

疫情期间微博网友的情感变化及其影响因素

——以人民日报相关微博及评论为对象的关联研究

马博宏 171860638

程珂 171860640

南京大学信息管理学院

摘要

【目的/意义】重大突发公共卫生事件背景下，群众情感变化规律及影响因素探究极为重要。这项研究既可以为突发事件中的长期舆论引导提供思路，又能为短期舆论纠正提供方法。而舆论调控的效果，直接影响到政府公信力的强弱，民众对各项政策的贯彻落实程度。因此，新冠肺炎防控期间以新浪微博为对象的网友情绪变化及影响因素研究，有着很强的必要性与广泛的应用空间。

【方法/过程】1. 收集了“人民日报”的部分疫情相关微博的评论和二级评论。2. 基于 SO-PMI 方法进行评论的情感分析，以天为单位生成情感指数。3. 以每日新增确诊为基础数据生成每日疫情指数，并总结梳理疫情期间发生的大事件、热点事件。4. 通过每日情感指数和每日疫情指数的对比，热点事件的引入，分析微博网友情感指数的变化及影响因素。

【结果/结论】1. 疫情发展对微博网友的长期情感水平有较大影响。2. 热点事件对微博网友的短期情感波动有较大影响。3. 政府行为对微博网友的情感变化具有强力影响。

关键词：关联分析、情感指数、疫情指数、微博

1. 介绍

基于算法的量化工作最早出现在上个世纪，于近几十年开始流行，为数据挖掘的应用提供了巨大潜力。情感分析已经提出并在金融、银行或网络应用等领域得到了广泛应用。其相关方法在评论预测、商业预测与热度预测中被广泛接受。

虽然已经有涉及了疫情的数据分析研究，但我们并未在各大文献库发现某一项研究侧重于疫情期间人民群众的情感分析。根据研究[1]，在一个现代社会，信息的传播很大程度上依赖于社交媒体，社交媒体用户在使用社交媒体或社区时会留下大量的数据。因此，有理由相信，重大突发事件的情感指数可以依赖社交媒体用户的评论、标记和流行论坛和网站上的其他数据而估计计算。本文尝试进行这种分析工作，分析新型冠状病毒肺炎防控期间（本文中指武汉封城当日至湖北清零后一日，即 1 月 23 日至 4 月 27 日）微博用户对“人民日报”机构在其新浪微博账号上发布的新冠疫情相关消息的评论，并运用自然语言处理与情感词典方法得到相应的情感指数。

本文得到情感指数后，将其与疫情发展情况和疫情期间的大事件做对比，对相关部门制定数据发布策略，网友从数据中获取实际信息，调节自身情感等都十分有意义。从社会科学层面上，重大突发事件发生时期的群众情感变化也值得关注。自新中国成立以来，国家第一次经历规模如此庞大的全国性传染病防治行动，疫情期间由于诸多原因，部分网友的情感出现过很大的波动。中国境内尤其是湖北病例清零后，对疫情期间整体的情感指数做以分析，并结合疫情进程与重大事件日历，对两者进行关联分析，对类似事件的应对具有重要意义。

2.情感分析

2.1 情感分析简介

随着互联网的发展，近年来网站文本量迅速增加，学者们开始以量化的有效方式挖掘复杂文本中的知识或有价值的信息。情感分析也随之而来，其中一种传统的情感分析方式是使用情感词典。

如论文[2]所言，情感词典最早出现在 1998 年惠塞尔的实验[3]。在实验中，惠塞尔要求他的 148 名志愿者用情感词来描述科学领域的术语。这个实验产生了第一个情感词典，并随后导致搜索情感分析的流行。

从那时起，一个典型的使用情感词典的方式被定义：先将文本切成几个单词，并匹配情感词典中的每个单词。再根据词典中记录的单词的情感倾向，将单词量化为一个数字。即文本背后的情感最终被定义为一个数字。

然而，情感词汇在本次疫情评论分析中的可用性上却处于劣势，主要是由于它总是在可数的字数上受到限制，并且无法附加实践使用的水平。如上文所述，词典的词汇通常来自某些特定领域，不像真正的词典那样全面和权威。为解决这一问题，一些学者提出了扩展情感词典的方法。

其中，论文[4]提供的方法是 PMI。PMI 是点对点的相互信息算法，在情感分析中代表两个词之间的关系。这两个词在文本中出现的频率越高，这两个词的关系就越密切。如果目标词与正（负）字高度连接，则更有可能为正（负）。如果目标词为 c ，则正（负）字为 p ，适用于文本挖掘的 PMI 的公式为：

$$PMI(c, p) = \log_2 \left(\frac{count(c, p)}{count(c) * count(p)} \right)$$

计数 (x) 是含有单词 x 的句子在文本中的个数。 $count(x, y)$ 是同时含有 x 与 y 的句子在文本中的个数。

2.2 PMI 算法的优化

PMI 算法可以使得情感词典得以扩张，但是在疫情评论数据的分析中，仍然很难得到有效的结果（即大量评论的情感标注与人工判定不匹配），PMI 仍不足以成为目标词的情感指数，我们认为因为它只代表一个词和目标词之间的关系。因此又出现了 SO-PMI。目标词情感指数，可以通过计算目标词和一组单词之间的 PMI 得来；有些单词的情感倾向是正面的，有些是负面的。对这正面词组与负面词组分别求和 PMI，SO-PMI 即为正面词组和减去负面词组和的差。它表示目标词与一组正词和一组负词之间关系的总和。SO-PMI 的公式是：（在文本中假定存在 k 个正面词和 m 个负面词，目标词为 c ）

$$SO - PMI(c) = \sum_1^k PMI(c, p) - \sum_1^m PMI(c, n)$$

在本文的实际计算过程中，目标词与疫情评论分词后的一些词之间的 PMI 可能为 0（即许多单词在文本中不会同时出现），但其评论明显有情感偏向，为解决这一问题，我们参考[5] [6]的研究中，改变了 SO-PMI 的公式，使用平滑方法增强了 SO-PMI 的合理性，使目标词的情感指数预测更加准确。

3.研究过程

3.1 收集数据

在情感分析过程中，所有的数据来自新浪微博中“人民日报”账号所发布的微博，但由于其在此期间发布的所有微博并非都与疫情相关，为统一前后时间维度上的数据分布，设置如下筛选方式：1. 采集每天统计全国累计确诊或新增确诊人数（包括新增无症状感染者）的相关微博的评论；2. 除 1 中情况外，采集与疫情相关且评论数大于 10000 的微博的评论；3. 每条微博最多采集前 20 条微博评论。4. 每条微博评论最多采集前 20 条二级评论。

每条微博的地址采集方式为手工采集，汇总地址后，使用 Python 方法进行微博评论及二级评论的采集。最终采集微博评论 1659 条，二级评论 14729 条。

3.2 数据预处理

在情感分析之前，我们删除空评论、以非中文做出的评论和表情符号，使用 R 对所有评论进行分词。每个评论被分成几个单词。切割过程加载了一个百度文库文档[7]中的新冠肺炎有关词汇列表，根据其中文列表确保评论中的某些疫情相关的专有名词不被切分。分词之后，使用百度提供的停用词表组合删除一些不相关的单词。最后，我们从评论中获取 26000 个单词列表组合，每个列表均代表一个评论，由若干个单词组成。

3.3 情感词典处理

I. 合并情感词典 Hownet、清华和 Ntusc 的组合，准备了一个情感词典。删除了重复的单词，并将词典分为两部分：正面词列表和负面词列表。情感词汇的基本情况如下：

	Hownet	Tsinghua	Ntusc	Combined lexicon
积极词汇数	4566	5567	2811	10220
消极词汇数	4370	4469	8276	13754

II. 根据文档[7]，收集了 70 个流行的新冠专有名词。

根据步骤 I 中的词表和 26000 个评论分词列表计算了每条评论的 SO-PMI。SO-PMI 的计算方法基于研究[4]，并通过研究[5]引入的平滑技术得到了改进。SO-PMI 的改进公式是：

$$SO - PMI(c_i) = \log_2 \left(\prod_{j=1}^n \frac{\text{count}(c_i, p_j) + 1}{\text{count}(c_i, r_j) + 1} \right)$$

在这个公式中， $\text{count}(c_i, r_j)$ 是情感词典中的负面词和目标词 c_i 都出现在全部评论中的次数； $\text{count}(c_i, p_j)$ 是情感词典中的正面词和目标词 c_i 都出现在全部评论中的次数。对 1 月 23 日至 4 月 27 日，每天的评论中的每个词都有一个 SO-PMI。

若 SO-PMI 是一个正数，标记目标词为正面情感倾向；如果为负数，则标记负面情感倾向。

将准备好的新冠专有名词事先带入情感词典中并且计算 SO-PMI。标记其对应的情感倾向并且加入到情感词典当中。最终构建的新的情感词典包含数千个条目。每个单词都一定存在于正词列表或者负面单词列表中。

3.4 情感指数生成

接下来对一段评论对应的词列表中的每个单词查找其情感标记。根据研究[8]，量化情感标签的数字，称为情感指数。量化方法：正面标记 1，负面 -1，对 7 个否定词（“不要”，“不用”，“不

必”) 量化为-1（根据研究[9]）；根据研究[10]，量化 14 个转折连接词（如“但是”,“只是”）为 1.2；最后，下表的方式方式量化了的所有程度副词：

（表来自研究[8]，第 3 页）

级别	词汇	权重
6 - 最	极度、非常、倍加、极其……	1.7
5 - 超	超过、过于、何止、过度……	1.5
4 - 很	尤其、特别、多么、格外……	1.3
3 - 较	进一步、较为、更加、更为……	1.1
2 - 稍	稍微、略微、稍许、些许……	0.9
1 - 欠	相对、不大、不怎么、微微……	0.7

最终，对一段评论的情感指数公式如下：

$$SI_{comment} = (-1)^{m_1}w_1q_1 + (-1)^{m_2}w_2q_2 + (-1)^{m_3}w_3q_3 + \cdots + (-1)^{m_n}w_nq_n$$

在此公式中， w_n 是评论中第 n 个词查询情感词典得到的情感指数； m_n 是 w_n 和 w_{n-1} 之间的否定词个数； q_n 是 w_n 和 w_{n-1} 之间的转折助词和称度副词的对应情感指数。

最后计算每一天的情感指数：

$$SI = (\sum SI_{comment})/l$$

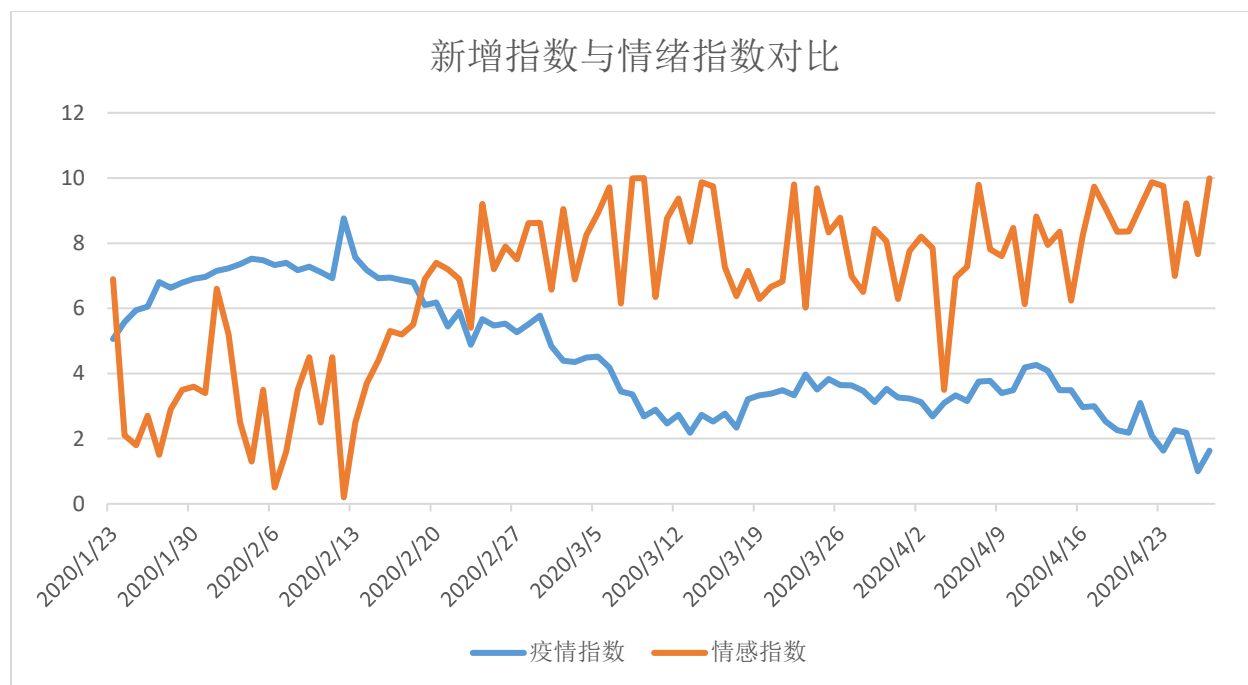
若某一天采集了超过一条评论，则计算两条评论的 SI 值的均值作为这一天的情感指数，最终得到 1 月 23 日至 4 月 27 日每一天的情感指数。

3.5 新增指数生成

生成情感指数后，我们将进入探究疫情发展情况与疫情期间的大事件对微博网友情感指数的影响，因此对比变量须反映疫情的发展进程，在本次研究中，我们最终选择单日新增确诊人数作为原始变量，对其进行对数运算等标准化方法后，生成 1 月 23 日至 4 月 27 日的每日疫情指数，与每日情感指数进行对比分析。

4.研究结果

根据第 3 部分的研究方法，2020 年 1 月 23 日至 4 月 27 日共计 96 天的每日疫情指数与情感指数对比图像如下：



可以看出，疫情指数自 1 月 23 日逐步上升，2 月 12 日达到最高点，对应 2 月 12 日单日新增确诊病例超过 1.5 万人，2 月 12 日后持续下降，至 3 月中旬稳定波动，3 月下旬又有所回升，至 4 月 12 日达到新的高点，对应 4 月 12 日单日新增确诊 108 例（其中 98 例为境外输入），4 月 12 日后持续下降，趋势中有小范围波动。

情感指数则明显可以分为两个阶段。

第一阶段是 1 月 23 日至 2 月中旬，这一时期微博网友的情感指数总体来说处于低位，负面情感占据主流，波动比较剧烈。几个比较重要的时期如下：1 月 24 日至 26 日，情感指数骤然下降到比较低的位置，这期间发生了武汉“封城”、全国诸多省份相继启动重大突发公共卫生事件一级响应等事件。1 月 27 日至 31 日，情感指数有所回升，这期间发生了武汉开建两所传染病防控医院等事件。2 月 1 日情感指数达到高点，可能由于 1 月 31 日世卫组织讲新冠疫情列为国际关注的突发公共卫生事件，国务院决定调拨全国物资支援湖北。2 月 1 日后一周，情绪指数持续回落，这期间主要发生了湖北省红十字会物资调拨广受质疑的事件，在这一期间，湖北省、武汉市、湖北省红十字会等单位党政领导相继被免职，调换。2 月 12 日情感指数达到新的低点，也是整个疫情期间的最低点，期间发生的事情主要有：2 月 12 日调整确诊标准，当日新增确诊病例约 1.5 万例，全国医务人员累计确诊近 2000 例，其中 6 人不幸死亡等。

第二阶段为 2 月中旬至 4 月 27 日，这一时期微博网友的情感指数总体来说处于高位，正面情绪占主流，也伴随有波动，主要的极值点与同期发生的主要事件有：3 月 22 日高点：意大利单日确诊逾 6600 人创单日最大增幅；4 月 4 日低点：留学生归国确诊事件多次曝出；4 月 9 日高

点：欧洲老人被迫同意放弃抢救事件发酵；4月25日高点：美国累计确诊超过80万例。

结合疫情指数与情感指数的发展，我们参考研究[11]将新冠肺炎防控期间，全国疫情与微博网友情绪的综合发展分为4个时期。

1月22日之前为潜伏期。在这一时期，微博网友关注的要点并非新冠肺炎，新冠肺炎的热度与影响力较小，除“武汉8人造谣事件”外并未有与新冠肺炎相关的事情引起热议，因此，微博中的情绪指数与新冠肺炎疫情发展关联程度较小。

1月23日至2月中旬为爆发期。在这一时期，新冠肺炎迅速成为参与人民日报评论的微博网友的首要关注事件，网络原生情绪反应正式进入爆发期。其主要特点为，微博网友关注度段时间急剧升高，媒体关注点高度聚焦，次生情感大范围传播，政府情感引导力量迅速加强。在该阶段，微博网友情感主体的不稳定性特征表现得十分突出，二次情感传播也变得更加活跃与复杂。事件处于爆发期时，整个微博网友的情感综合复杂而具有感染力。

2月中旬至3月中旬为平稳期。在这一时期疫情的发展进程主要是：随着抗疫的突进，全国新增确诊病例逐步减少，清零城市逐渐增多，中国式抗疫和制度优势得到微博网友认可。相对应的情感指数与发展主要为：微博网友仍持续关注新冠疫情，但原生和次生情绪力量都逐渐减弱，各大媒体的总结发声能力增强，能有效引导微博网友情绪，总体情绪变为积极正面。

3月中旬至4月27日为消散期，从这一时期的疫情进程来看，境外输入成为我国新增确诊病例主要来源，除输入病例引起聚集性传播的地区如黑龙江、广东部分地区外，全国大部分地区的生产生活恢复正常，微博网友对国内新冠肺炎疫情的关注度持续下降。从这一时期的情感指数来看，微博网友的情绪总体来说处于正面，主要的负面情绪来源为境外输入对我国疫情防控造成的压力。

5. 主要结论

5.1 疫情发展对微博网友的长期情感水平有较大影响

对情感指数的总体两阶段，容易看出其划分主要来自新冠疫情发展的影响，自1月23日武汉封城起，新冠疫情进入爆发期，期间网友的主要情感为：焦虑、担忧、恐慌。随着疫情在2月上旬前持续扩大，网友的情感水平也在波动中持续下降，情感指数在1月中旬到达最低点。

随着2月下旬起疫情度过高峰，国内正常生产生活不断恢复，疫情对微博网友的印象逐渐减弱，网友的情感指数平均值虽然还有波动，但总体均值回升至较高水平。

5.2 热点事件对微博网友的短期情感波动有较大影响

观察情感指数变化趋势中与左右相邻值差异较大的极值点，如 1 月 23 日、2 月 1 日、月 12 日、3 月 22 日、4 月 4 日、4 月 9 日、4 月 25 日等，如上文所述这些时间点周围均发生了正面或负面的热点事件，造成微博网友情绪短时间内的剧烈波动，但这种波动持续性较弱，情感指数的最终发展仍然受疫情发展的长期影响。

5.3 政府行为对微博网友的情感变化具有强力影响

政府行为与力量在重大突发公共卫生事件中扮演着至关重要的角色，从正面发声到侧面有道，政府力量的发挥客观地影响着微博网友情绪的波动。政府防控措施的布置直接影响微博网友情绪的发展，政府直接或间接的宣传教育会对微博网友的评论及情感产生直接的影响。

6.研究的不足与展望

6.1 精力所限，基础数据准备不足

从网友的选取上，本文仅选择对“人民日报”的相关微博进行评论的网友，对全体微博网友甚至全网网友缺乏代表性，后续研究可以扩大样本选取平台和范围，提高代表性。

在微博的选取中，由于人民日报每日发布微博数量较多（日均 10 条以上），且不同时期微博与疫情相关的比重并不相同，为统一前后比重，本文选择每日仅采集报送每日新增或累计确诊的微博，以及评论数一万以上的疫情相关微博。并未全面收集所有疫情相关微博，后续可以将这些微博更加全面的收集。

在评论的选取中，为提升采集代码的性能，限定每条微博最多提取 20 条微博评论，每条微博评论最多提取 20 条二级评论，则每条微博最多提取约 400 条评论，相比疫情相关微博的评论数平均在 1000-10000 的范围，缺乏代表性。若进行后续研究，可以考虑采集每条微博的全部评论。

6.2 方法所限，情感指数仍有漏洞

情感词典在准确性和效率方面仍然存在弱点，由于缺乏该领域的训练数据集注释，后续可以考虑结合情感词典与交叉训练的可能性，以提高情感分析的效率和准确性。

6.3 知识所限，综合分析较为片面

新冠疫情的发展是一个复杂的医学及公共卫生问题，舆论情绪的引导也是复杂的社会学问

题, 本文参考多种分析方法, 对疫情发展与网友情绪发展进行量化和对比, 一定程度上的得到部分结论, 但缺乏更深入的实验论证和更广泛的理论支撑。因此只能较为片面的反映两个研究主体的部分关联。后续研究将采用更科学的方法, 更严谨的论证过程, 以期得到更符合现实的结论。

参考文献

- [1]孙瑞祥. 当代中国流行文化生成机制与传播动力阐释[D]. 天津师范大学, 2009.
- [2]洪巍, 李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(04): 750-757.
- [3]Whissell C. Objective analysis of text:II.Using an emotional compass to describe the emotional tone of situation comedies[J].Psychological Reports, 1998, 82 (2) :643-646
- [4]Peter D. Turney,Michael L. Littman. Measuring praise and criticism[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4).
- [5]杜锐, 朱艳辉, 田海龙, 刘璟, 马进. 基于平滑 SO-PMI 算法的微博情感词典构建方法研究 [J]. 湖南工业大学学报, 2015, 29(05): 77-81. [6]姜伶伶, 何中市, 张航. 基于 Good-Turing 平滑 SO-PMI 算法构建微博情感词典方法的研究[J]. 现代计算机(专业版), 2018(10): 15-20.
- [7] 新冠肺炎有关词汇
[<https://wenku.baidu.com/view/4ecad0e8bdd126fff705cc1755270722182e595e.htm>
1]
- [8]傅魁, 刘玉洁, 陈美丽. 基于财经新闻情感倾向值的股票价格预测[J]. 北京邮电大学学报 (社会科学版), 2019, 21(01): 87-100.
- [9]刘相臣, 丁崇明. 近百年现代汉语否定副词研究述论[J]. 江西师范大学学报(哲学社会科学版), 2014, 47(06): 91-100.
- [10]邸鹏, 李爱萍, 段利国. 基于转折句式的文本情感倾向性分析[J]. 计算机工程与设计, 2014, 35(12): 4289-4295.
- [11]邢鹏飞, 李鑫鑫. 重大疫情防控中网络舆情形成机制及引导策略研究——基于新冠肺炎疫情期间网络舆情文本的质性分析[J/OL]. 情报杂志: 1-9[2020-04-28].