Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5053-5061
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118530

Hans汉斯

基于BERT模型的网络舆情情感分析

——以上海疫情为例

孙　洋[1*]，冷冠男[2]

[1]上海工程技术大学管理学院，上海
[2]上海工程技术大学电子电气工程学院，上海

收稿日期：2022年7月1日；录用日期：2022年7月27日；发布日期：2022年8月4日

摘　要

随着互联网时代的不断演进，各种社交软件为网络舆情的传播提供了主要平台。疫情发生后对网络舆情信息的搜集、分析和引导，对于相关部门开展舆情解析和稳定公众情绪具有现实意义。本研究基于新浪微博数据，以此次上海疫情为例，对疫情期间的微博相关数据信息进行提取和挖掘，首先用TF-IDF词频统计处理方法统计出关键节点日期的主题词，然后使用BERT模型对此次疫情的舆情情感进行分析。结果表明：在时间分布上，博文的发布数量整体呈“n”字形分布。在舆情情感上，负面感情前期一直占了较大的比例，中期无明显变化，后期民众的感情虽依旧起伏不定，但消极情绪总体上有了明显的下降趋势。本文研究结果可为日后疫情防控中网络舆情的引导和处置提供借鉴参考。

关键词

网络舆情，情感分析，上海疫情，BERT模型

Sentiment Analysis of Network Public Opinion Based on BERT Model

—Taking the Shanghai Epidemic as an Example

Yang Sun[1*], Guannan Leng[2]

[1]School of Management, Shanghai University of Engineering Science, Shanghai
[2]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai

Received: Jul. 1st, 2022; accepted: Jul. 27th, 2022; published: Aug. 4th, 2022

Abstract

With the continuous evolution of the Internet era, various social softwares provide the main plat-

*通讯作者。

文章引用: 孙洋, 冷冠男. 基于 BERT 模型的网络舆情情感分析[J]. 应用数学进展, 2022, 11(8): 5053-5061.
DOI: 10.12677/aam.2022.118530

form for the dissemination of network public opinion. The collection, analysis and guidance of online public opinion information after the outbreak of the COVID-19 is of practical significance for relevant departments to carry out public opinion analysis and stabilize public sentiment. This study is based on Sina Weibo data, taking the Shanghai epidemic as an example, to extract and mine Weibo-related data information during the epidemic. First, the TF-IDF word frequency statistical processing method is used to count the subject words of the key node dates, then use the BERT model to analyze the public sentiment of the COVID-19. The results show that: in terms of time distribution, the number of blog posts is distributed in the shape of "n" as a whole. In terms of public opinion and emotion, negative emotions have always accounted for a large proportion in the early stage, and there was no significant change in the mid-term. Although the people's emotions in the later period still fluctuated, negative emotions showed a clear downward trend in general. The results of this paper can provide reference for the guidance and disposal of network public opinion in the future epidemic prevention and control.

Keywords

Internet Public Opinion, Sentiment Analysis, Shanghai Epidemic, BERT Model

1. 引言

新冠肺炎自 2019 年年底爆发以来，全国各地人民的生活都受到了一定的影响，在整个疫情防控过程中，网民主要是通过各个社交平台来了解新增病例、病毒溯源、各地防控措施等疫情相关信息。由于网络舆情的传播具有话题密集性、演化多变性和衍生多因性等特点，多元主体在线上平台以及线下生活中会围绕相关热点进行交谈、协商和博弈[1]。在整个疫情防控过程中，如果对于网上的舆论没有加以合理地引导和治理，那么网络上的负面舆情就有可能成为大众的“迷魂汤”和社会的“分离器”[2]。因此，对疫情等非常规突发事件发生后的网络舆情信息的监控、处理和引导，对于相关部门安抚社会公众情绪具有重要意义。

现如今，国内外的学者对于非常规突发事件的网络舆情研究，大多是基于文本数据，在这其中，情感分析就成为了极其重要的一部分。黄仕靖等人基于 SnowNLP 情感分析工具得出，网民群体中受教育程度低的群体更容易受到舆论的影响[3]。陈娟等人认为政府官方信息的发布对于用户的情感倾向起到明显的正面影响[4]。凌海彬等人提出了一种多特征融合的文本情感分析模型，使其能更好地捕捉用户的情感语义[5]。胡思才等人则将情感极性值和词频信息融入到模型中，提高了情感文本分类的准确性[6]。最近几年，随着互联网的飞速发展，具有海量信息的新浪微博(以下简称微博)社交媒体平台为网络舆情的检测与分析提供了大量的数据[7]。在微博上，某条高点赞、高评论以及高转发的微博，会影响其他用户的行为，当某些个体认为这个观点合理时，就会表达出与之相似的看法[8]。微博上互动较多的微博条文以及热门评论在很大程度上体现了网民的观点，其带有明显的个人情感特征，体现出了不同网民的情感倾向性[9]。李宗敏等认为热门评论对于所接受到这个信息的其他用户来说，具有引导其情感态度的作用，如果热门评论的情感态度是积极的，那么在下面的评论中就大概率会形成良好的反馈，反之，则会带来消极的影响[10]。同时，Li 等人通过对获取的微博数据的分析，发现自从新

冠疫情爆发后，网民的负面情绪较多，对于娱乐等方面的事情关注度下降，在自己和家人的健康问题上则更加用心[11]。Gao 等人发现在疫情期间，人们接受的社交网络上的信息对于抑郁症和焦虑症等精神问题的发作存在着一定的关系[12]。

本文以 2022 年上海疫情为例，基于新浪微博数据，结合疫情相关态势的发展，研究此次疫情事件中网络舆情的发展特征以及民众的情绪变化，并将分析结果可视化呈现，得到微博关键词分布图、情感变化图等信息，为日后的非常规突发事件应急管理提供参考。

2. 研究背景

2.1. 上海本次疫情概况

上海疫情每日新增病例数数据由上海发布汇总获得，具体数据如图 1 所示。3 月 1 日至 3 月 22 日态势稳定，每日新增控制在千例以内。3 月 23 日起，疫情新增病例数据明显。3 月 29 日是本次疫情第一个拐点，疫情新增病例数量开始下降。然而 4 月 1 日起，病例新增数量暴增，4 月 4 日起日增已达万例。4 月 10 日、4 月 12 日、4 月 16 日均呈现出短暂拐点，随即又继续攀升。4 月 22 日在日增达到 23,300 例之后，开始出现明显的下降趋势，4 月 28 日虚晃一枪，然后便一直呈下降态势，4 月 30 日日增已下降到万例以下。随后便一直下降，到 5 月 29 日已下降到百例以内。

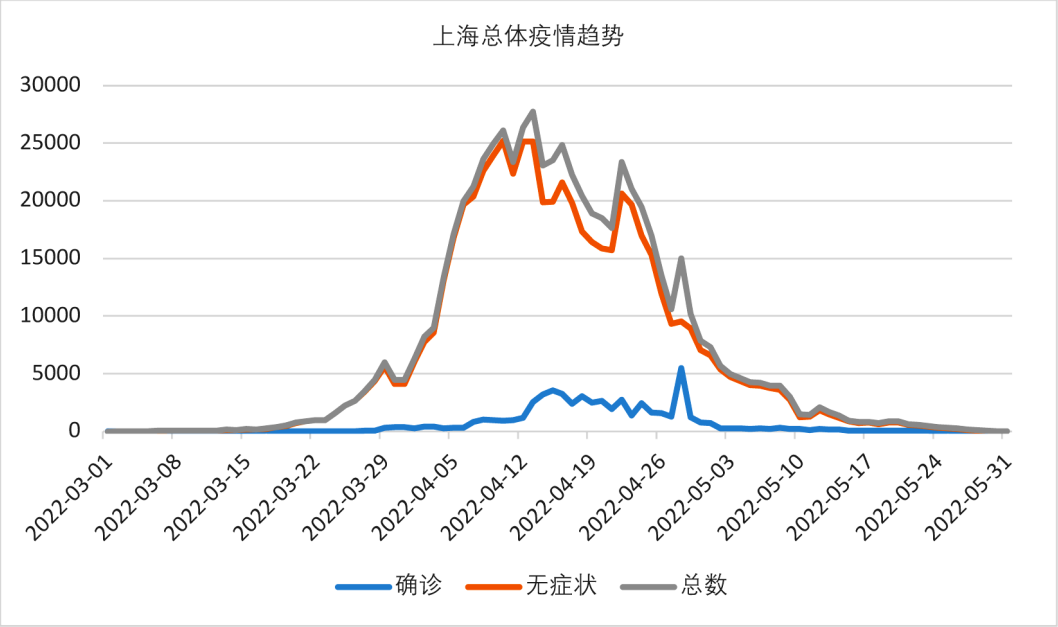

Figure 1. The overall trend of the epidemic in Shanghai
图 1. 上海疫情总趋势

2.2. 新浪微博数据总体分析

本次上海疫情 3、4、5 月份的微博发表数量如图 2 所示。整体来看，每日微博数起伏较大。3 月 5 日起开始递增，3 月 12 日有明显下降趋势，但 3 月 13 日发表的微博数是前一日的二倍之多，接下来这几天又开始减少，3 月 20 日起则在起伏中逐渐攀升，4 月 2 日达到顶点，有效日发微博高达 16,805 条。随后上升减少不断交替，4 月 23 日之后，总体是下降趋势，到 5 月 26 日，维持一般日常数据，日增仅 3000 条左右。

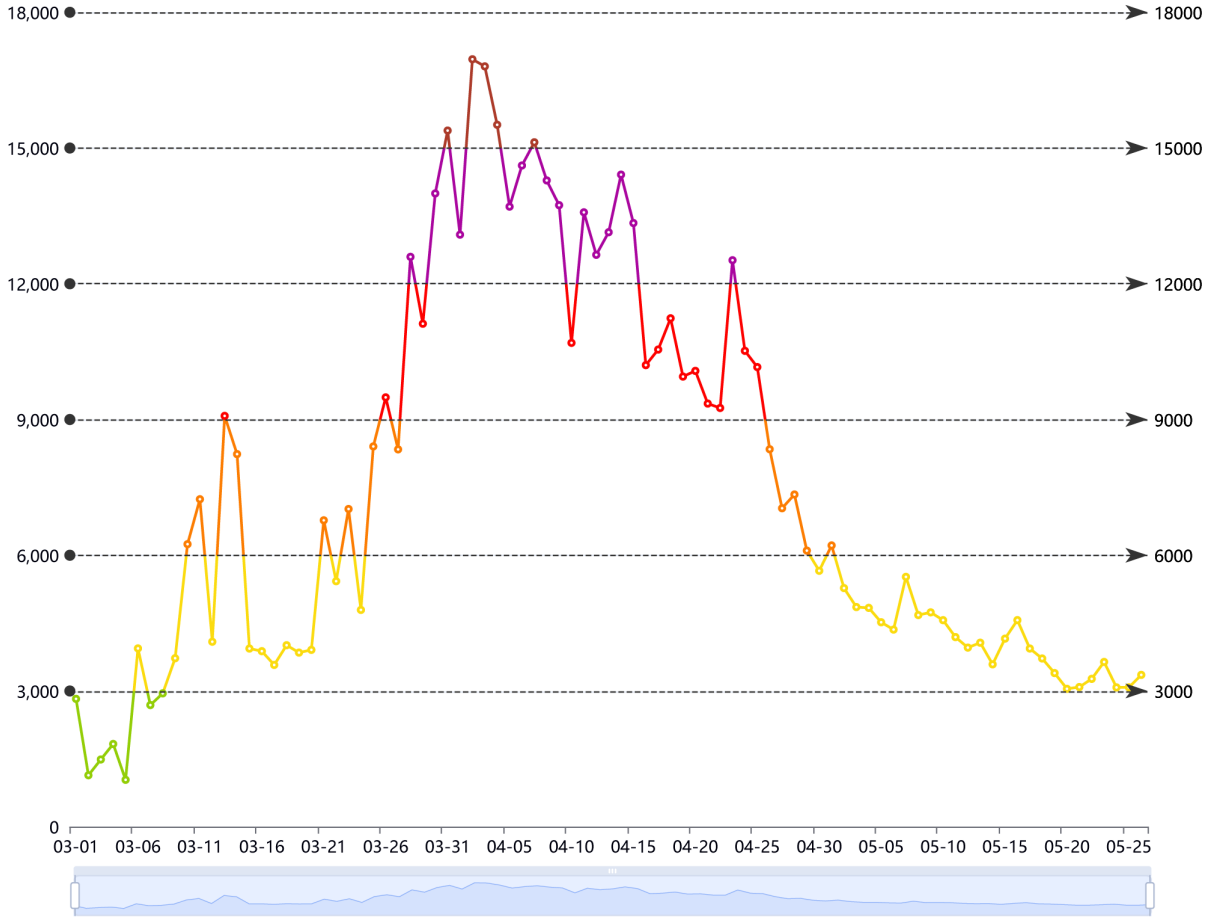

Figure 2. Daily number of Weibo during the epidemic in Shanghai
图 2. 上海疫情期间每日微博数

3. 研究方法

3.1. 数据来源

本论文的数据来自微博网页版客户端。新浪微博是基于用户关系的社交软件平台，用户可以通过文字、视频、图片等方式来传播信息并实现实时互动，具有互动性强、扩散度广等特征[13]。此次研究利用Python 的 Scrapy 分布式爬虫框架从微博获取上海疫情这一非常规突发事件的微博文本数据，选定时间为2022 年 3 月 1 日到 2022 年 5 月 29 日。采集的主要内容包括：用户名、微博内容、发表时间、点赞数、评论数以及转发数等，共爬取数据 652,840 条。经过数据预处理后，最终得到的有效数据共 83,254 条。

3.2. 词频统计方法

本论文在进行词频统计时，采用的是 TF-IDF 词频统计处理方法。TF-IDF (Term Frequency-Inverse Document Frequency)是一种统计方法，用来评估某个字或者词语对于整份文件的重要程度[14]。如果某个词语在一篇文章中经常出现，但是却很少出现在其他文章中，那么这个词语就是具有区分度的，一篇文章中出现的词语间的相互连接，表达出了该文章的核心思想，适合用作研究。

TF 是词频(Term Frequency)，是指一个词语在该文章中出现的频率。这个频率数字是对词语字数的归一化，避免它偏向过于长的文章。IDF 是逆向文件频率(Inverse Document Frequency)，某个特定词语的

IDF，可以用总文件数量除以包含这个词语的文件的数目，然后再将得到的结果取以 10 为底的对数得到，计算公式如下：

$$idf(t)=\log\left[(1+n)/(1+df(t))\right]+1$$

根据某一特定文件里的高词语频率，以及该词语在整个文章中的低文件频率，可以得出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

3.3. 情感分析方法

文本情感分析是指对富有感情色彩的具有主观性的文本进行分析、处理以及归纳的过程。传统的文本情感分析主要是根据机器学习和情感词典来处理，随着技术的不断发展，深度学习开始应用在了这项任务中[15]。

本文所采用的情感分析工具是 BERT 模型，BERT 是基于双向的 Transformer 编码器的语言预训练模型[16]。其中，Transformer 是 Ashish Vaswani 等在 Attention Is All You Need 中所提出的 sequence2sequence 模型，其主要任务是预先进行训练语言模型，然后再分配任务，进行分类、标记等工作[17]。为了弥补 RNN 运行慢等缺点，Transformer 使用自注意力(Self-Attention)机制来替代 RNN 的结构。当输入一个句子时，该句子中的每个词都与其他的词进行 Attention 计算，Attention 的计算公式如下：

$$Attn(Q,K,V)=Soft\max\left(\frac{QK}{\sqrt{d_k}}\right)V$$

其中，d_k 表示每个字的 query 和 key 向量的维度，SoftMax 是归一化指数函数。最终得到的 Attention 值是一个矩阵，矩阵的每一行代表输入句子中相应字的 Attention 向量，其中包含了这个句子中该词和其他位置的词的相互关系信息，是一个新的向量表示。

BERT 模型使用带有 Self-Attention 的双向 Transformer 模型获得了句子的前后语义关系，从而使句子的语义表达变得更好。在预训练过程中，BERT 模型还有所创新，即 MLM 任务(Masked Language Model)和 NSP 任务(Next Sentence Prediction) [18]。MLM 可以被理解为完形填空任务，程序会随机遮盖掉句子中 15%的词，然后预测被遮盖的位置是什么词汇，目的是让 BERT 模型能够实现深度的双向表示。例如，句子“本次疫情还处在上升阶段”中部分词被遮盖掉后变为“本次疫情还处在[MASK]阶段”。为了在微调阶段尽可能降低[MASK]标记带来的负面影响，模型采用的具体策略如下：

1) 80%的情况下采用[MASK]标记，即“本次疫情还处在上升阶段”，变为“本次疫情还处在[MASK]阶段”；

2) 10%的情况下采用一个随机词汇来替代被遮盖的词，即“本次疫情还处在上升阶段”可能变为“本次疫情还处在稳定阶段”；

3) 10%的情况下保持原句不变，即保持“本次疫情还处在上升阶段”不变。

NSP 任务是给定一些句子对(A，B)，其中 50%的数据中的 B 是 A 的下一句子，剩余 50%的数据中的 B 则是语料库中随机选择的。NSP 任务的目的是让模型能够对两个句子间的逻辑和因果关系进行理解。

Bert 模型拥有强大的语言表征能力和特征提取能力。在 11 项 NLP 基准测试任务中达到 SOT，再次证明了双向语言模型的能力更加强大。

4. 结果与分析

4.1. 博文特征分析

首先，根据疫情走势和微博数的变化来筛选三个关键节点进行分析，分别是 3 月 13 日、4 月 2 日和

5 月 16 日。其中“上海”、“隔离”、“疫情”、“防控”、“核酸”、“物资”等词高频出现。在三个不同的节点，“疫情”、“防控”等词都位居前列，可见整个疫情过程中，这类措施更引得民众关注。此次上海疫情整体数据微博词云图如图 3 示。

Figure 3. Weibo word cloud map
图 3. 微博词云图

本次选取的三个节点都属于统计过程中的关键节点，其中 3 月 13 日是整个疫情期间微博发表数的第一个高峰，4 月 2 日是整个统计过程中，微博数发的最多的一日，具有参考价值。而 5 月 16 日，上海疫情在经历暴增之后，累计确诊人数首次跌破千例，表明形式一片大好。具体数据如下表 1 所示。

Table 1. The distribution of “subject words” of netizens’ public opinion
表 1. 网民舆情“主题词”分布

日期	单词及所占比例
3 月 13 日	疫情(13.64%)、防控(6.96%)、新冠(6.41%)、病例(4.23%)、新闻(4.09%)、确诊(3.96%)、核酸(3.27%)、感染者(3.14%)
4 月 2 日	上海(26.32%)、防控(3.68%)、隔离(3.68%)、核酸(2.89%)、视频(2.37%)、防疫(2.37%)、医院(1.84%)、物资(1.84%)、小区(1.84%)
5 月 16 日	上海(20.53%)、疫情(17.66%)、防控(6.98%)、新闻(3.29%)、恢复(3.29%)、核酸(2.26%)、管控(1.85%)、隔离(1.85%)

4.2. 总体情感分析

本文使用 Hugging Face 官网提供的预训练检查点 BERT-Base。中文作为起始检查点，用于预训练过程。然后对模型进行微调，最后设计一个适合于本文数据的分类器，依据事实，将分类标签设为正面感情和负面感情两种。模型使用的数据集做训练时的参数如下表 2 所示。

实验表明，使用 BERT 预训练模型可以在非常短的时间内以一个非常快的速度就能够达到很高的正确率，在测试集上最终取得了 93.75%的情感分类准确率。将预处理后的微博数据放入模型进行分析，最终结果如图 4 所示，本文以负面感情变化为例。

Table 2. Parameter settings
表 2. 参数设置

参数	设定值/函数
Batch Size	16
Learning Rate	5e−4
Optimizer	AdamW
epoch	300
loss	Cross Entropy Loss

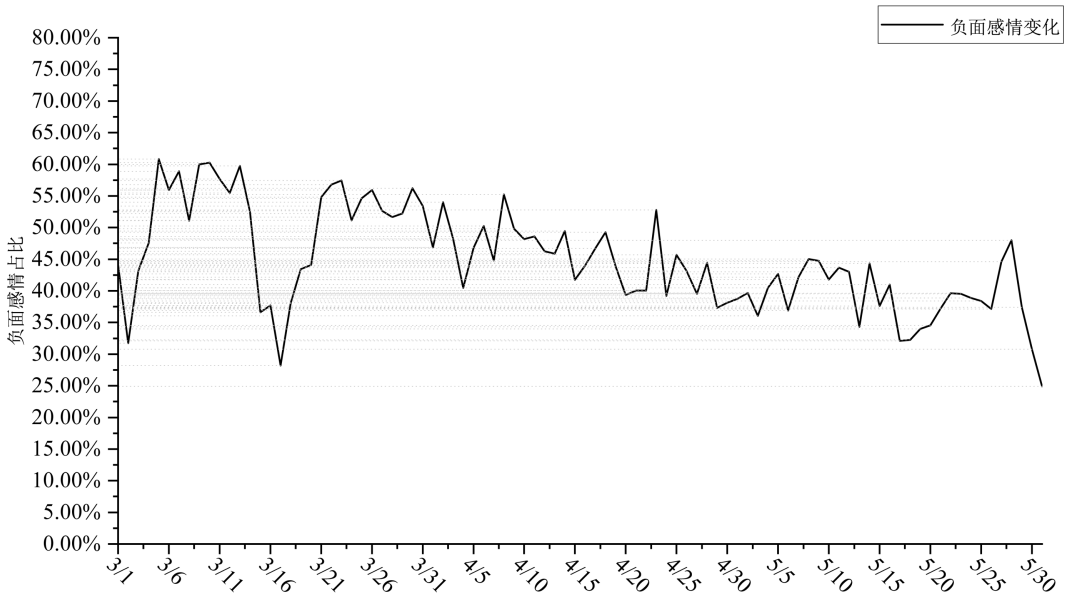

Figure 4. Sentiment analysis results
图 4. 情感分析结果

由图 4 可知，网民的负面情绪虽起伏不定，但总体比例较高。3 月 2 日至 3 月 6 日呈直线上升趋势，负面情绪的比例直逼 60%，也就是说大部分人都表达出了不满，其中，很大一部分原因就是民众对政策理解的不到位以及社区的后勤保障并没有跟上。此后十几天一直上下起伏，3 月 14 日有所下降，3 月 16 日随即上升，此后的时间内负面感情整体以较微弱的趋势下降，截至 5 月 26 日，已经降至 36%左右。整个过程中，民众负面情绪的峰值发生在疫情初期。

值得注意的是，负面感情第一次大幅度下降发生在 3 月 14 日左右，结合当天疫情的相关数据可以发现，在几日的连续攀升之后，3 月 13 日正好出现了短暂的下降趋势。结合之前的上海精准防控措施，广大网友一改前几天消极的情绪，微博的整体舆情消极情况锐减。例如某位网友的发表博文：“#上海六院疫情#还想去迪士尼，上海加油”。还有另一位网友发布：“#上海交大疫情#从来没遇到这么好的学校饭端在手上来一切要求都满足不接受任何负面消息我爱交大”。但此后疫情新增数据又有上升，负面感情便又开始回升。

5. 讨论

1) 在时间分布上，疫情发生后，整体的博文发表数量呈“n”字形曲线分布。疫情初期，偶尔出现

的几例新增并不能够引起人们的恐慌，关注度小。然而到中期疫情爆发时，由于一些居民被困在家里许久，且一些小区的后勤保障措施又不能及时到位等因素，不仅上海地区的用户发表的博文多，其他地区也都对这座城市高度关注。后期新增病例逐渐减少至千例甚至百例时，随着解封的步子越迈越大，相关用户的活跃度也降低，发表的博文明显减少。

2) 在对网络舆情变化进行分析时，能够发现其变化的趋势与疫情的变化趋势有一定的相关性。而疫情的变化趋势又体现在微博平台的热搜词条上。比如初期的“上海六院疫情”，主要集中在前一日的新增的地区，关注面窄。中期的热搜则聚焦在数字上，地区已经不能说明舆情走向，数字才最引起民众的关注，比如“上海新增 38 例本土确诊 2231 例无症状”、“上海新增本土确诊 311 例无症状 16766 例”等，此外，在这个时期，某些负面事情出现在热搜的次数也逐渐增多，也进一步加大了负面情感在整体的比重，比如“女子疫情抢购千元食物被大妈顺走”、“上海确诊病例自驾杭州来福士喝咖啡”等。在疫情末期，一切都逐渐向好，这时舆情的走向逐步迈向积极的方面，比如热搜词条“上海回来了”、“上海司机不加价送仨小伙去车站”、“上海本轮疫情的首个方舱医院休舱”等。

6. 结论

本文基于微博数据，以本次上海疫情为例，通过对 2022 年 3 月份、4 月份、5 月份的微博文本、发布时间、点赞数、评论数、转发数等信息进行提取和分析，研究此次疫情中网络舆情的时间分布特征；结合 BERT 模型，对本次疫情呈现出的舆论情感进行分析，寻找疫情中全国各地网民的情感演化差异，挖掘此次疫情的微博舆情与时间、实际疫情走向以及当前热点的相关关系，为新冠疫情舆情的引导和处置提供参考。

结果表明：首先在时间分布上，发微博的数量主要集中在疫情的高发期。此时，网络舆情信息传播量也较大，极易对网民产生影响。其次在舆情情感上，由于受到博文语境的大环境影响，负面感情的最高比例达到 60%，最低比例为 25%左右。最后从整体上看，在疫情高发阶段，网民情绪较偏激，而在疫情后期逐渐向好阶段，网民情绪则趋向积极稳定。本次疫情舆情初期表现出的多为负面感情；中期由于新增感染者数量的激增以及一些负面事件的爆出，负面感情的比例一直居高不下；末期由于疫情走势良好以及其他热点问题，负面感情的比例开始在波折中呈下降趋势。疫情以及舆情的逐渐向好离不开中央政府和地方工作人员在疫情舆情应对中的高水平发挥，其他地区医护人员的驰援以及蔬菜等物资的支援也使得上海人民对未来充满希望，“最美逆行者”等群体不仅唤起了群众的民族团结和凝聚力，也号召广大群众积极抗疫。

本文研究的网民情感波动较大，其稳定程度与政府疫情防控工作的稳定性密不可分，可以预测由于疫情发展的不确定性，在未来的一段时间内，网民的情感波动大概率还会出现反复，因此，在疫情常态化下，网络舆情监测和治理也将是一项长久的工作。

参考文献

[1] 邢云菲, 王晰巍, 韩雪雯, 张长亮. 基于信息熵的新媒体环境下网络节点影响力研究——以微信公众号为例[J]. 图书情报工作, 2018, 62(5): 76-86. https://doi.org/10.13266/j.issn.0252-3116.2018.05.009

[2] 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]. 情报杂志, 2020, 39(7): 67-74+158.

[3] 黄仕靖, 吴川徽, 袁勤俭, 夏镜然. 基于情感分析的突发公共卫生事件舆情时空演化差异研究[J]. 情报科学, 2022, 40(6), 149-159. https://doi.org/10.13833/j.issn.1007-7634.2022.06.019

[4] 陈娟, 刘燕平, 邓胜利. 政府辟谣信息的用户评论及其情感倾向的影响因素研究[J]. 情报科学, 2017, 35(12): 61-65+72. https://doi.org/10.13833/j.issn.1007-7634.2017.12.011

[5] 凌海彬, 缪裕青, 张万桢, 周明, 武继刚. 多特征融合的图文微博情感分析[J]. 计算机应用研究, 2020, 37(7): 1935-1939+1951. https://doi.org/10.19734/j.issn.1001-3695.2018.12.0929

[6] 胡思才, 孙界平, 琚生根, 王霞, 龙彬, 廖强. 基于扩展的情感词典和卡方模型的中文情感特征选择方法[J]. 四川大学学报(自然科学版), 2019, 56(1): 37-44.

[7] 薄涛, 李小军, 陈苏, 王玉婷, 祁国良. 基于社交媒体数据的地震烈度快速评估方法[J]. 地震工程与工程振动, 2018, 38(5): 206-215. https://doi.org/10.13197/j.eeev.2018.05.206.bot.024

[8] Zhao, Y. and Fan, B. (2018) Exploring Open Government Data Capacity of Government Agency: Based on the Resource-Based Theory. *Government Information Quarterly*, **35**, 1-12. https://doi.org/10.1016/j.giq.2018.01.002

[9] Mergel, I., Kleibrink, A. and Sörvik, J. (2018) Open Data Outcomes: US Cities between Product and Process Innovation. *Government Information Quarterly*, **35**, 622-632. https://doi.org/10.1016/j.giq.2018.09.004

[10] 李宗敏, 张琪, 杜鑫雨. 基于辟谣微博的互动及热门评论情感倾向的辟谣效果研究——以新冠疫情相关辟谣微博为例[J]. 情报杂志, 2020, 39(11): 90-95+110.

[11] Li, S., Wang, Y., Xue, J., Zhao, N. and Zhu, T. (2020) The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International Journal of Environmental Research and Public Health*, **17**, 2032. https://doi.org/10.3390/ijerph17062032

[12] Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., Dai, J., *et al*. (2020) Mental Health Problems and Social Media Exposure during COVID-19 Outbreak. *PLOS ONE*, **15**, e0231924. https://doi.org/10.1371/journal.pone.0231924

[13] 齐珉, 齐文华, 苏桂武. 基于新浪微博的 2017 年四川九寨沟 7.0 级地震舆情感分析[J]. 华北地震科学, 2020, 38(1): 57-63.

[14] 张娜, 张琨, 张先国, 张佳慧, 蒋彤彤, 方悦. 基于主题词与信息熵编码的文本零水印算法[J]. 计算机与数字工程, 2021, 49(8): 1612-1618.

[15] Lin, B., Zampetti, F., Bavota, G., Di Penta, M., Lanza, M. and Oliveto, R. (2018) Sentiment Analysis for Software Engineering: How Far Can We Go? *Proceedings of the* 40*th International Conference on Software Engineering*, Gothenburg, 27 May-3 June 2018, 94-104. https://doi.org/10.1145/3180155.3180195

[16] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30: *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 5998-6008.

[18] Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.Y. (2020) Mpnet: Masked and Permuted Pre-Training for Language Understanding. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 16857-16867.