 词频前 100 位的词汇频率核密度估计（KDE）

图三 疫情相关词汇词云图

词云图结果显示，“武汉”成为微博用户讨论的最大焦点，这符合疫情爆发初期全国网民集中关注武汉在疫情防控、患者救治、物资供给等方面议题的现实。此外，“中国”、“加油”等词汇则体现了网民在抗击疫情过程中团结一致的乐观态度。偶然出现的数字采集自每日发布的疫情发布数据，体现了网民对每日疫情数据的关心。

结合直方图结果与核密度估计，除上述提到的少量超高频词之外，剩余词汇出现的频次较低，且呈现较为均匀的分布。这现实了文本数据当中词维度的相似度较低，有较好的研究价值。

三、 基于 LDA 主题生成模型的微博主题提取

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）主题生成模型，借助贝叶斯定理和两次概率抽取进行文本的相似性评估，并根据文本间相似性提取存在较大共现可能的主题词，从而实现对文本主题词的提取，发现文本中潜在的共同主题。

对分词结果创建词语词典，并对每个单独的词语赋予索引。借助索引，将语料转换为用索引值标记的矩阵，并在此矩阵上进行 LDA 模型训练。因为语料数据庞大，所以选择对语料仅进行 5 次遍历。

1. 主题个数选择

模型参数中需要人为设定文本中包含的主题数，通常这一设定基于数据的分布情况和研究者的个人经验。本次研究中引入困惑度指标（Perplexity）和模型一致性（Coherence），对不同主题个数下训练出的 LDA 模型的效度进行评估。

$$\mathrm{Perpexity}(\overline{\mathcal{W}}|\mathcal{M}) = \prod_{m=1}^{M} p\,(\widetilde{\overline{w}}_{\bar{m}}|\mathcal{M})^{-\frac{1}{N}} = \exp\left(-\frac{\sum_{m=1}^{M}\log p(\widetilde{\overline{w}}_{\bar{m}}|\mathcal{M})}{\sum_{m=1}^{M} N_m}\right)$$

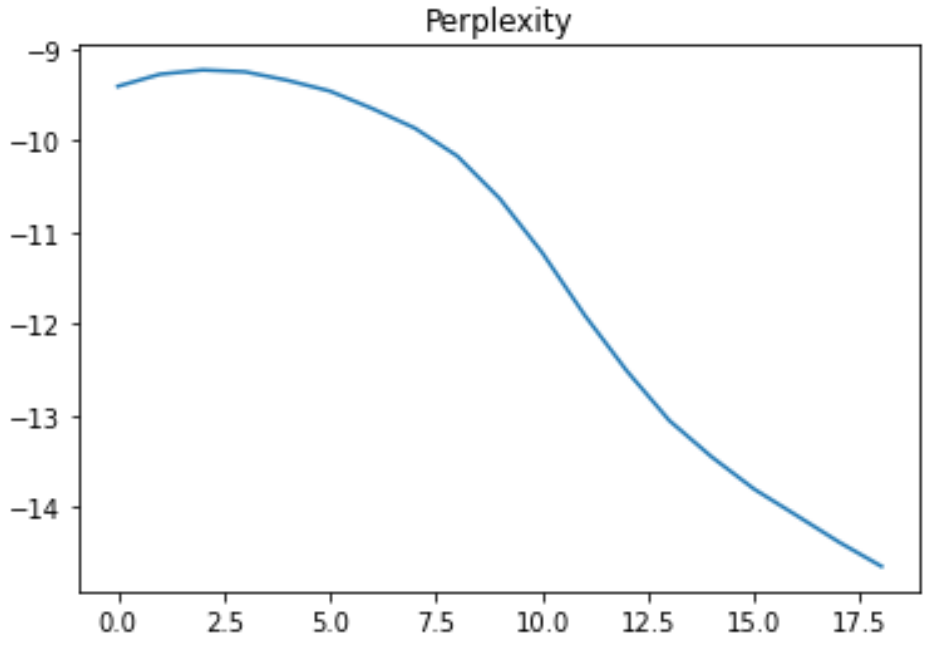
困惑度 Perplexity 的计算

$$\mathrm{Coherence}(z; S^z) = \sum_{n=2}^{N}\sum_{l=1}^{n-1}\log\frac{D_2(w_n^z, w_l^z)+1}{D_1(w_l^z)}$$

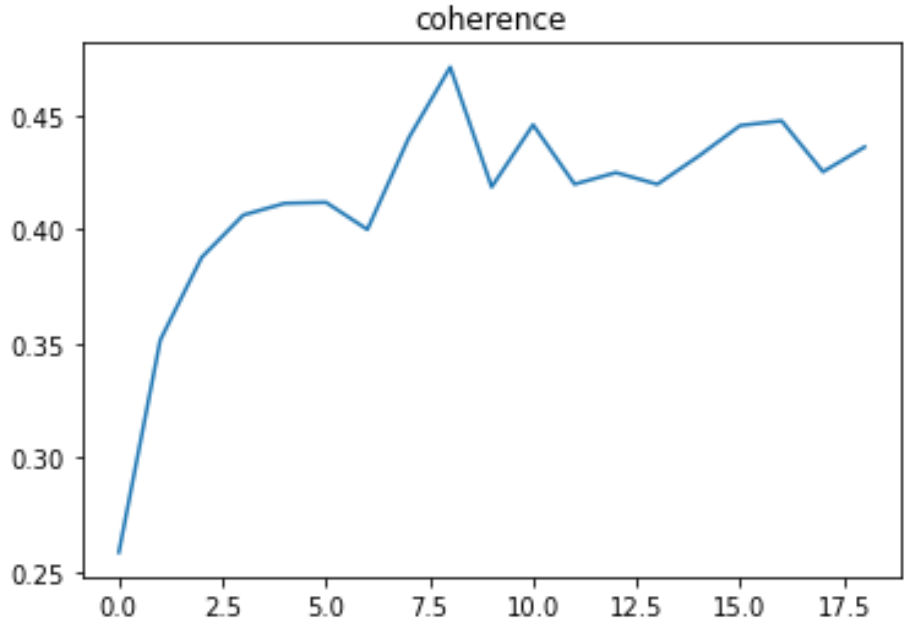
模型一致性 Coherence 的计算

通常，困惑度指标（Perplexity）越低，一致性指标（Coherence）越高，暗示该主题数下的 LDA 模型拟合效果越好。

本次研究度量了主题数从 1 到 20 的参数设定下，不同 LDA 模型的困惑度和一致性。



图四 不同主题个数模型的困惑度 横坐标：主题个数 纵坐标：困惑度



图五 不同主题个数模型的一致性 横坐标：主题个数 纵坐标：一致性

结果显示，困惑度在主题数超过 7 时出现连续且显著的下降趋势，暗示过拟合情况将变得越来越严重。一致性度量则显示，当主题数选择 7 时一致性最好，选择 5 时一致性尚可。最终，对主题数为 5 和 7 的 LDA 模型分别进行了评估和分析。

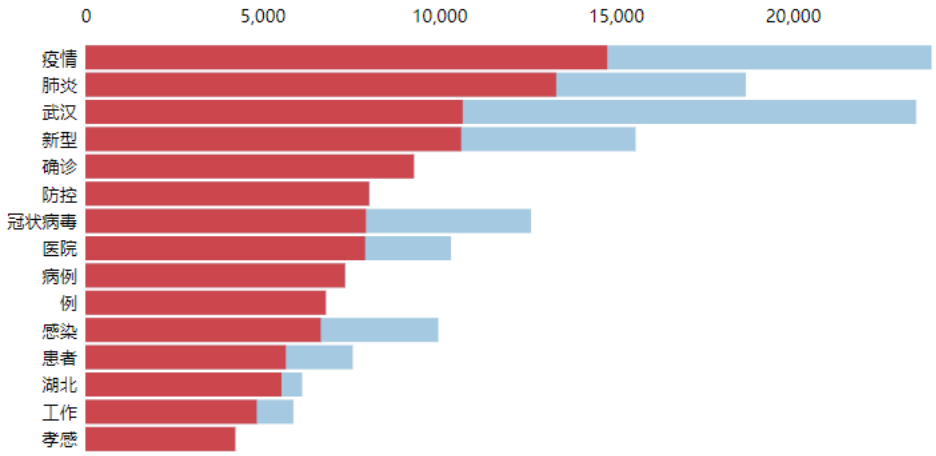
2. 主题评估与分析

2.1 主题分布情况评估

图六 5 个话题（左）和 7 个话题（右）下的模型分布情况

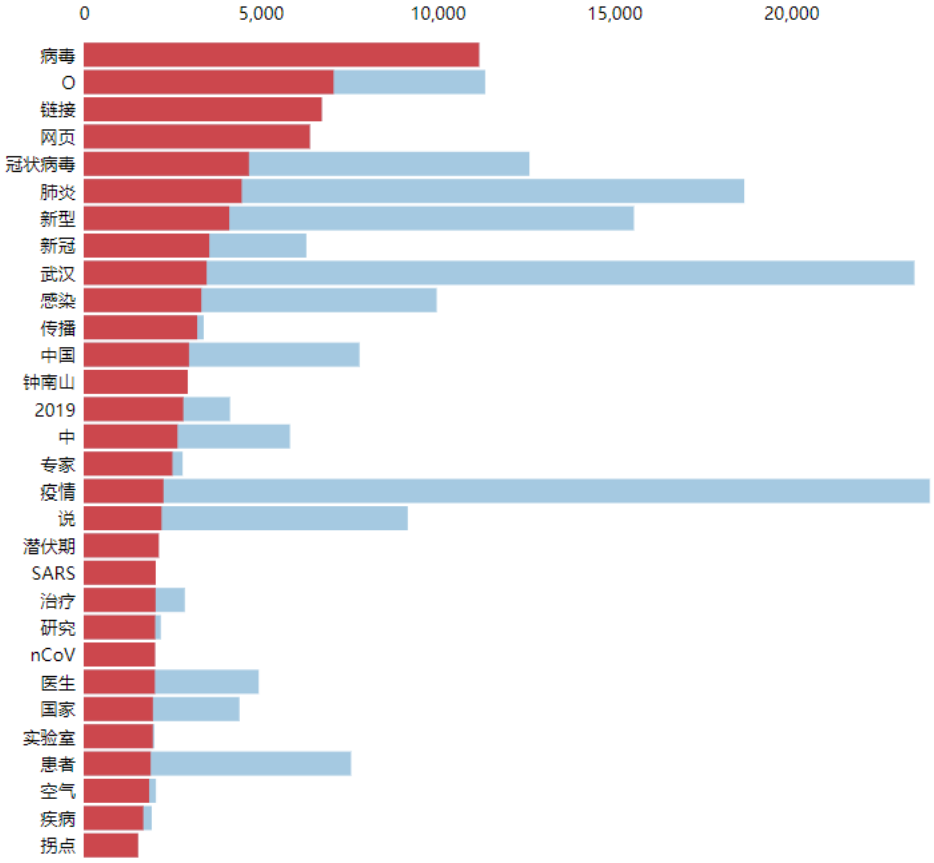
上图显示了 5 个话题和 7 个话题下，不同话题的分布情况。其中圆形的大小表示该话题的总体出现频数，圆形之间的相对位置关系表示话题间的联系，距离越近，关联越明显。某些话题间可能出现重复。评估结果表明，5 个话题和 7 个话题下，话题间虽然存在一定程度的重合，但仍表现出较好的独立性。其中，7 个话题使得某些较为相近的小话题得以发掘，使得数据得以进一步细化。

2.2 主题内容分析



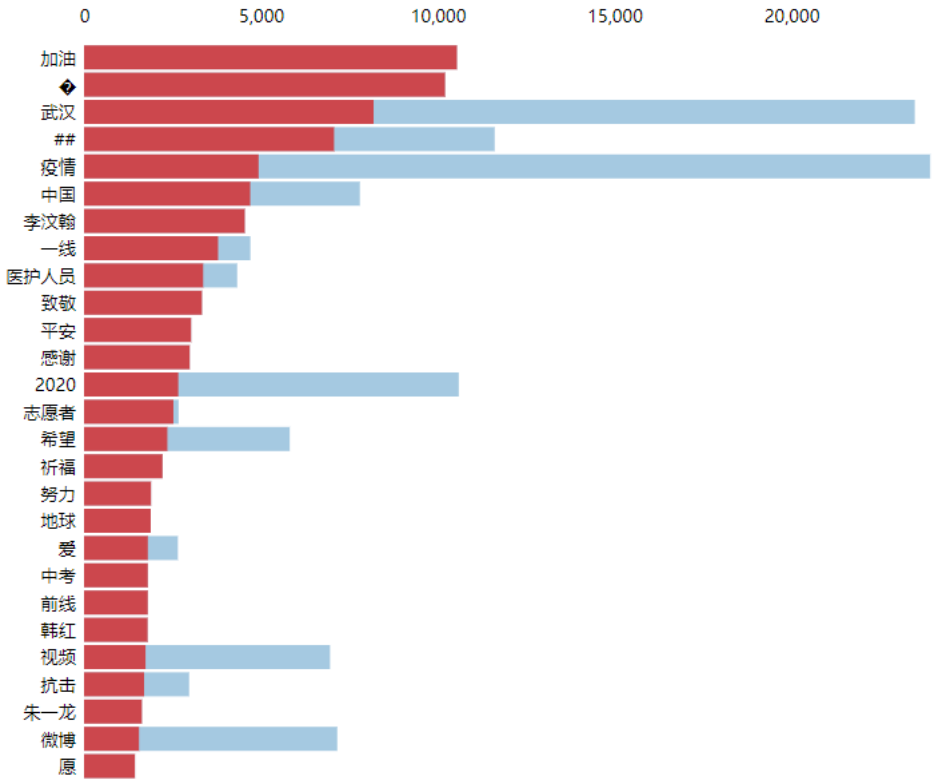
图七 话题一：疫情数据发布包含的重点词汇及频数

话题一：疫情数据发布。图七中，话题中的词语组合暗示出文本包含大量形式统一、严谨的新冠肺炎相关术语，较可能是官方发布的疫情数据。这一话题指出疫情期间关于疫情发展变化、感染确诊人数、地理位置发展等问题受到人们的广泛关注。



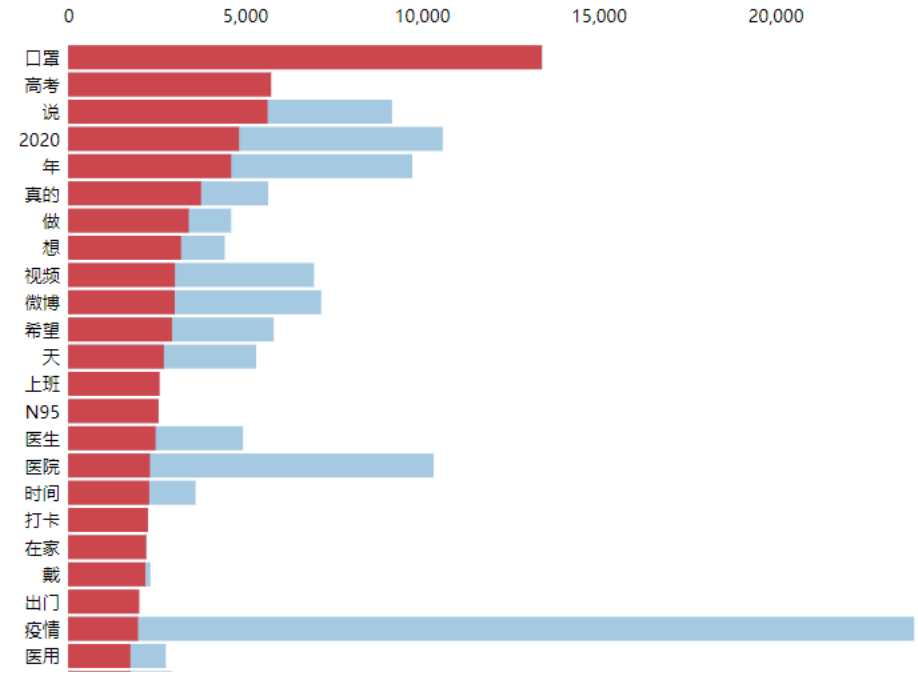
图八 话题二：科学问题研讨包含的重点词汇及频数

话题二：科学问题研讨。相比于话题一，话题二中“SARS”、“专家”、“潜伏期”、“实验室”等词汇的出现，暗示了话题中涉及关于肺炎类型、肺炎传播途径、肺炎发病原理、肺炎病毒溯源等问题的公共讨论。值得注意的是，钟南山院士在话题当中高频出现，这和新冠疫情期间他的积极公众互动和广泛的社会信任有关。



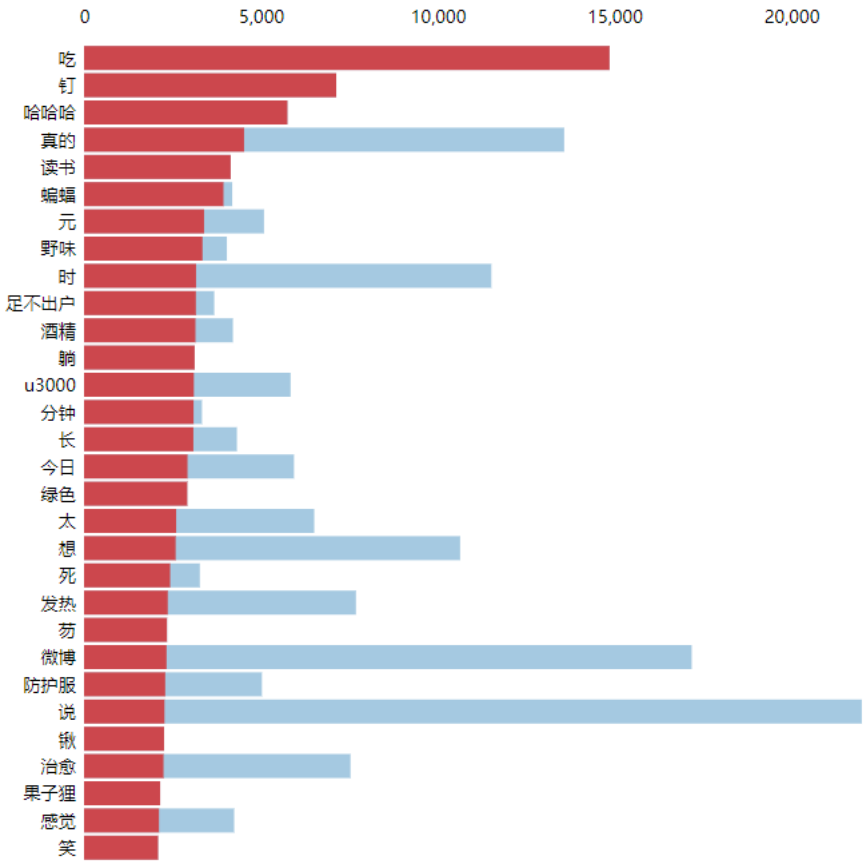
图九 话题三：公益与社会情感参与包含的重点词汇及频数

话题三：公益与社会情感参与。话题三中出现了大量的积极情绪词汇，如“加油“、“致敬”、“平安”、“希望”、“祈福”等，表达了社会公众对于疫情得到控制、患者得到救治的美好祝愿。“韩红”、“朱一龙”、“李汶翰”等词语则暗示了明星人物在疫情期间承担社会责任、参与社会公共事业的行为收获了广泛的社会讨论。



图十 话题四：个人病毒防护包含的重点词汇及频数

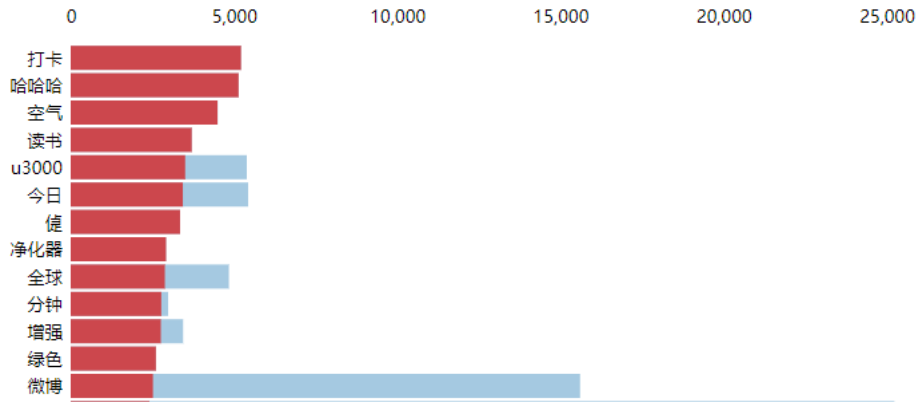
话题四：个人病毒防护。话题四中“口罩”高居首位，除此之外“在家”、“N95”、“戴”等词暗示了疫情期间公众居家隔离、自我防护的个人记录，也包括对于居家保护措施的科普和探讨。



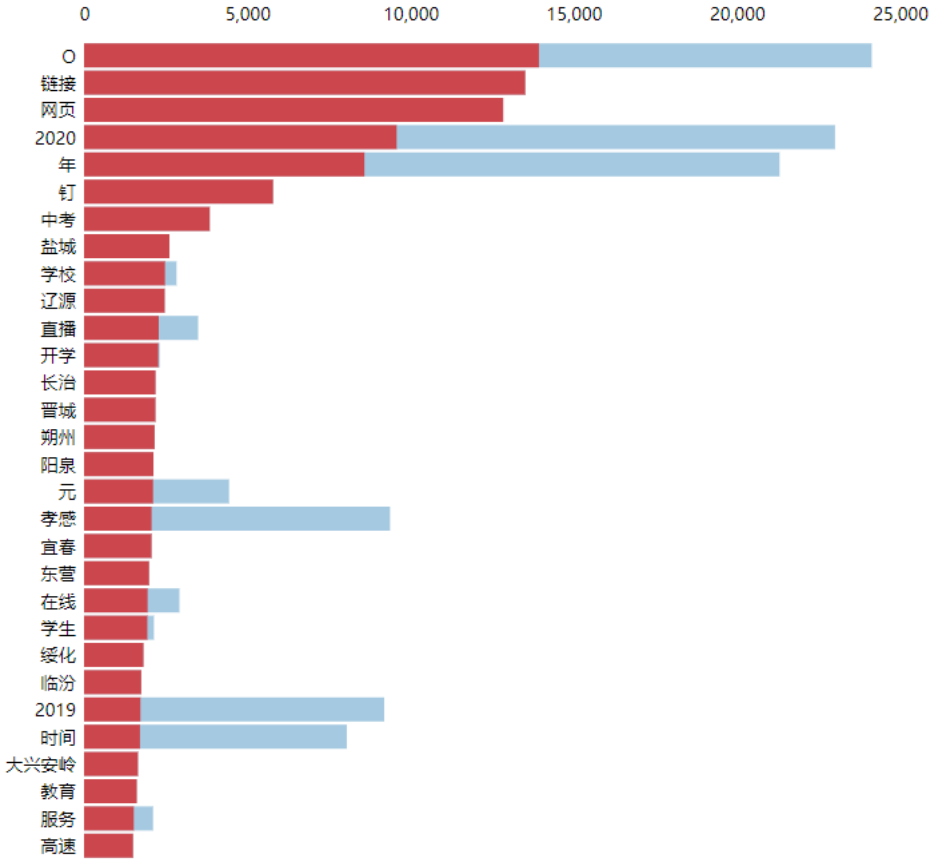
图十一 话题五：居家日常生活包含的重点词汇及频数

话题五：居家日常生活。话题五包含的词汇较为短小，且较为生活化，反映了疫情期间的居家生活，当然，“吃”字毫无悬念的高居榜首，而“钉”则暗示了在线办公、在线学习的“钉钉”平台在生活当中的广泛使用。除此之外“哈哈哈”、“读书”、“足不出户”等词汇也反映了疫情期间努力丰富生活、排解无聊、苦中作乐的生活图景。

七话题模型在上述五个话题之外挖掘出了两个规模较小、分类较细的新话题，其中话题六与话题五的本质内容相近，予以合并处理。话题六的词汇分布如下：



图十二 话题六包含的重点词汇及频数



图十三 话题七：各地在线授课政策包含的重点词汇及频数

话题七：各地在线授课政策。“钉”、“中考”、“学校”、“直播”、“开学”等词暗示了该话题与教育问题高度相关，而各地地名的出现则表明在疫情爆发初期，各地的开学时间问题是学生群体的重点关注议题。

话题提取结果表明，2020 年 1 月 1 日至 2020 年 2 月 18 日，在疫情初期，公众主要讨论了六个不同的议题，包括疫情数据发布、科学问题研讨、公益与社会情感参与、个人病毒防护、居家日常生活、各地在线授课政策等疫情相关议题，

涵盖了疫情数据、病毒科学研究、个人防护、居家生活、公益捐赠与社会参与、在线教学与办公等各个方面。在疫情初发时，社交媒体上对于科学信息的高度讨论、公益事业的广泛参与、高度自觉的居家隔离和自我防护，这些都预示着科学态度、人道主义和牺牲精神在中国抗击疫情的斗争中会发挥至关重要的作用。

四、 基于朴素贝叶斯模型的微博文本情感分析

1. 模型训练

原始数据训练集中，提供了 100k 条有情感标签的文本数据，其中‘-1’表示消极情感倾向，‘0’表示无明显情感倾向，‘1’表示积极情感倾向。按照情感标签将数据分类为积极文本和消极文本，作为正负样本。SnowNLP 库采用了朴素贝叶斯模型，通过计量正负文本中不同词汇的显示频率实现对新文本的分类。本次研究使用新冠疫情相关的正负语料对 SnowNLP 库中的语料样本进行了替换，并重新训练出针对新冠疫情数据的文本分类器，产生 0-1 的数值评估。数值越接近 1，则文本越可能表达积极情感；数值越接近 0，则文本越可能表达消极情感。

2. 结果评估

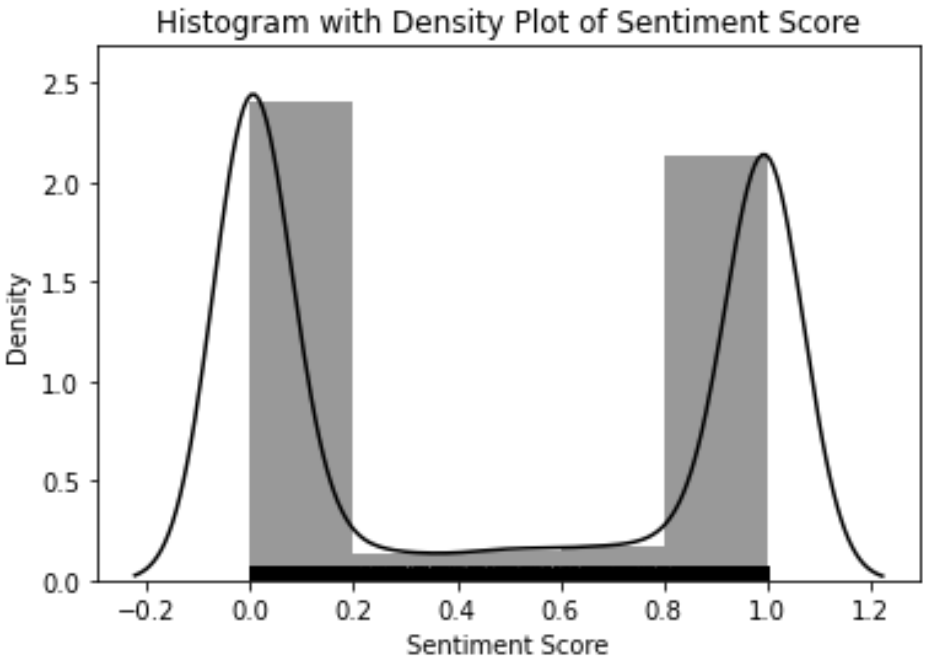
将情感分类器应用于有标签的情感数据，并和原标签进行比对。在假设原标签全部正确的前提下，在 30k 个总样本中，错误分类样本共 4374 个，平均分类误差为 0.5053。

对数据进行浏览后发现，错误分类样本主要集中在原数据中被标注为无明显情感的样本上。事实上，许多标注为无倾向的样本依然存在一定程度上的情感倾向。同时，由于原标注的取值是高度离散的，所以较大的分类误差几乎不能避免。综合以上分析，这一分类器提供的结果是可接受的。

3. 结果分析和聚类

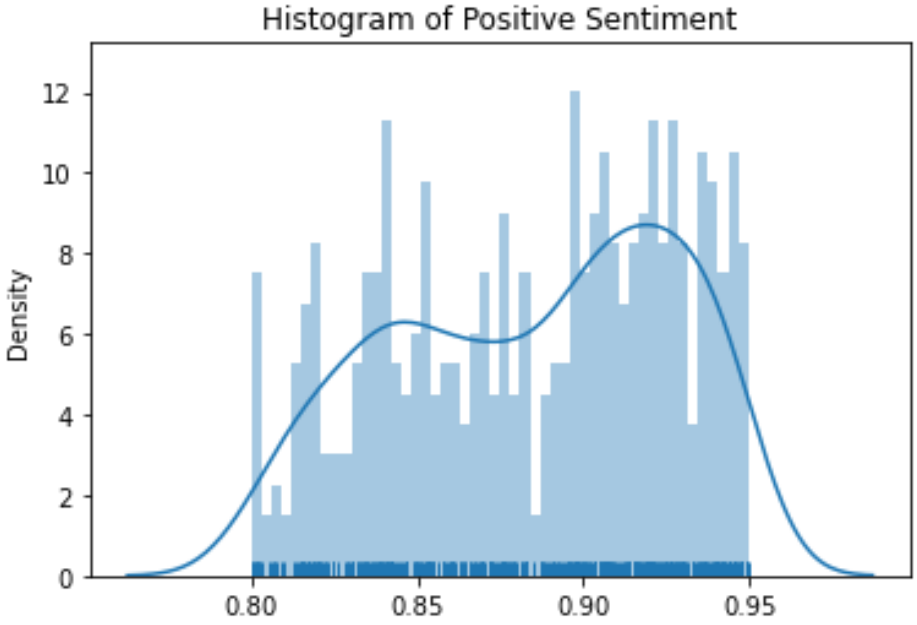
3.1 直方图和核密度估计

对上述情感分析的结果绘制直方图并求取核密度估计。



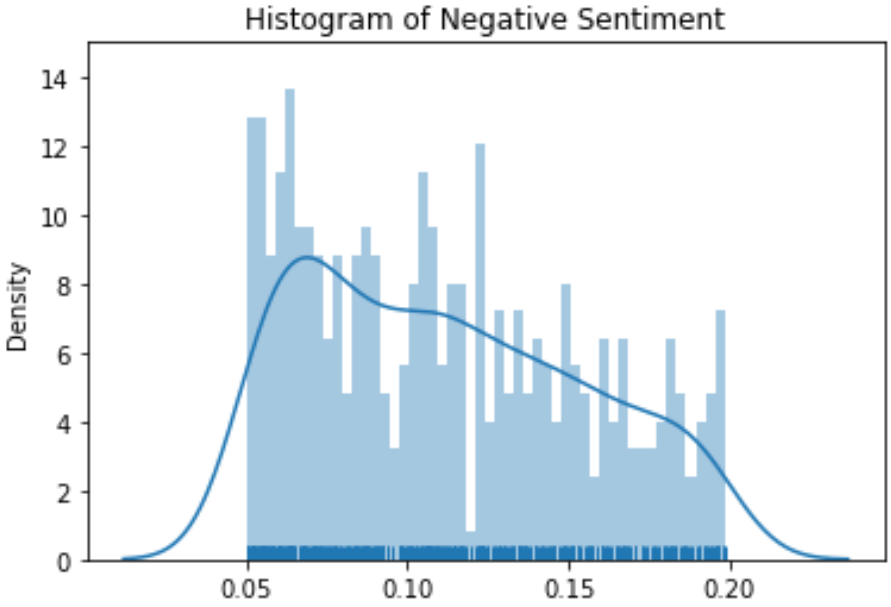
图十四 情绪得分直方图和核密度估计

对全部数据的直方图和核密度估计均表现出了显著的双峰性，大量数据集中在两种极端情绪附近，而温和情绪的分布稀疏且平坦。这反映出疫情持续期间，人们更加愿意表达情绪激烈的观点或转发有明显倾向性的事实。疫情扩散的恐惧和居家隔离的苦闷则助长了这种态势。



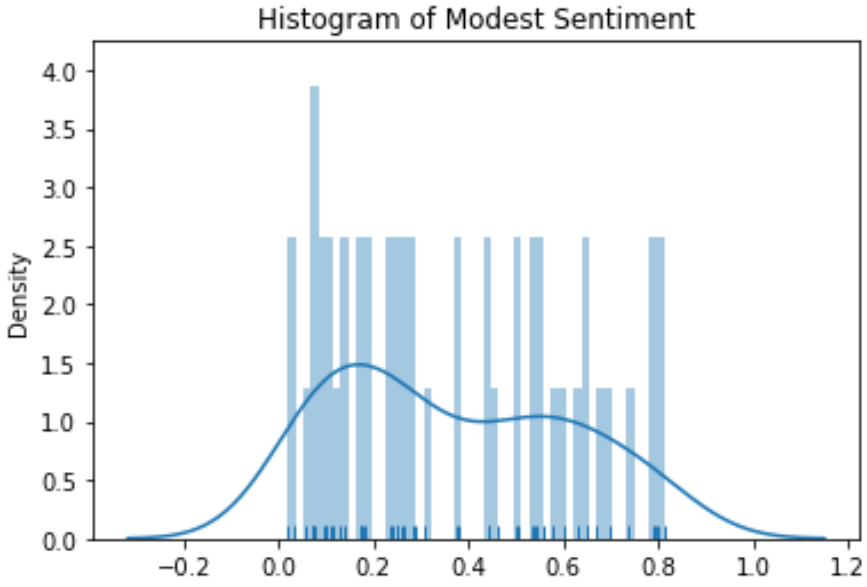
图十五 积极情绪得分直方图和核密度估计

上图显示了积极情绪文本（情绪评分介于 0.8-0.95 分）的情感分布情况，数据显示在积极文本当中，有相当一部分表现出强烈的积极情绪，但不可忽视的是在 0.83 附近仍出现一个小峰。这说明在疫情背景下，积极消息也可能受到消极情绪的制衡而部分失去积极性。



图十六 消极情绪得分直方图和核密度估计

消极情绪的表现则较为不同。除了在极端数值附近仍有聚集外，消极情绪分布相对均衡且无峰，但向温和情绪的偏移极不明显。这表明疫情爆发初期，社交网络参与者调节自身情绪的能力相对较弱，人们沉浸于恐惧和怀疑的情绪当中，更容易受到极端的消极情绪的感染。



图十七 温和情绪得分直方图和核密度估计

上图描绘温和情绪的分布情况。温和情绪的分布表现出更强的均衡性，但依然在数值较低（悲观）的部分出现峰值。这表明情绪温和的网民在疫情初期普遍抱持着谨慎悲观的态度。

综上，疫情爆发初期的网民情绪出现了严重的极化，绝大多数样本表现出强烈的极端情绪。网民的态度总体悲观，积极情绪和温和态度也会因为疫情背景的

影响，而出现向消极情绪的倾斜。

3.2 聚类与分类数量选择

如上所述，极端样本和非极端样本在数量上的悬殊使得基于总体的研究会过度放大极端样本而忽略温和样本。为了改善这一问题，本研究采用基于 k-means 的聚类方法对情绪样本进行分类。

为了寻找最佳聚类数量，引入两个评价指标对不同聚类模型的效果进行评估。

3.2.1 轮廓系数（Silhouette Coefficient）

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, a(i) < b(i) \\ 0 \quad\quad\quad, a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1, a(i) > b(i) \end{cases}$$

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

$S(i)$接近 1，则说明样本i聚类合理；

$S(i)$接近-1，则说明样本i更应该分类到另外的簇；

若$S(i)$近似为 0，则说明样本i在两个簇的边界上。

轮廓系数S为所有样本$S(i)$的均值

聚类结果的轮廓系数的取值在$[-1,1]$之间，值越大，说明同类样本相距约近，不同样本相距越远，则聚类效果越好。

3.2.2 整体平方和（Inertia）

$$\text{Cluster Sum of Square(CSS)} = \sum_{j=0}^{m}\sum_{i=1}^{n}(x_i - \mu_i)^2$$

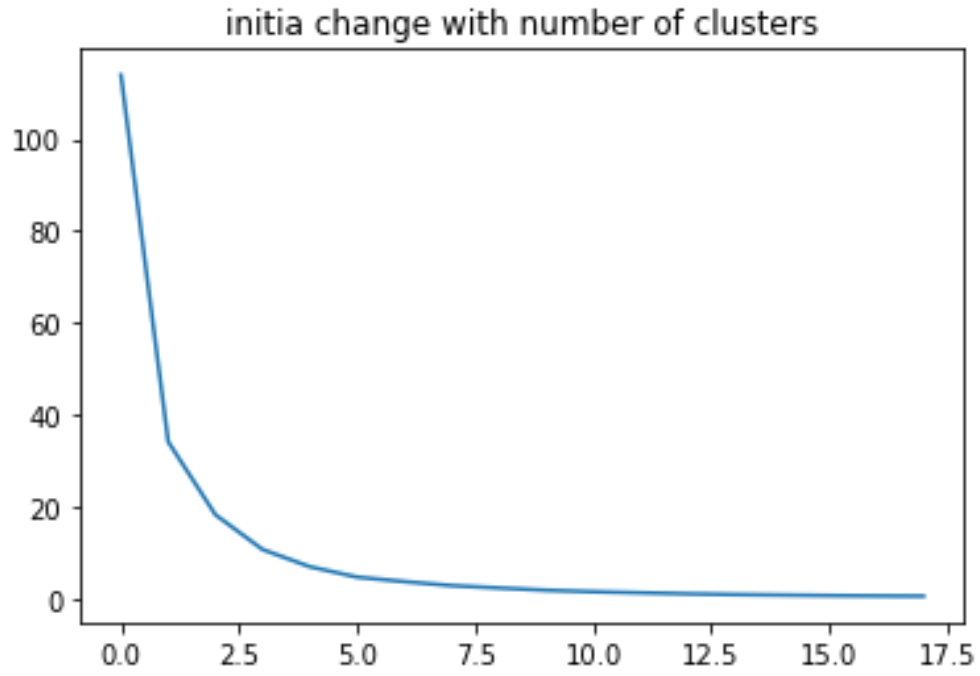
$$\text{Total Cluster Sum of Square} = \sum_{l=1}^{k} CSS_l$$

整体平方和(Total Cluster Sum of Square)为簇内平方和之和，又叫做 total

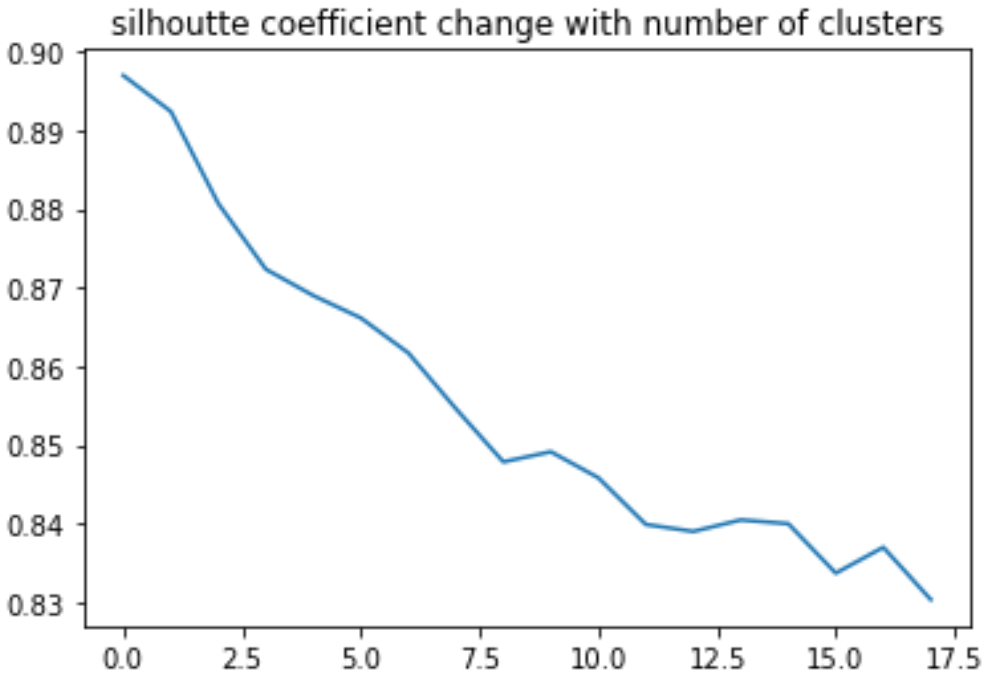
inertia，TSSE。Total Inertia 越小，代表着每个簇内样本越相似，聚类的效果就越好。

3.2.3 聚类数量选择

对于 sci-kit learn 中提供的 Kmeans 方法，采用聚类数量从 1 到 20 遍历，计算出每种聚类模型下的轮廓系数和整体平方和，并根据肘部准则选择的聚类数。



图十八 整体平方和随聚类个数的变化



图十九 轮廓系数随聚类个数的变化

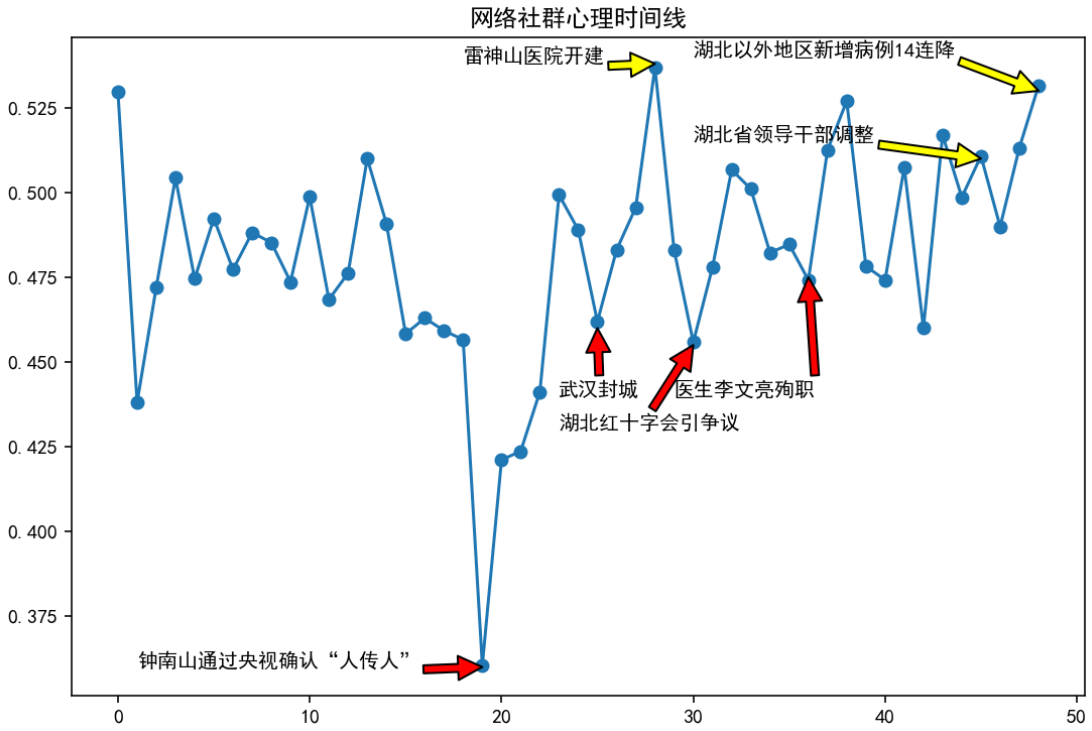
通过观察，当聚类数量大于 5 时，整体平方和不再出现显著下降，此时增加聚类数量对改善分类效果的作用有限，同时，轮廓系数在此时仍保持较高的水准。因此选择聚类个数为 5 是较为合理的。

3.2.4 聚类结论

通过有关评价标准判定的聚类个数为 5，即基于正负情绪判断，网民的疫情表达情绪应分为 5 类。根据有关社会心理学研究，人类的情绪分类方法和分类数的理论非常广泛，其中与本次聚类结果较为接近、且能够得到广泛人口的是美国研究人员 Ekman 和 Friesen 的情绪分类方法。他们认为人类有六种基本情绪即快乐、悲伤、恐惧、惊讶、愤怒、嫉妒。考虑到本次研究的背景，将嫉妒类进行删除。从而得出新冠疫情爆发初期，网民在微博内容中主要表现出快乐、悲伤、恐惧、惊讶、愤怒等五种情绪，且情绪的极性较强，负面情绪的感染力较大。

4. 基于时间序列的描述性分析

在获取每条文本的情感倾向后，本研究着力了探索公众情绪在疫情初期的波动情况，即特定疫情消息的产生是否对公众整体情绪会产生较为显著的影响。按日为单位，将文本重新聚合，求出样本中每日情绪的均值，并结合新冠疫情消息的时间线绘制折线图。



图二十 网络社群心理时间线

横坐标：时间（2020 年 1 月 1 日至 2 月 18 日）

纵坐标：当日平均情绪评分

本次研究中，共挑选了 7 个在疫情爆发初期的关键事件节点，红色箭头指示负面新闻，黄色箭头指示正面新闻。这些事件和群体心理波动的趋势产生了很好

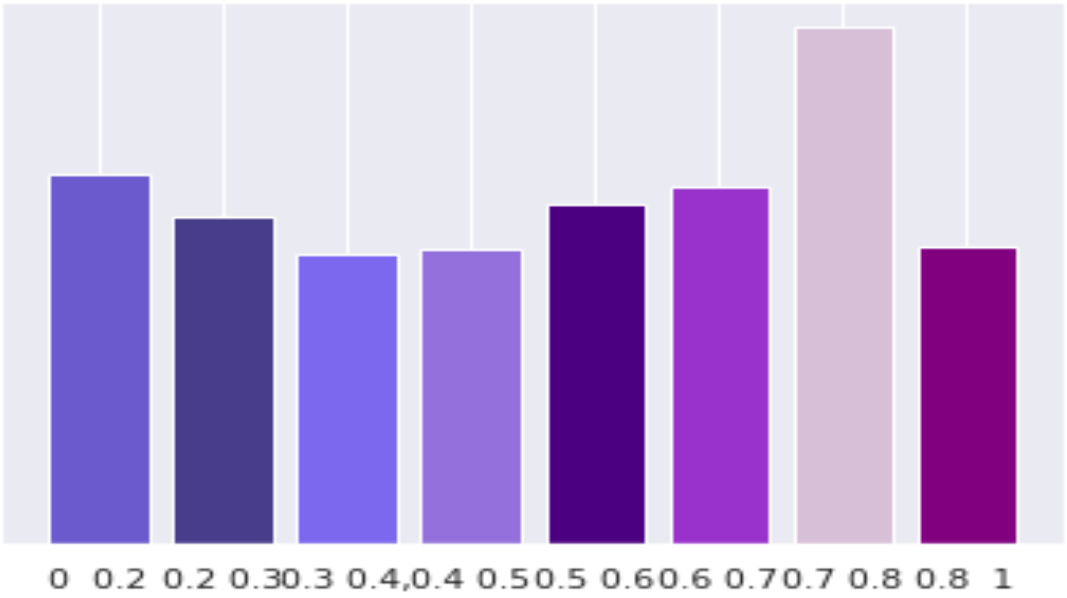
的拟合。这也从侧面验证了情感分析模型的可靠性。

从折线图中应该清晰地看到，在疫情爆发初期，钟南山院士在央视直播中对“人传人”现象的确认，引发了网民对于未知传染病的高度关注和恐慌。随后，武汉封城的消息进一步加重了网民对病毒传染性和疫情现状的担忧。湖北红十字会被指责物资发放不利、大量物资囤积的消息则造成了公众对其落后的物流分发体系和应急管理能力的愤慨，而李文亮医生的殉职则唤起了公众对于公共信息公开和言论自由的关注，也引发了关于“吹哨人”的热议。

此外，雷神山、火神山医院的开建、因应疫情防控要求对湖北省领导干部进行调整、疫情防控初现成效的消息，则改善了疫情爆发出去低迷的社群情绪，提升了公众对于疫情防控工作的信心。

5. 情感极性和用户关注的潜在联系

将情感分析模型用于疫情期间的微博热门话题的文本处理，并绘制文本情感和浏览量的柱状图。横坐标为话题情感得分，越接近 1 越可能表示积极情感，越接近 0 越可能表示消极情感。纵坐标表示该话题的浏览次数。



图二十一 不同情感的微博热搜话题浏览量统计
横坐标：情感得分
纵坐标：浏览量

直方图显示，接近 0 和 1 两个极端值的话题浏览量较高，随着情感极性的减弱（值接近 0.5），浏览量出现了一定程度的下降。这表明疫情期间，极端乐观与悲观的情绪相比温和的情绪更容易吸引网民的关注，而情感表述模糊的话题则显得较为冷清。在极端情绪中，积极的消息的吸引力更高，这也反应了人们对于疫情利好消息的期待和对于抗击疫情终将成功的信心。

五、 利用 root 词构建领域情感词典

情感分析是自然语言处理领域的重要话题，传统的情感分析方法中，包括借助情感词典对文本情感进行测量的方法。文献显示，构建情感词典是这种方法中难度最大、成本最高的部分。查阅 github 开源代码时，发现刘焕勇博士开发出一个利用少量根词和文本语料，开发出领域情感词典的方法，从而为情感分析方法提供重要的一句。本次研究中，将主题提取中获得的 44 个词汇进行手动正负向情感分类，并利用刘焕勇博士开发的 wordexpansion 库创建了专门针对新冠肺炎疫情研究的情感词典，为他人的研究提供帮助和参考。根词词汇和情感词典参看附录和另附文件 pos_candi.txt 和 neg_candi.txt。

六、 成果总结

本次研究的成果主要包括四个方面。一是对微博文本数据的清洗、切分和词频统计、词云展示。二是对微博文本使用 LDA（隐含狄利克雷分布）主题生成模型方法进行主题提取，获得了六个疫情初期的热议话题。三是利用朴素贝叶斯模型训练出针对新冠疫情相关讨论的文本情感分类器，并获得了较好的效果。利用这一情感分析模型，分析了网民的极端性表达规律。通过对微博热搜文本的情感分析，获得了极端情绪和社会话题对网民的吸引程度计量。同时通过聚类分析将网民情感分为 5 类，并结合疫情时间线，分析了社交媒体在新冠疫情初期的日间情绪变动。四是训练出一个在新冠肺炎疫情的社会研讨方面有显著效果的领域情感词典，为他人的研究提供帮助。

Reference：

[1]杜增文. 基于狄利克雷回归的微博主题检测模型研究[D].中国科学院大学(中国科学院大学人工智能学院),2020.

[2]凌鑫元. 基于在线社会网络的用户情感分析研究与实现[D].南京邮电大学,2020.

[3]杨德生,程慧,叶绮娜. 重大突发事件对群体情绪的影响测度及预警干预研究——以新冠肺炎疫情为例[A]. 中国统计教育学会.2020 年（第七届）全国大学生统计建模大赛优秀论文集[C].中国统计教育学会:中国统计教育学会,2020:54.

[4]庄穆妮,李勇,谭旭,毛太田,蓝凯城,邢立宁.基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真[J].系统仿真学报,2021,33(01):24-36.

[5]常甜甜.基于文本挖掘的网络舆情研究进展述评——使用 CiteSpace 的可视化图谱研究(2000—2020)[J].新媒体研究,2021,7(03):5-7.

[6] 如何使用 Python 快速构建领域内情感词典_大邓和他的 Python-CSDN 博客

[7【Python】LDA 模型中文文本主题提取丨可视化工具 pyLDAvis 的使用_刷题小狂魔的博客-CSDN 博客]

附录：

根词（root word）

一线 pos 保卫 pos 韩红 pos 医疗队 pos 抵抗力 pos 战疫 pos 治愈 pos 祈福pos

钟南山 pos 拐点 pos 院士 pos 丁香 pos 力量pos 捐赠pos 希望pos

医护人员 pos 平安 pos 抗疫 pos 加油 pos 感谢 pos 爱 pos 钉钉 pos 志愿者 pos

致敬 pos 愿 pos 戴 pos 阻击战 pos 支持pos 野味neg 蝙蝠neg 野味neg

果子狸 neg 疫情neg 确诊neg 肺炎neg 病毒neg 感染neg 冠状病毒 neg