【网络经济】

基于文本挖掘的网络舆情主题发现与情感分析

邱泽国　贺百艳

（哈尔滨商业大学，黑龙江哈尔滨 150028）

[摘　要] 随着网络的飞速发展，社交网络成为网络舆情传播的主要平台，用户可以随时随地通过文本、图片、视频等方式表达对热点事件的看法。分析用户对突发事件的情感态度，可以发现舆情演变规律和潜在风险，为舆情的引导和控制提供决策支持。利用 python 对爬取的文本数据进行数据预处理，从多语言多源数据的角度出发，挖掘网络舆情中高价值的舆情主题，并利用数据可视化方法来研究网民的情感倾向，分析舆情的时空演化规律。研究结果能够清晰的表明网民的情感态度和舆情每个阶段的特征。

[关键词] 网络舆情；多源数据；情感分析；演化

[中图分类号] G250　　[文献标识码] A　　[文章编号] 2095-3283(2021)02-0076-04

Topic Discovery and Sentiment Analysis of Internet Public Opinion Based on Multi Source Data Mining

Qiu Zeguo　He Baiyan

(Harbin University of Commerce, Harbin 150028)

Abstract: With the rapid development of the network, social network has become the main platform for the dissemination of network public opinion. Users can express their views on hot events through text, pictures, videos and other means anytime and anywhere. By analyzing users' emotional attitude towards emergencies, we can find the evolution rules and potential risks of public opinion, and provide decision support for the guidance and control of public opinion. This paper uses Python to preprocess the crawled text data. From the perspective of multi language and multi-source data, it mines high-value public opinion topics in network public opinion, and uses data visualization method to study the emotional tendency of Internet users and analyze the temporal and spatial evolution of public opinion. The results can clearly show the characteristics of netizens' emotional attitude and public opinion in each stage.

Key Words: Internet Public Opinion; Multi Source Data; Sentiment Analysis; Evolution

一、引言

随着近几年互联网和信息技术的飞速发展，微博微信等社交平台已经成为人们获取新闻信息的重要来源。据中国互联网络信息中心 (CNNIC) 发布第 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，较 2019 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]，越来越多的人通过网络获取新闻等热点事件。如新浪微博、微信等已经成为人们社交生活中不可或缺的一部分。在社交平台上，用户可以通过点赞、评论、转发等形式参与到发生的热点事件中，不受约束的与众多用户互动沟通。由于社交平台具有开放性、便捷性和匿名性等特点，导致新闻信息在社交网络中的传播广度、传播深度和传播速度都有着惊人的潜力，舆论会在短时间内发酵达到最后形成网络舆情，引起社会大众的广泛关注。因此，十分有必要动态跟踪网民对舆情事件话题讨论内容以及情感的变化，了解网民对于舆情事件的主观看法和情感倾向性，对于整体把握舆情事件的发展方向，引导和控制舆情有重要的意义。

二、研究现状

关于微博话题发现，学者们的研究主通过计算机领域，改进经典聚类算法来提高主题发现的有效性。Chen 等人设计开发了一个增量聚类框架来检测识别新的主题，并利用文本的内容和时间特征来及时发现热门主题 [2]；Stilo 等人基于时间序列的相似性，提出了一种在微博中用于词聚类的新方法 [3]；Hu 等人从用户评论中挖掘用户的观点看法 [4]；李亚星等人改进了 Single-Pass 算法，

[作者简介] 邱泽国（1981 － ），男，黑龙江人，博士，哈尔滨商业大学副教授，研究方向：管理决策与信息系统，贺百艳（1997 － ），女，天津人，哈尔滨商业大学在读研究生，研究方向：网络舆情管理

[基金项目] 2020 黑龙江省哲学社会科学研究规划项目（20JYB031）。

提出一种基于实时共现网络的微博话题发现模型[5]；宋莉娜等人提出了 SOM 聚类方法用于微博的话题发现，研究表明该方法可以有效改善传统文本聚类不准确的缺点，从而有效的发现微博话题[6]。

情感分析，又被称为观点挖掘，是一种分析、处理、归纳和推理具有情感色彩的主观文本的过程[7]。情感分析主要包括机器学习和基于情感词典两种方法。分析研究用户发布的观点看法在很多领域有着非常重要的作用，对于用户情感的挖掘研究具有广泛的应用价值，目前对此国内外已有诸多学者开展了研究。在国外，对于网民情感态度的研究主要集中于 Twitter、Facebook 等社交平台上，Bollen 等人对发布在 Twitter 平台上的推文进行情感分析，并以日为单位计算时间轴上的情绪向量，进而对网民的情感态度进行分析与预测[8]。由于基于中文环境的微博与基于英文环境的 Twitter 在语言表达习惯上存在着很大的差异，因此用于微博文本的情感分析工具与 Twitter 平台上的情感分析相比有很大不同。刘智等人从集成学习的角度出发，设计了一种基于样本空间动态划分的机制，在此机制上构建了微博文本情感分类器，通过实验实现了大规模评论集的情感分析以及用户观点挖掘[9]。史伟等人提出了一种基于 KBANN 的情感分析方法来解决没有情感关键词存在的文本，通过构建隐性知识来推测文本的情感状态[10]。

众多研究学者为微博话题发现和舆情文本情感分析注入了新的研究方法和思想理念。而基于多源数据挖掘与融合来研究舆情文本情感与舆情演化规律的研究很少。故本文从多源数据角度出发，利用文本情感分析技术，对不同数据源中的网络舆情情感状况进行分析，实现对网民情感的挖掘，为网络舆情的引导和控制提供有益借鉴。

三、数据采集与预处理

（一）数据源选取

在中文语言环境中，舆情案例的数据源一般都来自新浪微博。它是一个为大众提供信息交流共享和娱乐休闲的平台。据央视财经统计，截止 2020 年第三季度，微博的月活跃用户数达 5.11 亿。因此，以新浪微博为数据源进行的研究具有一定的代表性。

（二）数据采集

在明确研究对象和数据来源后，要对舆情案例的相关数据进行采集。根据新浪微博平台的数据开放程度和网页结构特点，采用 Python 软件通过网络爬虫的方式获取文本数据，并且有针对性的编写 Python 脚本抓取微博文本数据。

利用新浪微博的高级搜索功能，选定时间范围为 2019 年 3 月 1 日到 2019 年 8 月 30 日，以“经贸磋商”为搜索关键词，编写 Python 爬虫程序进行数据采集，采集的主要字段包括：用户名、发布内容、发布时间。共采集到 17436 条微博文本数据。

（三）数据预处理

由于微博平台具有大众化，不受任何的时空限制，灵活度较高的特点，用户在发表博文的过程中，不会受到文字格式的约束，因此文本内容中往往包含大量噪声数据，如网址 HTML 标签、话题标签、无用的表情符号等。这些噪声数据对文本的分词和词频统计都会造成影响，所以在数据预处理阶段要对这些无意义的信息进行清洗。

使用正则表达式对文本内容数据进行清洗，删除重复的文本数据，删除 @、数字、无用网址、表情等无关内容，提取文本内容，再将清洗后的数据进行分词处理，利用 python 中的 JIEBA 分词工具包，对文本内容逐条进行分词，去除停用词、标点符号等无意义的词。对处理好的数据进行高频词统计并绘制词云图，其结果如表 1 和图 1 所示。

表 1　微博文本词频 Top10

词语	频次	词语	频次
中国	1238	拭目以待	563
加油	1126	奉陪	515
强大	969	祖国	502
理性	653	光明	423
吓倒	621	冷静	359

数据来源：根据采集的微博文本数据统计整理而得。

由高频词可以看出，网民支持国家做出的决定，纷纷为国家加油打气，表示中国绝不会被此事件吓倒，此事件的发生会让国家变得越来越强大，不畏惧对方提出的挑战，表现出了网民的爱国主义情怀。

三、情感词典构建

情感词典包含基础词典和基于特定事件情境下的情感词典。利用大连理工大学开发的情感词典作为基础词典，但在针对某一特定事件的研究，只利用基础词典中的情感词往往不够准确，因此在研究特定事件中网民的情感态度时，需要加入有关于该事件情境下的高频词汇。因此，通过人工筛选，对比大连理工大学情感词典本体库对情感词的打分情况，构建经贸磋商事件情境下的特定情感词典。最终统计得到情感词包括“中国”、“中美”、“经贸磋商”等在大部分文本中都出现的词

语，权重较高，因此需要去除这些词语。利用大连理工大学情感词典本体库进行对比，如词库中某个词为积极情感词，而计算后为消极情感词，则对其分数进行校正。若校正之后大于 0，则归入积极情感词典中，若校正后仍然小于 0，则继续留在消极情感词典中。将校正后的分数作为该词的最终情感分数。表 2 中序号 1~10 为排名前十的积极情感词，序号 11~20 为排名前十的消极情感词。

表 2　积极情感词 Top10(1~10) 与消极情感词 Top10(11~20)

序号	词语	情感得分	序号	词语	情感得分
1	效率	0.99	2	无所畏惧	0.98
3	加油	0.88	4	强大	0.86
5	振兴	0.84	6	志同道合	0.82
7	繁荣	0.80	8	支持	0.78
9	包容	0.76	10	幸福	0.75
11	严重	−0.99	12	霸权	−0.89
13	影响	−0.85	14	打压	−0.83
15	制裁	−0.78	16	崩溃	−0.74
17	危机	−0.71	18	压倒	−0.70
19	纠纷	−0.68	20	限制	−0.66

数据来源：与大连理工大学情感词典本体库情感分值对比计算而得。

四、文本情感分析

（一）微博文本情感强度计算

基于中文文本情感词典，计算 17436 条微博文本数据的情感得分。情感得分取值范围为 [−1,1]，若情感得分大于 0 则判定该文本情感为积极倾向，情感得分小于 0 则判定该文本情感为消极倾向，情感得分等于 0 则判定该文本情感为中性。根据计算结果，最终得到 13526 条积极情感微博，占比为 77.6%；消极情感微博 3298 条，占比为 18.9%，中性情感微博 612 条，占比为 3.5%，图 2 为微博情感极性分布结果。

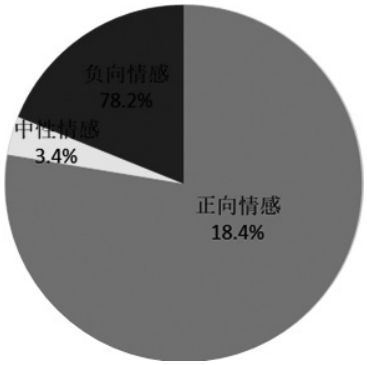

图 2　微博情感极性分布结果

数据来源：对采集的微博文本进行情感强度计算而得。

（二）微博发文趋势分析

图 3 为微博积极情感强度时序图，可以从图中看出 2019 年 5 月 ~2019 年 8 月期间网民对经贸磋商结果的情感强度高于 2019 年 2 月 ~2019 年 5 月期间的情感强度。且网民的积极情感强度在 2019 年 5 月 15 日达到峰值，当天积极情感博文为 2669 条。次高峰发生于 2019 年 5 月 23 日，博文数量为 2352 条。

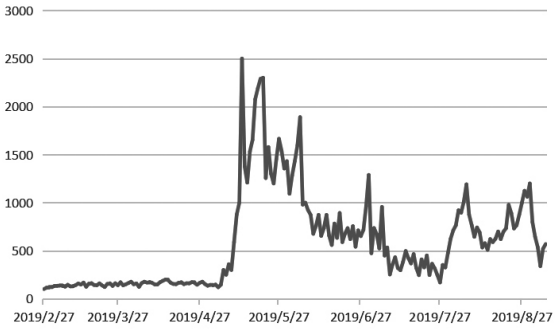

图 3　微博积极情感强度时序图

数据来源：根据每日发博数量和情感分析统计整理而得。

图 4 为微博消极情感强度时序图，整体的变化趋势与积极情感强度时序图呈现的效果一致，同样在 2019 年 5 月 15 日消极情感强度到达低谷，当天发文数量为 1130 条。次谷值同上也发生在 2019 年 5 月 23 日，发文数量为 1091 条。但消极情感强度的分值低于积极情感强度分值，大约为积极情感强度分值的一半，经分析得到在经贸磋商期间，微博上网民表现出的积极情感占多数，并且积极情感强度要远大于消极情感强度。

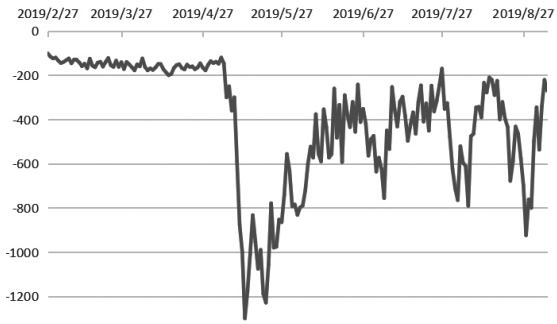

图 4　微博消极情感强度时序图

数据来源：根据每日发博数量和情感分析统计整理而得。

五、网络舆情主题聚类分析

（一）主题的确定与发现

通过上述分析可以发现在微博平台上网民的积极情感占多数。由整个事件可以看出，随着事件的发展在主要时间节点上网民的情感状态会产生波动，由于两国之

间存在着文化差异，导致双方的观点立场不同，造成情感倾向的主要原因也会不尽相同。

通过对情感分析之后的文本进行主题聚类分析，挖掘每种情感下的子主题。通过构建 LDA 主题模型，将有关“经贸磋商”的文本进行聚类和主题提取。LDA 主题模型是通过给出每个主题下的高频词来确定当前的主题内容，利用每个主题的主题词还原网民讨论的热点话题。由于 LDA 主题模型没有明确的主题个数，因此要经过不断调试与对比分析才能得出最优的主题数量。经过调试最终确定积情感为 5 个讨论主题。表 3 为 LDA 主题模型提取的各个主题关键词。

表 3　微博各主题关键词

微博积极情感主题关键词

Topic1	华为	中国	加油	雄起	国家	支持	利益	公司
Topic2	创新	企业	国家	安利	市场	发展	鼓励	行业
Topic3	贸易	环境	周边	优化	大国	国家	开放	氛围
Topic4	中国	市场	经济体	世界	创造	自信	依赖	克服
Topic5	发展	企业	结构	快速	产业	调整	贸易	促使

数据来源：根据 LDA 主题模型提取结果整理而得。

从微博积极情感主题 1 可以看出，国民表示支持华为、华为加油等，为民族企业加油打气。主题 2 反映了鼓励大众进行创新，不畏惧挑战。主题 3 反映了中国不断优化对外贸易环境，营造了良好的对外贸易氛围，塑造了大国形象。主题 4 反映了中国可以克服自身不足，摆脱对其他国家的技术依赖，在世界经济体系中更加自信自强。主题 5 反映了中国的产业结构因此会做出调整，使得企业可以快速发展。

（二）微博信息分析

对爬取到的数据分析发现，原创微博的占比为 39.7%，转发占比为 60.3%。其中 39.7% 的网民利用微博平台，发表原创信息表达对此事件的看法和意见。对网民的情感分析可以发现，大部分网民能够理性看待该事件发生的前因后果，60.3% 的网民通过转发官方微博的方式表达自己对该事件的态度，将该事件话题传播的范围扩大，并引导其他网民支持自己国家所做的决定，进一步提高了该事件的积极影响力和传播效果。

六、结论

通过对网民的情感分析可以得到，网民对于事件的情感变化会受到主流媒体报道、周围用户和新闻内容的影响，因此相关部门和政府应该充分利用主流媒体，把控好网民情感变化的节点，有针对地对网络舆情进行引导管控。舆情信息爆发快、蔓延广、消散期后舆情信息不断，相关管理部门要加强对突发事件网络舆情的信息管理。在事件舆情突发期，把握舆论信息导向，引导网民参与正向的、积极的舆论讨论中；在蔓延期应发挥意见领袖作用，主流媒体应及时发布信息资讯，避免舆情传播的过程中谣言的产生；在消散期应重视各大网站的信息推送，保证推送信息的准确性，避免出现衍生舆情。

[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.

[2] CHEN Y,AMIRI H,LI Z,et al. Emerging topic detection for organizations from microblogs[C]// Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval.Dublin: ACM,2013:43-52.

[3] STILO G,VELARDI P.Efficient temporal mining of micro-blog texts and its application to event discovery[J].Data mining and knowledge discovery,2016,30(2):372-402.

[4] HU M,LIU B. Mining and summarizing customer reviews[C]//Proceedings of the ACM SIGKDD international conference on knowledge discovery & data mining.Seattle:ACM 2004,168-197.

[5] 李亚星 , 王兆凯 , 冯旭鹏 , 等 . 基于实时共现网络的微博话题发现 [J]. 计算机应用 ,2016,36(5):1302-1306.

[6] 宋丽娜 , 冯旭鹏 , 刘利军 . 基于 SOM 聚类的微博话题发现 [J]. 计算机应用研究 ,2019,35(3):671 -679.

[7] 赵妍妍 , 秦兵 , 刘挺 . 文本情感分析 [J]. 软件学报 ,2010,21(8):1834-1848.

[8] BOLLEN J,MAO H,PEPE A.Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena[J].Computer science,2009,44(12) :2365-2370.

[9] 刘智 , 杨宗凯 , 刘三 (女牙), 铁璐 . 一种基于样本空间动态划分的中文情感识别方法 [J]. 计算机应用研究 ,2013,30(5):1443-1447.

[10] 史伟 , 王洪伟 , 何绍义 . 基于 KBANN 的文本情感识别研究 [J]. 情报理论与实践 ,2015,38(3):112-115,121.

（责任编辑：顾晓滨　刘晓辉）