系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
41(10) (2021, 10), 2906–2918

基于 PCA-Spectral-LDA 的网络舆情聚类和情感演进分析: 一个微博文本挖掘研究*

邱泽国[1,2] 贺百艳[1]

(1. 哈尔滨商业大学, 哈尔滨 150028; 2. 黑龙江省文化大数据理论应用研究中心, 哈尔滨 150028)

摘要 互联网成为网络舆情传播的主要媒介, 分析突发事件的情感发展态势, 可以探究舆情演变规律并识别潜在风险, 为舆情的引导和控制提供决策支持. 文章对爬取的微博文本数据进行预处理, 基于文本数据的高维特征, 首先利用主成分分析方法进行降维, 然后采用谱聚类算法, 并提出结合潜在狄利克雷分析模型提取文本主题的方法, 对每类主题进行情感分析. 通过数据可视化方法研究网民的情感倾向, 得到网络舆情传播中情感的时空演化规律. 研究结果能够清晰地表明网民的情感态度和舆情走向, 文章的研究方法为微博舆情的研究提供了新视角.

关键词 网络舆情, 谱聚类, LDA 主题模型, 情感分析, 演进.

MR(2000) 主题分类号 62P25, 68T50

Analysis of Internet Public Opinion Clustering and Sentiment Evolution Based on PCA-Spectral-LDA: A Research on Weibo Text Mining

QIU Zeguo[1,2] HE Baiyan[1]

(1. *Harbin University of Commerce, Harbin* 150028; 2. *Heilongjiang Province Cultural Big Data Theory Application Research Center, Harbin* 150028)

Abstract The Internet has become the main medium for the dissemination of online public opinion. Analyzing the emotional development trend of emergencies can explore the evolution of public opinion and identify potential risks, providing decision-making support for the guidance and control of public opinion. This paper preprocesses the crawled microblog text data. Based on the high-dimensional features of the text data, first uses the principal component analysis method to reduce the dimensionality, then uses the spectral clustering algorithm, and proposes to combine the

* 黑龙江省哲学社会科学基金项目 (20JYB031) 资助课题.
收稿日期: 2021-03-25, 收到修改稿日期: 2021-06-29.
编委: 李振鹏.

potential Dirichlet analysis model to extract the text topic The method, sentiment analysis for each type of topic. The data visualization method is used to study the emotional tendency of netizens, and obtain the temporal and spatial evolution law of emotion in the spread of online public opinion. The research results can clearly indicate the emotional attitudes and public opinion trends of netizens. The research method in this article provides a new perspective for the research on Weibo public opinion.

Keywords Internet public opinion, spectral clustering, latent Dirichlet allocation, sentiment analysis, evolution.

1 引 言

据中国互联网络信息中心 (CNNIC) 发布的第 45 次《中国互联网络发展状况统计报告》显示, 截至 2020 年 3 月, 我国网民规模为 9.04 亿, 相较 2018 年底新增网民 7508 万, 互联网普及率达 64.5%, 手机网民规模为 8.97 亿, 网民使用手机上网的比例达 99.3%[1], 由数据可知越来越多的人通过网络渠道来获取新闻等热点事件. 微博因其具有操作简单, 发布查阅信息快捷方便等特点, 逐渐成为主流的社交媒体和舆论场. 用户可以在微博上即时表达对某一事件的观点、看法、情感态度等, 这极大地推动了社会突发事件的舆论传播, 使得网络舆情应运而生 [2]. 网民的情绪不但能影响消息的传播速度, 还可以将自身情绪快速地传递给其他人进而导致舆论的爆发 [3]. 因此, 有必要针对微博社交媒体进行数据挖掘与分析, 了解人们对于某一事件的主观看法、态度和情感倾向, 并根据分析结果实时引导舆论发展, 避免错误和偏激的舆论传播对社会造成不良影响.

近年来, 随着社交平台的飞速发展, 对于主题挖掘和情感分析的研究越来越受到国内外学者的重视. 微博文本作为典型的文本数据之一, 具有很高的研究价值. 文本挖掘方法可以用来发现文本中的潜在信息, 文本聚类分析是文本挖掘的核心内容之一. 文本聚类分析是通过对文本内容进行归纳处理并将文本对象自动分组的过程 [4]. 在文本聚类方面, 有专注于技术的改进与实现的, 例如 Stilo 和 Velardi[5] 基于时间序列的相似性提出了一种在微博中对词语进行聚类的新方法, 并通过实验进行了验证, 结果表明该方法具有较好的可靠性; Yan 等 [6] 研究了一个聚类框架来识别新的主题, 并根据其内容和时间特征及时检测热门主题; Huang 等 [7] 提出了一种新的主题检测技术, 利用隐含狄利克雷分析模型 (latent Dirichlet allocation, 以下简称 LDA) 代替了传统的向量空间模型, 提取微博中的主题信息; 唐晓波和王洪艳 [8] 基于 LDA 主题模型对微博主题进行挖掘, 并采用依存句法分析对传统文本相似度矩阵进行改进, 以提高聚类精度; 王静茹和陈震 [9] 针对不同类型文本数据, 提出了一种基于 LDA 的主题提取效果对比评价方法, 实验结果证明 LDA 模型在处理语义逻辑关系清晰, 信息明确的长文本数据时具有较好的主题提取效果. 有专注于技术研究与应用的, 例如马莹雪和赵吉昌 [10] 采用机器学习方法对微博有效数据进行了提取, 利用深度学习方法对微博文本进行聚类分析, 并使用复杂网络分析方法对微博信息传播模式进行了研究, 研究结果可以为灾害期间的舆情发现和管理提供一定启发和支持; 张琛等 [11] 基于 Single-Pass 聚类算法对微博文本进行聚类分析, 识别疫情热点话题, 利用自然语言处理技术构建了一个面向社交媒体评论文本的舆情分析框架, 为重大公共事件的舆情研究提供了理论支撑和研究思路.

在对舆情进行分析时, 通常将情感分析与主题挖掘相结合来更全面的研究网络舆情. 例如朱晓霞等 [12] 结合 TF-IDF 和 K-means 聚类方法, 挖掘评论中主题和情感的分布与联系, 并利用情感词典和点互信息的方法计算主题词的情感值从而进行情感分类; Chen 等 [13] 提出了隐含情感模型 (latent sentiment model), 该模型将主题划分为三种带有不同情感的特殊主题, 从而实现对文本的情感分析; 王秀芳等 [14] 提出了一种基于主题聚类和情感强度计算的微博舆情分析模型, 该模型可以实现对微博话题的快速聚类和情感强度量化计算, 并通过对时间序列的回归分析来追踪预测热点话题的情感变化.

综上所述, 国内外学者对于网络舆情中情感和主题的分析方法较为丰富. 本文针对微博文本的高维稀疏特性进行研究, 以文本挖掘为基础, 通过 PCA-Spectral-LDA 的综合方法进行主题聚类和提取, 并对所有微博文本进行情感分析. 最终研究目的, 旨在提高主题发现效率, 降低文本的高维性和稀疏性的影响; 研究舆情事件在整个生命周期中话题内容变化的状态, 分析网民的情感态度及走向, 正视和积极引导网民的网络行为, 帮助相关部门及时把控舆情动态, 充分利用网络舆情的积极作用.

2 网络舆情数据分析方法

2.1 PCA-Spectral-LDA 的数据分析策略

PCA (principal component analysis) 即主成分分析, 它不仅可以对高维数据进行降维, 更重要的是能够消除数据冗余和噪声 [15]. 如果最终聚类的维数很高, 由于降维的幅度不够, 谱聚类的聚类效果和运行速度均欠佳. 因此利用 PCA 降维后, 在少数几个综合指标或维度的基础上实行聚类分析, 最终的效果会更简洁、明了.

谱聚类 (spectral-clustering) 是一个被广泛应用的聚类算法, 它对数据分布的适应性相较于其它聚类算法而言更强, 聚类效果更优秀, 因其实现简单最近几年越来越受学者们的青睐. 谱聚类具有的优点如下 [16]: 1) 谱聚类算法只需要数据间的相似度矩阵, 对于稀疏数据的聚类效果非常好; 2) 相较于传统聚类算法, 由于谱聚类算法使用了降维, 因此在处理高维数据聚类时的复杂度更低; 3) 谱聚类算法基于谱图理论, 它能够在任意形状的样本空间上进行聚类并且收敛于全局最优解.

LDA 是由 Blei 等人 [17] 于 2003 年提出, 是一种用来推测文档的主题分布的模型. LDA 主题模型是一个 “文档－主题－词” 的多层贝叶斯生成式模型 [18]. 其核心是认为一篇文档 (Document) 包含有多个主题 (Topic), 而每个词 (Word) 都由一个固定主题生成. 利用 LDA 主题模型可将文本信息转化为数字信息, 使文档集中每篇文档的主题以概率分布的形式给出, 通过抽取文档中的主题, 便可根据主题分布实现主题聚类或文本分类, 目前已有诸多学者使用该方法来提取社交媒体文本中的主题 [19, 20]. 但 LDA 主题提取在面对社交媒体语料库时存在缺点. 首先, 由于社交媒体语料库比较杂乱噪声较大, 某些主题中包含较多与事件不相干的词语, LDA 会提取得到较为嘈杂的主题, 主题间区别不明显. 其次, LDA 主题模型是通过在单文档层面获取共现词对来发现主题的, 与 LDA 擅长处理的长文章级别的文档相比, 微博文本较短, 导致单条文本中的共现信息不足, 不利于 LDA 发现共现词对并生成高质量的主题 [21]. 而利用 PCA-Spectral 可以有效克服文档矩阵的高维性和稀疏性, 保留语料库中的重要共现信息以提高主题生成的质量.

2.2 网络舆情传播中的情感分析

文本情感分析也称为意见挖掘, 是指利用自然语言处理、文本挖掘及计算机技术等方法对带有情感色彩的文本进行提取、分析和处理的过程, 从而帮助用户获取有效信息 [22]. 通过情感分析能够了解到网民对某个事件的看法, 识别出网民发布内容的情感趋势, 例如: 支持、反对、快乐或难过, 通过情感分析可以进一步预测情感随时间的演化规律. 文本情感分析主要有基于规则 (情感词典) 和机器学习两种方法. 基于词典的方法是通过制定情感词典对预处理后文本中的情感词进行匹配得到文本的情感倾向, 此方法不需要人工标注, 但是如果在不同语境下同一词汇可能会表达出不同的情感倾向. 基于机器学习的方法是将文本情感分析转化成有监督的分类问题来处理, 虽然其准确率高, 但是需要人工对原始数据进行标注, 不仅耗时费力, 而且人工标注数据的结果会直接影响到分析的准确性.

本研究基于事件语料库, 运用大连理工情感词汇本体库、ROSTCM6 分词系统做基础词库的同时, 通过观察和收集整理出包括有关本事件专业用语以及网络流行词语的自定义词库, 并将自定义词库导入 ROSTCM6 文本分析工具的分词字典中 [23]. 通过对微博文本进行情感分析, 获得微博网民的情感倾向性分布, 为后续的舆情管理提供有效对策.

3 研究设计

本研究的主要流程可分为 3 部分: 数据预处理、主题挖掘与情感分析及舆情演化阶段划分, 具体技术路线如图 1 所示.

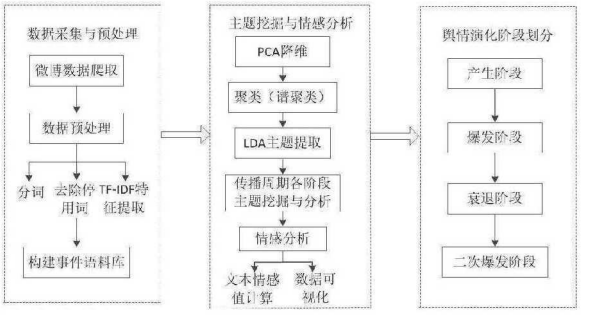

图 1 技术路线图

(Figure 1 Technology roadmap)

3.1 数据收集及预处理

2020 年 5 月 22 日, 艺人 TZ 在直播中自曝曾为考心仪大学将往届生身份修改成应届生, 该言论一出立刻引发热议, 网民就此事在微博上展开激烈讨论. 本文以 “TZ 往届生改应届生” 为搜索关键字, 从新浪微博中按照相关性爬取了 2020 年 5 月 30 日至 2020 年 6 月 23 日有关 “TZ 往届生改应届生” 的微博正文, 最终得到 3660 条微博文本数据.

1) 去除重复数据

有些营销号为了赚取流量会随意复制他人评论, 甚至带相关话题发布一些与本事件无关的内容, 这些大量重复的文本会降低主题提取的精确率, 因此要将这些数据剔除. 本文采用比较删除法去重, 若两条文本比较后发现完全相同, 就予以删除. 而针对无关的内容文本,

则采用人工筛选的方法. 在人工清洗后最终去除重复文本 326 条, 无意义文本 210 条, 得到有效数据 3124 条.

2) 分词及去除停用词

文本分词是将文档中的每一子句按照一定的规则拆分为单个词语的过程, 是语义理解的关键步骤. 本文使用 python 中的 Jieba 分词工具包进行中文分词, 得到分词后的数据集. 停用词是指去除如语气词和特殊符号等对后续工作无实际意义且主题不明显的字或词. 文本预处理流程见图 2.

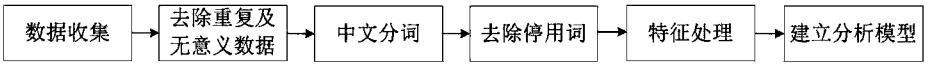

图 2 数据预处理流程

(Figure 2 Data preprocessing process)

为了更清晰地了解 “TZ 往届生改应届生” 一事网民所讨论的主要话题, 本文对预处理后的所有微博文本进行词频分析, 排名 Top20 的高频词如表 1 所示. 将高频词生成词云图以更直观地展现事件发生后网民讨论的热点话题, 如图 3 所示.

表 1 Top20 高频词表

(Table 1 Top20 high-frequency vocabulary)

词语	频次	词语	频次
坑爹	1243	热搜	671
应届	1125	学籍	663
高考	1134	犯罪	645
毕业证	1088	违反	611
往届	1012	任职	586
自曝	886	舞弊	509
反腐	862	学籍	467
撤职	750	弄虚作假	404
学历	38	撤销	352
临汾市	715	作废	330

图 3 文本词云图

(Figure 3 Text word cloud)

由词云图可以看出, 在微博文本词频前 600 的词语中, 出现频次越高, 字体就越大. 词频达到 1000 以上的按照次数多少依次为 “坑爹” “应届” “高考” “毕业证” “往届”, 具体次数为 1243 次、1125 次、1134 次、1088 次和 1012 次. 这些出现频率比较高的词语在一定程度上构建了 “TZ 往届生改应届生” 的焦点事件.

3) 特征提取

在分词及去除停用词之后, 需要对文本进行特征提取, 因为每个词语对实体的贡献度不同, 所以需要对这些词语赋予不同的权重. 本研究采用 TF-IDF 方法计算词项在向量中的权重, 利用 scikit-learn 工具调用 CountVectorizer() 和 TfidfTransformer() 函数计算 TF-IDF 值. 由于文档的维数过多, 可以设置固定的维度, 同时可以利用 PCA 进行降维操作, 通过对参数进行不断调试取最优, 最终选取 n-components=3.

3.2 网络舆情数据挖掘分析

3.2.1 Spectral-LDA 分析

本文利用聚类评价指标 (轮廓系数 $S(i)$) 来确定聚类主题数量. 由谱聚类的特点可知, 如果聚类主题数 K 过小, 则会导致数据本身分不开, 影响聚类效果. 但随着主题数 K 增大, $S(i)$ 值会越来越小, 聚类效果会变得越来越差. 为了验证聚类结果的有效性, 选取 K-means、层次聚类以及本文使用的谱聚类三种算法进行实验. 因为三种方法均为无监督学习算法, 所以统一选用轮廓系数作为最终的效果评价指标, 对比结果如图 4 所示.

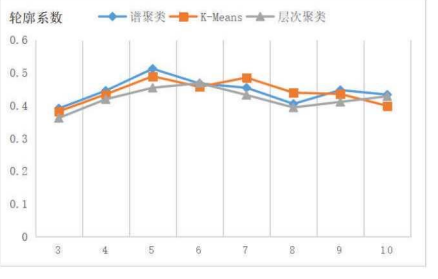

图 4 聚类算法结果对比

(Figure 4 Comparison of clustering algorithm results)

由上图可以看出, 谱聚类算法取得了最优的轮廓系数值, 此时 $S(i) = 0.512$, 主题数 $K = 5$, 其聚类效果如图 5 所示.

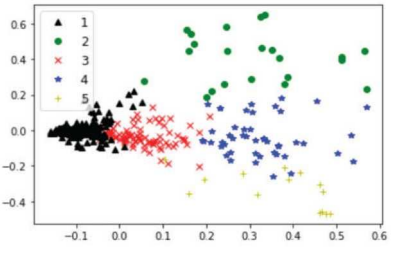

图 5 谱聚类效果图

(Figure 5 Spectral clustering effect)

为了检验 Spectral-LDA 策略的可靠性, 选用 PLSA 主题模型与 LDA 主题模型做对比. PLSA 是广泛应用于文本挖掘、话题识别领域的一种主题提取模型. 通过 F-score 和聚类结果纯度 (Purity) 两个指标来衡量基于 LDA 模型和 PLSA 模型的文本聚类效果. F-score 是一个用来评价聚类效果的综合指标, 它是精确率 (Precision) 与召回率 (Recall) 调和平均值. 给定聚类 i 和类别 j, F-score 的计算公式如下

$$F\text{-}score = \frac{(1+\beta^2)Precision * Recall}{\beta^2 Precision + Recall}, \tag{3.1}$$

其中 β 为 F-score 的调和系数, 一般取 $\beta = 1$. 在聚类有效性评价中, 整体的 F-score 值可通过类间 F-score 值的加权平均求得, 计算公式如下

$$F\text{-}score = \sum_i \frac{n_i}{n} max_j \left(\frac{2Precision * Recall}{Precision + Recall} \right). \tag{3.2}$$

聚类结果的纯度 (Purity) 是一种简单直观的聚类评价指标, 其计算公式如下

$$Purity = \sum_{i=1}^{k} \frac{m_i}{m} p_i, \tag{3.3}$$

其中 $p_i = \max(p_{ij})$, $p_{ij} = \frac{m_{ij}}{m_i}$, m_i 是在聚类 i 中所有成员的个数, m_{ij} 是既属于聚类 i 又属于聚类 j 的成员个数. 利用谱聚类算法选取主题个数为 5 后, 比较两个主题提取模型的文本聚类效果, 结果如图 6 所示.

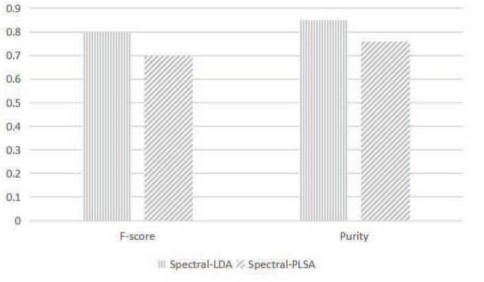

图 6 聚类效果比较

(Figure 6 Comparison of clustering effect)

利用 LDA 主题模型对微博文本进行主题提取, 得到各主题的关键词, 提取每个主题中权重前 8 的关键词, 结果如表 2.

表 2 5 个主题前 8 个高频词

(Table 2 Top 8 high-frequency words in 5 topics)

主题编号	主题特征词							
Topic1	坑爹	高考	事件	应届生	往届生	直播	继父	改为
Topic2	招生	普通高校	统一	调查	管理中心	回应	全国	官微
Topic3	撤销	综合	办公室	中央戏剧学院	舞弊	毕业证	介入	调查
Topic4	临汾市	处分	山西省	仝天峰	相关人员	撤职	父亲	涉事
Topic5	手段	不正当	违背	惩罚	严厉	公平	利用	往届生

从表 2 可以总结网民关于 5 大主题的讨论, 如表 3.

表 3　主题内容提取

(Table 3　Topic content extraction)

内容		微博文本示例	数据量/条
Topic1	“TZ 高考往届生改应届生” 一事坑爹, 对其父亲造成了很大的影响.	厉害了, 实力坑爹, 反腐在路上, 你立功了.	1653
Topic2	在事件发生后招生管理中心立即介入调查, 并做出相关回应.	山西省招生考试管理中心已作出 TZ 参加山西省 2013 年普通高校招生全国统一考试各阶段、各科成绩无效的处理决定.	581
Topic3	中央戏剧学院针对此事件, 撤销了 TZ 的毕业证书.	中央戏剧学院发布通报, 撤销 TZ 的毕业证书.	626
Topic4	TZ 父亲仝天峰被撤职, 以及相关涉事人员被处分.	从自爆往届生改应届生, 到被挖不到年龄入党, 到继父被撤职相关人员相应被调查处分. 一句话断送了多少人原本平静的生活和大好前程. 一声唏嘘可惜啊 … 也是咎由自取吧 …	110
Topic5	TZ 利用不正当手段修改往届生身份, 违背了高考公平原则, 应该受到重罚.	这种走关系利用不正当手段的上学的方式真的对我们这些平民学子也太不公平了, 都说寒门出贵子, 那也得是在公平竞争的良性环境下啊, 必须得到法律的惩罚.	154

为了观察该事件话题在时间上的演变, 本文将事件舆情随时间发生的变化按照话题热度划分成 2020 年 5 月 30 日 – 2020 年 6 月 6 日、2020 年 6 月 7 日 – 2020 年 6 月 15 日和 2020 年 6 月 1 6 日 – 2020 年 6 月 23 日 3 个阶段, 舆情话题热度变化趋势如图 7 所示. 整体而言, 第一阶段为舆情产生阶段, 网民主要围绕主题 1、2 展开讨论; 第二阶段为舆情爆发阶段, 网民参与程度高, 话题的讨论数量随之增多, 网民主要针对事件的处理结果发表看法, 主题 1、3 和 4 的讨论热度升至顶峰; 第三阶段为舆情的衰退阶段, 此时事件随着时间的推移, 发博数量也逐渐降低, 围绕主题 1 和 5 的讨论较多. 根据每个话题的发博数量可以分析出, 网民讨论热点主要集中在主题 1、2 和 3 上. 由此而言, 该事件舆情讨论的重点是当事人及其相关涉事人员的处理结果.

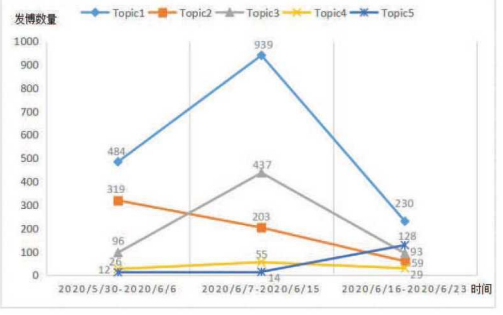

图 7　聚舆情话题热度变化趋势

(Figure 7　Trends in the popularity of public opinion topics)

3.2.2 情感分析

ROSTCM6 软件是由武汉大学开发的一款开源的文本挖掘工具, 可对文本内容进行分词、词频统计、情感分析以及聚类分析等. 该工具的情感分析原理是对文本中每一个子句的情感词计算权重, 将所有权重相加得到整个文本的情感得分. 它可将文本情感分为积极情绪、中性情绪和消极情绪三类, 各类情感值区间分别为 $(5,+\infty)$、$[-5,5]$ 和 $(-\infty,-5)$.

将聚类后的数据作为输入, 选择 ROSTCM6 软件中的情感分析功能模块, 计算每条微博文本的情感值, 最终得到积极情绪文本 1764 条, 占 56.47%; 中性情绪文本 509 条, 占 16.29%; 消极情绪文本 851 条, 占 27.24%, 网民的情绪分布对比如图 8 所示.

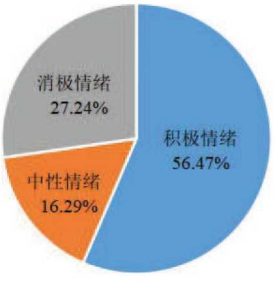

图 8 网民情绪分布对比

(Figure 8 Comparison of netizens' emotion distribution)

分别对每一主题的情感极性占比进行统计分析, 情绪类别结果见表 4 所示.

表 4 各主题情绪类别结果

(Table 4 Results of emotion categories of each topic)

Topic1 1	类型	数量/条	占比/%	Topic2	类型	数量/条	占比/%
	积极情绪	813	49.18%		积极情绪	428	73.67%
	中性情绪	364	22.02%		中性情绪	57	9.81%
	消极情绪	476	28.80%		消极情绪	96	16.52%
Topic3	类型	数量/条	占比/%	Topic4	类型	数量/条	占比/%
	积极情绪	402	64.22%		积极情绪	96	87.27%
	中性情绪	65	10.38%		中性情绪	10	9.09%
	消极情绪	159	25.40%		消极情绪	4	3.64%
Topic5	类型	数量/条	占比/%				
	积极情绪	25	16.23%				
	中性情绪	13	8.44%				
	消极情绪	116	75.33%				

由表 4 可以发现, 网民发表的博文虽然有积极、消极情绪之分, 但是总的情感倾向是积极的. 每个主题所讨论的内容不同情感倾向也有很大的差别. 前 4 个主题的积极情绪占比都大于消极情绪占比, 说明即使发生了这样的负面新闻, 只要当事人及时承认错误, 相关部门反应迅速, 处理结果合理, 网民的情绪是很容易被积极引导的. 只有主题 5 消极情绪多于积极情绪, 根据主题所讨论内容可以看出网民极其痛恨高考舞弊行为, 因为高考是万千学子改

变命运的可能, 网民希望消灭社会不正之风, 引领教育的正面竞争. 这也是网民消极情绪多于积极情绪的原因所在.

为了便于可视化, 本文利用 Python 中的中文自然语言处理工具包 SnowNLP 的情感分析模块计算 3124 篇博文的情感得分, 该模块可以将文本分为积极和消极两类, 即预测输入的文本属于积极和消极的概率. 对文本中出现的情感词汇进行加权分析, 返回的情感分析结果接近 1 为积极, 接近 0 为消极. 以日为单位, 对所有微博文本进行情感分析, 取平均值为该日的情感得分, 图 9 是以时间顺序, 显示了对该事件的情感分析结果. 从整体上看, 积极情绪 (情感得分大于 0.5) 的数量占多数, 消极情绪 (情感得分小于 0.5) 的数量较少, 大部分博文的情感极性较强, 使得最终的分类结果较为直观.

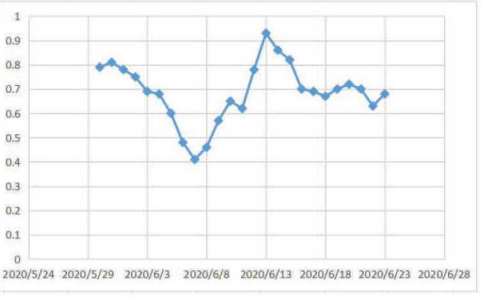

图 9 情感极性变化图

(Figure 9 Change of emotional polarity)

4 舆情演化阶段分析

研究舆情演化规律必然涉及到事件发展过程中的动态变化, 因此研究发表博文的时间序列是分析工作中的一个重要环节. 根据网络舆情的周期演化理论, 以日为单位对网民发表的博文进行分析, 得到发博数量的时间序列图, 如图 10 所示. 由时间序列图可以发现, 尽管该事件在网络上持续了一个相对较长的时间, 但是针对当事人的发博数量并非是爆发性增长的. 同时由时间序列图可以看出, 网民的发博数量出现两次峰值. 与常见的舆情传播模式不同, 本研究按照每日发博数量, 将该事件舆情发展划分为 4 个阶段: 1) 产生阶段 (2020 年 5 月 30 日 –2020 年 6 月 6 日), 这期间事件的热度并不是很高, 参与讨论的网民不多. 当事人发布手写道歉信, 部分网民表示接受当事人的道歉, 但称其要得到应有的惩罚, 希望能好好改正. 结合图 9 情感极性变化图来看, 网民的积极情绪占主导. 2) 爆发阶段 (2020 年 6 月 6 日 –2020 年 6 月 8 日), 官方政务媒体 @ 央视新闻发布微博 “TZ 涉嫌高考舞弊调查进展” 称: 工作人员一问三不知回应 TZ 事件. 网民就工作人员的不作为, 调查就是走走形式展开激烈讨论, 纷纷发表微博表示强烈不满, 对相关部门的办事效率提出质疑, 博文数量迅速飙升, 舆情传播速度也急剧上升. 结合图 9 情感极性变化图可以看出网民的积极情绪迅速减少, 消极情绪占据主导位置. 3) 衰退阶段 (2020 年 6 月 9 日 –2020 年 6 月 12 日), 网民对该事件的关注度逐渐降低, 发表博文的数量也逐渐减少. 4) 二次爆发阶段 (2020 年 6 月 12 日 –2020 年 6 月 15 日), @ 央视新闻再次发布微博公布 TZ 事件处理结果: TZ 高考各科成绩无效, 中戏撤销其毕业证, 6 人因为 TZ 办理虚假转学手续被处理, 网民发表博文数量再次增多. 并表示对

处理结果很满意, 大快人心, 通过图 9 情感极性变化图可以看到网民积极情绪占比再次回升, 到达顶峰.

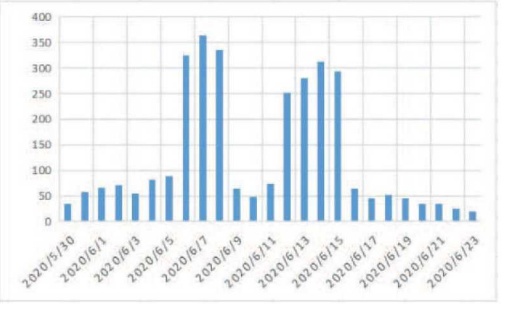

图 10 发表博文数量时间序列变化

(Figure 10 Time series changes in the number of published posts)

从事件整个发展过程来看, 舆情的起始阶段网民情绪波动较大, 如图 11 所示. 主要因为相关部门还未做出回应, 网民彼此间互相讨论, 情绪易被周围人影响. 但随着时间的推移, 官方微博公布处理结果后, 网民的情绪逐渐趋于平缓稳定, 情绪多由消极转向积极. 由此得出, 每次官方政务媒体发布微博做出解释后会不同程度地激发网民参与讨论的热度, 官方回应行为在舆情传播中扮演着重要的角色, 是刺激舆情发展的重要因素, 官方政务媒体不同的回应内容可能会导致不同的舆论走向. 因此, 相关管理部门应重点对突发事件舆情进行监督和干预, 快速并有针对性地发布权威解释, 避免错误和偏激的舆论传播对社会造成不良的影响.

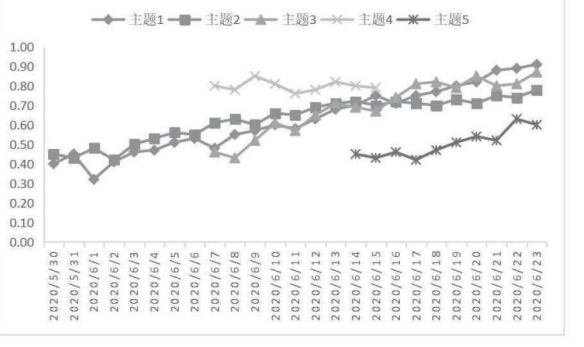

图 11 各主题情感演化趋势图

(Figure 11 Emotional evolution trend of each topic)

5 结论与展望

本文基于网络爬虫技术和网络文本挖掘方法, 爬取微博文本信息, 在利用 python 对数据进行清洗和预处理操作基础上, 提出了文本分析的 PCA-Spectral-LDA 策略, 进行主题聚类与提取, 并利用 ROSTCM6 软件对 “TZ 往届生改应届生” 事件各主题的情感进行探究. 通过分析每个主题的情感结果发现: 大部分网民对于该事件的处理结果持积极态度, 主要原因是主流媒体对事件及时采取了正面回应, 对相关责任人、贪腐官员追责问题给予了处理结

果, 这也是网民的情感态度多为积极情绪的原因. 通过对舆情演化各阶段的分析, 总结了影响微博网民情绪变化的主要原因有: 当事人回应的内容、相关部门及时有效采取措施、官方政务媒体的公信力对网民的情感极性有直接影响. 这有助于相关部门在舆情事件中获得更深层的信息和解决方法, 并为舆情的管理与决策提供可参考的建议.

如何在海量的微博文本中快速准确地提取出网民讨论的热点话题, 使不明所以的网民在最短时间内全面了解突发事件的前因后果具有重要的研究和现实意义. 网民了解事件经过后, 会引导舆情正向发展, 为相关部门更好地管控舆情奠定了坚实的基础. 本研究将 LDA 与 PCA-Spectral 相结合的方法用于舆情主题文本挖掘中. 首先利用 PCA 对文本矩阵进行降维, 采用谱聚类算法降低文本矩阵的稀疏性并确定主题数量, 再使用 LDA 主题模型提取各个聚类的主题, 该策略有效地减少了无效高频词对主题分析的影响, 提高了主题提取精度. 同时又进行了情感分析, 其结果能够清晰地反映各主题情感强度在整个舆情生命周期中的变化, 为研究文本挖掘和聚类提供了一个新思路.

本文还存在一些局限之处. 第一, 由于微博信息噪音过大, 搜索条件适当从严, 所以数据量在筛选之后变少, 后续需要收集更多可靠的中性评论和负向评论做进一步分析. 第二, PCA 对于本研究的数据是比较适用的, 但是是否能够用于其他的文本分析还有待验证, 可以在处理高维数据上进一步寻求非线性的方法, 优化特征选择的过程.

参 考 文 献

[1] 中国国家互联网信息办公室. 第 45 次中国互联网发展状况统计报告 [R/OL].(2020-04-28)[2020-05-17].www.cas.gov.cn/2020-04/27/cf589535470378587:pdf.
(China National Internet Information Office. The 45th Statistical Report on China's Internet Development[R/OL].(2020-04-28)[2020-05-17].www.cas.gov.cn/2020-04/27/cf589535470378587:pdf.).

[2] 左蒙, 李昌祖. 网络舆情研究综述: 从理论研究到实践应用. 情报杂志, 2017, **36**(10): 71–78, 140.
(Zuo M, Li C Z. Review of Internet public opinion research: From theoretical research to practical application. *Journal of Intelligence*, 2017, **36**(10): 71–78, 140.)

[3] 张鹏, 兰月新, 李昊青, 等. 基于 HAYASHI 数量化理论的网络谣言分类应对策略分析. 情报杂志, 2016, **35**(1): 110–115.
(Zhang P, Lan Y X, Li H Q, et al. Analysis on the classification and coping strategies of Internet rumors based on Hayashi's quantitative theory. *Journal of Intelligence*, 2016, **35**(1): 110–115.)

[4] 王刚, 邱玉辉. 基于本体及相似度的文本聚类研究. 计算机应用研究, 2010, **27**(7): 2494–2497.
(Wang G, Qiu Y H. Research on text clustering based on ontology and similarity. *Application Research of Computers*, 2010, **27**(7): 2494–2497.)

[5] Stilo G, Velardi P. Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery*, 2016, **30**(2): 372–402.

[6] Chen Y, Amiri H, Li Z, et al. Emerging topic detection for organizations from microblogs. International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2013, 43–52.

[7] Huang B, Yang Y, Mahmood A, et al. Microblog topic detection based on LDA model and single-pass clustering. Proceedings of the International Conference on Rough Sets and Current Trends in Computing, 2012, **7413**: 166–171.

[8] 唐晓波, 王洪艳. 基于潜在语义分析的微博主题挖掘模型研究. 图书情报工作, 2012, **56**(24): 114–119.
(Tang X B, Wang H Y. Research on the topic mining model of microblog based on potential semantic analysis. *Library and Information Service*, 2012, **56**(24): 114–119.)

[9] 王静茹, 陈震. 基于隐含狄利克雷分布的文本主题提取对比研究. 情报科学, 2018, **36**(1): 102–107.
(Wang J R, Chen Z. A comparative study on text topic extraction based on implied Dirichlet distribution. *Information Science*, 2018, **36**(1): 102–107.)

[10] 马莹雪, 赵吉昌. 自然灾害期间微博平台的舆情特征及演变 —— 以台风和暴雨数据为例. 数据分析与知识发现, 2021, **5**(6): 66–79.
(Ma Y X, Zhao J C. Characteristics and evolution of public opinion on microblog platform during natural disasters: A case study of typhoon and rainstorm data. *Data Analysis and Knowledge Discovery*, 2021, **5**(6): 66–79.)

[11] 张琛, 马祥元, 周扬, 等. 基于用户情感变化的新冠疫情舆情演变分析. 地球信息科学学报, 2021, **23**(2): 341–350.
(Zhang C, Ma X Y, Zhou Y, et al. Analysis of the public opinion evolution of the new crown epidemic based on user emotion changes. *Geo-Information Science*, 2021, **23**(2): 341–350.)

[12] 朱晓霞, 宋嘉欣, 孟建芳. 基于主题 - 情感挖掘模型的微博评论情感分类研究. 情报理论与实践, 2019, **42**(5): 159–164.
(Zhu X X, Song J X, Meng J F. Sentiment classification of microblog comments based on topic emotion mining model. *Information Studies: Theory & Application*, 2019, **42**(5): 159–164.)

[13] Chen Z P, Shen S, Hu Z N, et al. Emojipowered representation learning for cross-lingual sentiment classification. The World Wide Web Conference. San Francisco, CA, USA: ACM, 2019, 251–262.

[14] 王秀芳, 盛姝, 路燕. 一种基于话题聚类及情感强度的微博舆情分析模型. 数据分析与知识发现, 2018, **2**(6): 37–47.
(Wang X F, Sheng S, Lu Y. A microblog public opinion analysis model based on topic clustering and sentiment intensity. *Data Analysis and Knowledge Discovery*, 2018, **2**(6): 37–47.)

[15] 肖李明, 周玲, 张小龙, 等. 基于 PCA 谱聚类分析的无功分区方法研究. 陕西电力, 2016, **44**(12): 23–28.
(Xiao L M, Zhou L, Zhang X L, et al. Research on reactive power partition method based on PCA spectral clustering analysis. *Smart Power*, 2016, **44**(12): 23–28.)

[16] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001, **42**(1–2): 177–196.

[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2013, **3**(7): 993.

[18] Zhang Y, Eick C F. Tracking events in Twitter by combining an LDA-based approach and a density-contour clustering approach. *International Journal of Semantic Computing*, 2019, **13**(1): 87–110.

[19] Wang H C, Jhou H T, Tsai Y S. Adapting topic map and social influence to the personalized hybrid recommender system. *Information Sciences*, 2018, **575**: 762–778.

[20] 彭云, 万常选, 江腾蛟, 等. 基于语义约束 LDA 的商品特征和情感词提取. 软件学报, 2017, **28**(3): 676–693.
(Peng Y, Wan C X, Jiang T J, et al. Product feature and emotional word extraction based on semantic constraint LDA. *Journal of Software*, 2017, **28**(3): 676–693.)

[21] 高慧颖, 刘嘉唯, 杨淑昕. 基于改进 LDA 的在线医疗评论主题挖掘. 北京理工大学学报, 2019, **39**(4): 427–434.
(Gao H Y, Liu J W, Yang S X. Online medical review topic mining based on improved LDA. *Transaction of Beijing Institute of Technology*, 2019, **39**(4): 427–434.)

[22] 陈苹, 冯林. 情感分析中的方面提取综述. 计算机应用, 2018, **38**(S2): 84–88, 96.
(Chen P, Feng L. A review of aspect extraction in sentiment analysis. *Journal of Computer Applications*, 2018, **38**(S2): 84–88, 96.)

[23] 杨帆. 网络舆论事件中微博评论情感倾向及程度研究 —— 以 “于欢案” 为例. 传媒观察, 2018, (11): 60–66.
(Yang F. Research on the emotional tendency and degree of microblog comments in Internet public opinion events — Taking “Yu Huan case” as an example. *Media Observer*, 2018, (11): 60–66.)