新型冠状病毒肺炎相关的推特主题与情感研究

梁帅龙
新加坡科技设计大学
新加坡

黄辉
澳门大学计算机与资讯科学系
澳门

张岳
西湖大学工学院
杭州

shuailong_liang@mymail.sutd.edu.sg
derekfw@um.edu.mo
zhangyue@westlake.edu.cn

摘要

我们基于从2020年1月22日至2020年4月30日在推特社交平台上抓取的不同国家和地区发布的50万条推文，研究了有关2019新型冠状病毒肺炎相关的主题和人们的观点，发现了不同国家之间推特用户的普遍关切和看法之间存在着异同，并且对不同议题的情感态度也有所不同。我们发现大部分推文中包含了强烈的情感，其中表达爱与支持的推文比较普遍。总体来看，人们的情感随着时间的推移逐渐正向增强。

关键词： 主题模型；社交媒体分析；新冠肺炎；

Exploring COVID-19-related Twitter Topic Dynamics across Countries

Shuailong Liang[1], Derek F. Wong[2], Yue Zhang[3]
[1]Singapore University of Technology and Design
[2]Department of Computer and Information Science, University of Macau
[3]School of Engineering, Westlake University
shuailong_liang@mymail.sutd.edu.sg
derekfw@um.edu.mo
zhangyue@westlake.edu.cn

Abstract

We investigate the topics and sentiment concerning COVID-19 from half-a-million tweets across different countries between January 22 and April 30, 2020, finding similarities and differences between the general concerns and feelings between countries, and varying attitudes towards different issues. Strong sentiments are found in the tweets, yet love and support also gain much popularity. In general, a trend of increasingly positive sentiment was observed.

Keywords: Topic Model , Social Media Analysis , COVID-19

1 引言

2019年新型冠状病毒肺炎（新冠肺炎）是由2019年新发现的冠状病毒所引起的传染病。由于高死亡率和高传染性，它在全世界已经造成了近千万例感染和近五十万例死亡，213个国家受到影响（截至2020年6月27日）。许多国家都采取了严格的措施来防止这种疾病的传播，包括封锁、居家隔离、社交隔离和旅行禁令等。新冠病毒对世界的经济和政治产生了前所未有的影响。了解人们如何看待和响应政府的政策，他们的关注点、态度和观点以及他们的健康信息需

第十九届中国计算语言学大会论文集，第707页-第718页，海南，中国，2020年10月30日至11月1日。

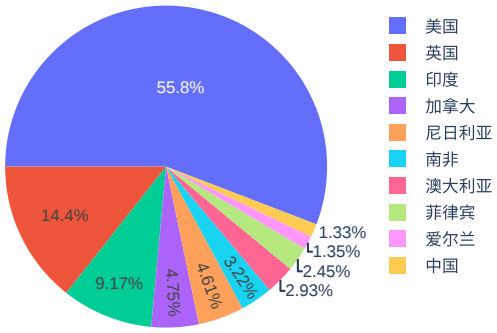

Figure 1: 饼图显示了推文数前十国家推文的比例

国家	推文数
美国	235122
英国	60434
印度	38591
加拿大	20008
尼日利亚	19424
南非	13574
澳大利亚	12348
菲律宾	10310
爱尔兰	5668
中国	5589

Table 1: 推文总数前十的国家的推文数

求和健康寻求行为是至关重要的。从社交媒体中获取的信息甚至可以帮助预测疾病的传播和增长(Turiel and Aste, 2020; Liu et al., 2020; Skiera et al., 2020)。

推特是一个流行的社交平台，有3.3亿月活跃用户，每天有5亿条推文被发出，并且有92.23%的联合国会员国的公民拥有推特账号（截至2019年第一季度）(Aslam, 2020)。与新闻相比，推特是一种更加动态和民主的信息来源，已被广泛用于社交媒体和自然语言处理研究。例如，Fung et al. (2014)和Lancet (2014)使用推特数据研究了埃博拉疫情。与官方新闻文章相比，推特在了解危机的真实叙述方面更具基层性和动态性。我们使用隐含狄利克雷分布（LDA）(Blei et al., 2003)来分析Chen et al. (2020)使用推特应用程序接口收集的从2020-01-22至2020-04-30期间与新冠肺炎相关的推文的主题。我们针对推文数量最多的前十个国家进行分析，分析了几个具有代表性的主题以及人们对这些主题的情感及其动态演变过程。

我们发现在情绪和情感的话题方面，包含了情绪的两个极端：仇恨言论和爱与支持言论。总的来看，人们的情绪是稍微偏积极的，并且随着时间的变化，情绪的积极程度也逐渐提高。不同国家和地区的推文所反映的情绪也各不相同：中国，爱尔兰和印度的情绪相对积极，而美国，澳大利亚和南非的情绪相对消极。人们对于特定主题的情绪，例如**隔离生活**和**病毒起源**，也随着时间而改变。我们还发现新冠肺炎对美国的选举和政治生活有着较为负面的影响。

2 数据集与实验设置

我们使用**COVID-19-TweetIDs**数据集(Chen et al., 2020)获取与COVID-19相关的所有推文ID。截至2020年6月5日，推文总数为1.44亿条，涵盖多种语言，其中英语占60%以上。我们使用Twarc [0]工具从推特应用程序接口获取推文全文（此过程称为hydration）。我们成功收集了从1月22日到4月30日（共100天）总计115,010,623条推文，平均每日超过100万条。

获取全文以后，我们通过将推文对象的*lang*字段限制为*en*来提取英语推文，并仅保留带有*place*属性的推文以包含位置信息。一共获得了498,852条推文。 [1]要研究各个国家的推文主题，我们选择数据集中推文数量最多的前十个国家。表1和图1中显示了按国家和地区分类的推文总数。图2中显示了前十个国家的每日推文计数。在所有国家中，与新冠肺炎相关的具有地理位置信息的推文中，美国的推文数量最多（每日超过2000条），其次是英国和印度。

预处理 过滤出包含地理位置信息的推文后，我们通过删除所有链接和用户提及（@）来对推文进行预处理，然后将主题标记符号#删除。我们使用*emoji*库将表情符号转换为标准文本（例如，将大拇指表情符号转换为*thumbs_up*）。最后，我们将所有文本转换为小写。图3中显示了完整的数据处理流程图。对于每个经过预处理的推文，我们使用NLTK工具包(Bird and Loper, 2004)将其分词，并删除数字，停用词和太短的词（少于三个字符）。此外，我们还删除了直接描述新冠肺炎关键词，例如*COVID19*，*cronavirus*，*COVID*，因为这些词无法提供有关新冠肺炎本身的更多信息。之后，我们去掉了文档频率少于5的词和出现在50%以上的推文中的词。

[0]https://github.com/DocNow/twarc

[1]还有其他可用的COVID19 Tweet数据集，例如 Qazi et al. (2020)，它们可以在基于地名词典的方法中推断地理信息。我们将其留给以后的工作。

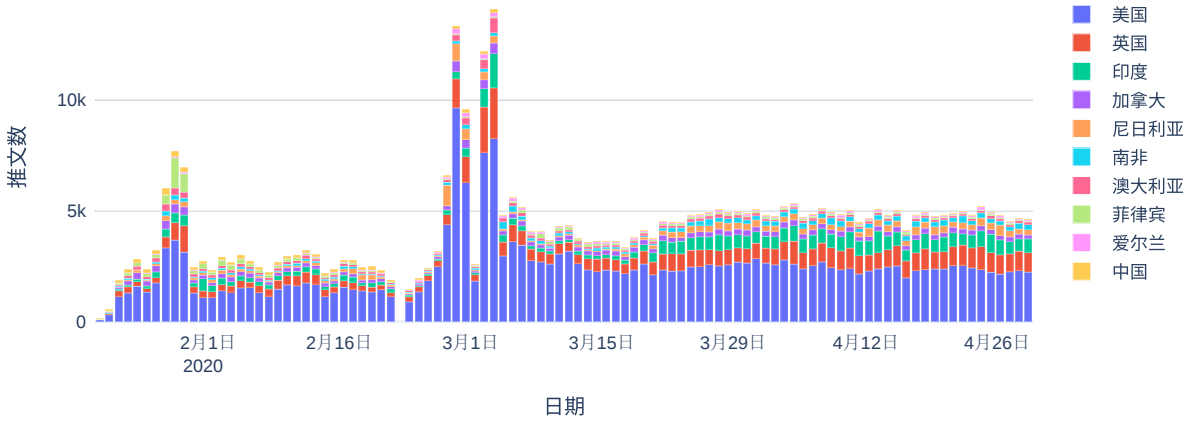

Figure 2: 推文总数前十的国家的每日推文计数。推文最多的国家位于最下方，而推文最少的国家位于最上方。

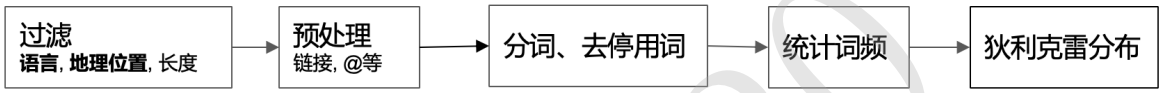

Figure 3: 推文处理流程图

主题模型 我们使用隐含狄利克雷(Blei et al., 2003)模型为每条推文计算话题分布。隐含狄利克雷模型将文档视为词袋（单词顺序无关紧要），它的概率模型的生成过程如下：对于每个文档，首先生成一组主题，然后每个主题选择自己的一组单词。主题的概率分布提供了文档的明确表示。参数估计通常基于变分方法和Gibbs采样。我们使用Gensim(Řehůřek and Sojka, 2010)的MALLET LDA框架(McCallum, 2002)[2]在所有推文上训练主题模型。MALLET框架使用了Gibbs采样的快速且高度可扩展的实现，并提供了非常有效的方法来进行文档主题超参数优化，并提供了工具来根据经过训练的模型推断主题。

为了找到最佳主题数目，我们将主题数目分别设置为$5, 10, 15, 20, 25, 30, 50, 100, 150, 200$，并计算每个主题模型的连贯性得分$C_V$(Röder et al., 2015)。迭代次数设置为2000。我们用每个主题的前20个关键词计算模型的主题连贯性得分，如表2所示。从表2中我们发现主题数为50的模型具有最高的主题连贯性得分。此后增加主题数将逐渐降低连贯性评分。当主题数为20时，相关度得分0.514略低于最高分数0.547，因此在分析中，为了方便研究，我们选择主题数20，没有牺牲太多模型质量。

情感分析 为了调查人们对特定主题的情绪及其随着时间的发展，我们使用了TextBlob [3]提供的基于词典的情感分析工具。对于每条推文，TextBlob返回的情感极性得分在-1和1之间，其中-1代表最负面的情绪，而1代表最正面的情绪。我们计算每日特定主题的推文的情感得分均值。

3 结果与分析

在本节中，我们首先介绍主要主题建模结果，包括提取的主题及其形成关键词。然后，我们讨论随着时间的推移总体和特定国家/地区主题的发展。最后，我们针对特定主题进行情感分析，并比较各国之间推特用户对某个主题的情感。

3.1 话题关键词

表3中显示了为这20个主题提取的关键词。对于每个主题，我们对主题关键词进行汇总，然后发现每个关键词簇都反映了特定主题。**主题名称**列是我们的归纳总结出来的。为了弄清楚每

[2]http://mallet.cs.umass.edu/topics.php
[3]https://textblob.readthedocs.io/en/dev/

话题数量	5	10	15	20	25	30	50	100	150	200
主题连贯性(C_V)	0.350	0.427	0.468	*0.514*	0.522	0.534	**0.547**	0.543	0.528	0.505

Table 2: 具有不同主题数的隐含狄利克雷主题模型的主题连贯性得分。

#	主题名称	主题关键词
0	情绪情感	good thing feel thought bad lot doesn change make sad hear idea hope head guy wow
1	隔离生活	stayhome quarantine eye socialdistancing day fire morning staysafe quarantinelife
2	仇恨言论	shit fuck man gonna lol guy fucking game damn real play joke catch yeah beer ain
3	投票与政治	house woman stupid party man white ppl won hate wrong left racist black power vote
4	特朗普与美国反应	trump president american cdc pandemic response leader lie hoax democrat called
5	封城日子	lockdown day week today month end state hour ago april start coming nigeria lock
6	爱与支持	friend live video lockdown watch night love family today tonight share watching movie
7	保健与医疗	health hospital care test patient public positive doctor medical worker testing tested
8	必需品供给	mask food open face free panic order buy place essential run local store street wear
9	工作	time pandemic work life great working long job hard year good making find start lost
10	爆发与旅行禁令	china country travel flight italy outbreak south japan korea australia iran ship canada
11	信息与媒体	news read medium fear fact question true information story real tweet check article
12	印度反应	india government lockdown fight govt sir situation nation action minister support
13	病毒来源	china wuhan outbreak city control war government problem usa animal bat lab threat
14	流行病与疫苗	people flu year epidemic disease die human vaccine kill million died cure sick outbreak
15	经济危机	business money global market economy pay company crisis big deal stock impact cut
16	祈祷	home stay safe god folded_hands love family hope happy save pray healthy microbe
17	预防与保护措施	social distancing hand spread stop wash measure place avoid protect prevent water
18	学校与学生	school community student team plan kid support child online event group class join
19	情况报告	case death number confirmed update person report rate state infection infected total

Table 3: 每个主题的关键词。由于空间有限，我们仅显示部分关键词。

个主题的主要程度，我们找到每个推文的主要主题。总主题分布显示在图4中。在这里，我们按贡献的多少的顺序讨论这些主题。

仇恨言论 如表3中所示，该主题由关键词*shit*，*fuck*，*damn*，*joke*，*play*等定义，并且包含强烈的个人负面情绪。此主题在大多数推文中表现显著。推特用户使用大写字母表达自己的强烈感情。例如：

> *I'm just convinced that we all gonna die anyways SINCE EVERYBODY AND THEY MOMMAS DONT KNOW HOW TO STAY THE FUCK HOME AND STOP HANGING OUT WITH THEIR FRIENDS. (*我只是相信，我们所有人都会死去，因为每个人和他们的妈妈都不知道如何待在他妈的家中并停止与他们的朋友们闲逛。*)*

上面的示例显示了推文作者对那些不执行居家隔离不停止社交聚会的人们的强烈消极情绪。另一个例子是

> *FUCK COVID-19 ITS PISSING ME OFF FIRST NIGGAS DONT WANNA VOTE CUZ "boo hoo I don't wanna get sick" NOW THEY'RE PLAYING WITH MY DAMN MONEY FUCK 2020 SO FAR MAN THIS SHIT IS RIDICULOUS* （新冠肺炎真是气死我了，黑人不想投票因为"我可不想得病"，现在他们在挥霍我的钱，到现在为止*2020*真是太荒谬了！）

显示了新冠肺炎对人们投票行为的影响。

由于隔离政策和城市关闭，许多人的生活变得不稳定或难以预测。因此，人们表达了强烈的消极情绪。这种现象与之前关于推特仇恨言论研究的一些文献是一致的(Kshirsagar et al., 2018; Hillard, 2018; Drum, 2017).

情绪情感 这个主题与仇恨言论类似。这是一个通用主题，包括诸如*feel*，*think*，*hear*和*hope*之类的关键词。该主题下的推文是关于人们分享其情感，思想和观点的。例如：

> *The best way to not feel hopeless is to get up and do something. Don't wait for good things to happen to you. If you go out and make some good things happen, you will fill the world with hope, you will fill yourself with hope.* （不感到绝望的最好办法就是起来做一些事情。不要等待美好的事情发生在你身上。如果您出去做一些美好的事情，您将使世界充满希望，您也将充满希望。）

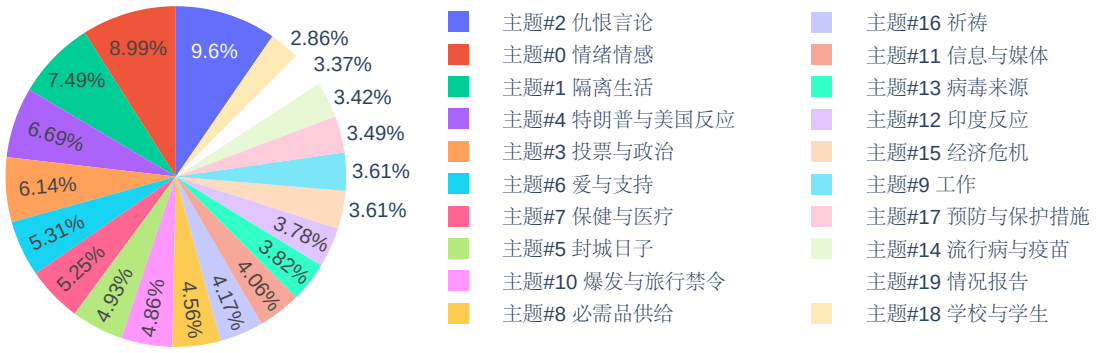

Figure 4: 所有新冠肺炎相关推文的主题分布，按每个主题所对应的推文数量进行排序，并按逆时针顺序显示。

隔离生活/封城日子 人们谈论了居家和社交距离，以及如何度过沉闷的居家时光。我们可以通过*stayhome*，*staysafe*，*quarantinelife*和*socialdistancing*等关键词来发现这个主题。人们焦急地等待封锁的结束。

> *From Thursday, 23 January 2020 to Tuesday, 14 April 2020, there are 82 days That's 2 months and 22 days*（从*2020*年*1*月*23*日星期四至*2020*年*4*月*14*日星期二，一共共*82*天，也是*2*个月*22*天）
>
> *Today was supposed to be the end of this bloody lockdown*（今天本应该是这场该死的封锁结束的日子）

特朗普与美国反应 有关美国总统唐纳德·特朗普和美国对大流行病的回应的讨论占总推文的6.7%，使其成为第四大热门话题。该主题由关键词*leader*，*lie*，*hoax*，*blame*和*truth*标识。我们随机选择几个示例，如下所示。

> *#DonaldTrump Dangerously Suggest Injecting Disinfectant As Treatment For #COVID19...*（特朗普危险地建议将注射消毒剂作为新冠肺炎的治疗方法...）
>
> *Trump failed America by dismantling the CDC pandemic response group in 2018, ignoring warnings, and claiming it was a hoax. The deranged imbecile Trump wasted precious time and lives.*（特朗普在*2018*年撤掉了疾病预防控制中心流行病应对小组，并无视警告，声称这是一个骗局。他辜负了美国。愚蠢的特朗普浪费了宝贵的时间和生命。）

不可避免地，美国民主党和共和党之间进行了大量的政治辩论和指责。我们将在第 3.4部分对美国的情况进行详细分析。

投票与政治 该主题的主要关键词包括*house*，*party*，*black*，*white*，*left*，*power*和*won*。这种流行病显然对2020年美国总统大选产生了根本影响。例如：

> *1. Pandemic emerges that disproportionately kills elderly 2. Masses of elderly congregate at polls during pandemic outbreak to vote for Joe Biden 3. Joe wins primary 4. Elderly voters are culled by pandemic 5. Trump wins in historic landslide*（*1.*新冠大流行导致更多的老年人丧生*2.*在大流行爆发期间的在民意调查显示大量的年长者倾向于投票支持乔·拜登（*Joe Biden*）*3.*乔赢得初选*4.*娇年长的选民在新冠大流行中丧生*5.*特朗普在历史性滑坡中获胜）
>
> *It was "unacceptable to hold an election ... in which people were forced to choose between their safety and voting," wrote new Democratic justice-elect.*（新当选的民主党大法官写道：*"*在选举中人们被迫在安全和投票之间作出选择，这样的选举是不可接受的。*"*）

爱与支持 人们向遭受大流行的国家表示支持，呼吁人类团结在一起度过难关，体现了国际主义的精神。

> *We stand by Italy during these trying times. Share your Support for our Italian friends, They are our colleagues, friends and family. Cari amici, siamo con voi.*（在这些艰难的时刻，我们支持意大利。分享你对我们的意大利朋友们的支持，他们是我们的同事，朋友和家人。亲爱的朋友，我们与您同在。）
>
> *I'm in Sanya, cheering Wuhan on. Hainan rice noodles is also cheering on Wuhan hot-dry noodles!!*（我在三亚，为武汉加油。海南米粉也为武汉热干粉加油！）

保健与医疗 在这场突发公共卫生事件中，人们倾向于讨论自己国家的卫生保健系统，批评其弊端，或者担心医院床位有限，无法适应迅速增加的新冠肺炎感染人数，并关注医务人员的必要保护设备。这个话题的关键词包括*test*，*positive*，*medical*，*worker* 等。

Doctors who have not been provided with Personal Protective Equipments have said that treating patients without the protective gear and masks is akin to a suicide mission. （没有个人防护设备的医生表示，没有防护装备和口罩为患者治疗无异于自杀。）

Uganda Virus Research Institute has the necessary equipment &reagents to test &confirm any suspected COVID-19 sample in country. So far samples from 10 persons who presented with signs & symptoms similar to that of COVID-19 have been tested. All tested negative. （乌干达病毒研究所拥有必要的设备和试剂来测试和确认该国任何可疑的新冠病例。到目前为止，已经测试了*10*名出现类似新冠症状的人的样品。所有检测均为阴性。）

爆发与旅行禁令 此主题的关键词列表中有多个国家/地区名称：中国、意大利、韩国、日本、澳大利亚、伊朗和加拿大。这些国家是最早爆发的国家之一，并在3月初被列为高风险国家。这些国家的公民被拒绝进入许多国家。

British Airways to cancel some Italy, Singapore, South Korea flights: 56 roundtrip flights from Heathrow and Gatwick airports to several destinations in Italy, including Milan, Bologna, Venice and Turin, between 14 - 28 March. (BBC) 英国航空公司取消了意大利，新加坡和韩国的部分航班：*3*月*14*日至*28*日之间，从希思罗机场和盖特威克机场到意大利的多个目的地（包括米兰，博洛尼亚，威尼斯和都灵）的*56*次往返航班。（英国广播公司）

基本服务 对于佩戴口罩是否可以降低新冠感染的风险这个话题引起了大量讨论。由于经济和供应链受到严重影响，食品和卫生产品等必需品供应也受到关注。

Can we spend a moment on stopping the panic buying/hoarding over the coronavirus? Warehouse clubs and stores are being emptied of masks, cleaning products, toilet paper, bread, and eggs in massive amounts for no reason at all. You don't need six months of canned goods either. （我们可以花点时间停止针对冠状病毒的恐慌购买和屯积吗？仓库和商店里大量的口罩，清洁用品，卫生纸，面包和鸡蛋被无故抢购一空。您也不需要六个月的罐头食品。）

FACE MASKS ARE NOT EFFECTIVE PROTECTION AGAINST THE CORONA VIRUS Face masks are effective for preventing spread when worn by those infected, and not by healthy people. Even that, the specialised face masks, N95 respirator, is what's recommended, not a regular surgical mask （口罩不是针对冠状病毒的有效防护。口罩可有效防止被感染者而非健康人群佩戴时扩散。即便如此，还是建议使用专用面罩*N95*呼吸器，而不是常规的手术口罩）

其他主题 由于篇幅所限，我们在此处简要讨论其余主题。主题16 祈祷与主题6 爱与支持较为接近。主题11与新闻、故事和事实的讨论有关，而主题19情况报告是关于确诊病例和死亡人数的每日更新。人们还谈论传染病带来的经济危机（主题15）:大流行导致许多小企业倒闭，许多国家的国内生产总值预期为负。如主题9所示，许多人失业，一些人不得不远程工作。学校和大学关闭，在主题18中得到了讨论。此外，还讨论了新冠肺炎的起源（主题13）和流行病学，包括死亡率、疫苗（主题17）以及预防和保护措施（主题14）。最后，除了美国以外，还有其他特定国家/地区的主题，例如主题12印度反应。

总之，从表3中，我们可以观察到推特上讨论的主要话题，这些主题反映了以下事实：新冠肺炎导致世界上许多国家封锁并要求人们呆在家里以隔绝病毒。结果，商业和经济受到严重影响，人们在家工作，许多人失业。人们将不可避免地表达他们对病毒以及对政府政策的情感和态度，包括责骂和仇恨言论。另一方面，他们也向朋友、家人和医务人员表示爱与支持。人们通过转发新闻、故事和事实（包括假新闻）来获取信息。关于病毒起源的阴谋论也大规模传播。人们也对寻找疫苗、保护自己和控制病毒的措施的有效性感兴趣。在所有国家中，中国作为疫情的首次爆发源地受到了广泛关注。美国拥有最多的推文用户和推文，并且确诊人数也迅速增加。因此，有关唐纳德·特朗普及其政府的讨论也有很多。随着印度确诊病例的增加，印度推特用户还讨论了很多有关印度局势及其政府政策。

3.2 总体话题流

图5中显示了20个主题的推文数量。第一个高峰出现在1月26日（图5点**A**，4千条推文）和1月30日和31日（图5点**B**，9千条推文）。最高峰大约在3月初（图5点**C**，1万6千条推文）。1月26日，中国国务院总理李克强领导一个预防和控制大流行的领导小组，然后决定延长中国春节假期，以遏制大流行(Xinhua, 2020)。最主要的主题是主题13病毒起源，主题2仇恨言论和主题19 情况报告。由于这是最初的爆发点，人们开始谈论病毒的起源，并密切关注病毒的发展。

1月30日，世界卫生组织宣布冠状病毒爆发为国际关注的突发公共卫生事件（PHEIC），并敦促所有国家为遏制疫情做好准备(WHO, 2020)。美国也宣布了国家进入“公共卫生紧急状

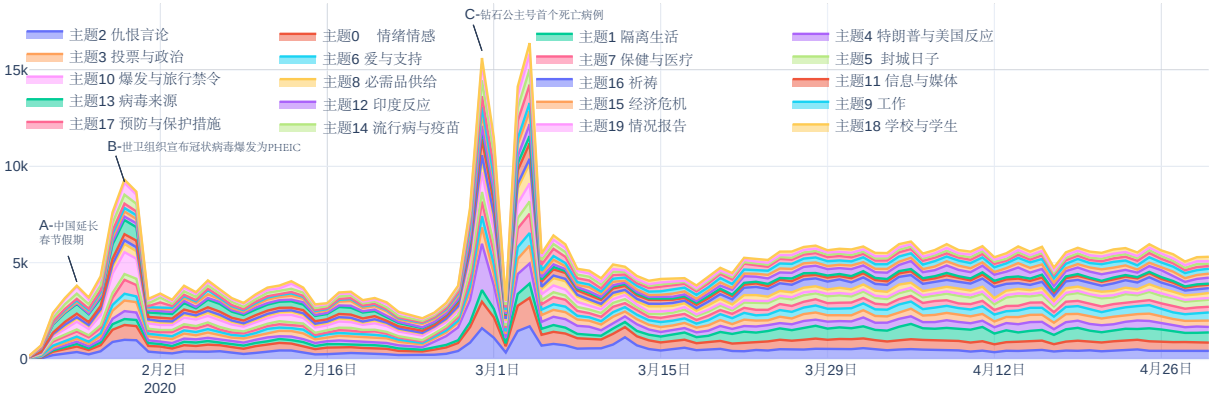

Figure 5: 20个主题所对应的推文的数量随时间的变化趋势。A，B和C是三个峰值。顺序（从下到上）与图4相同。

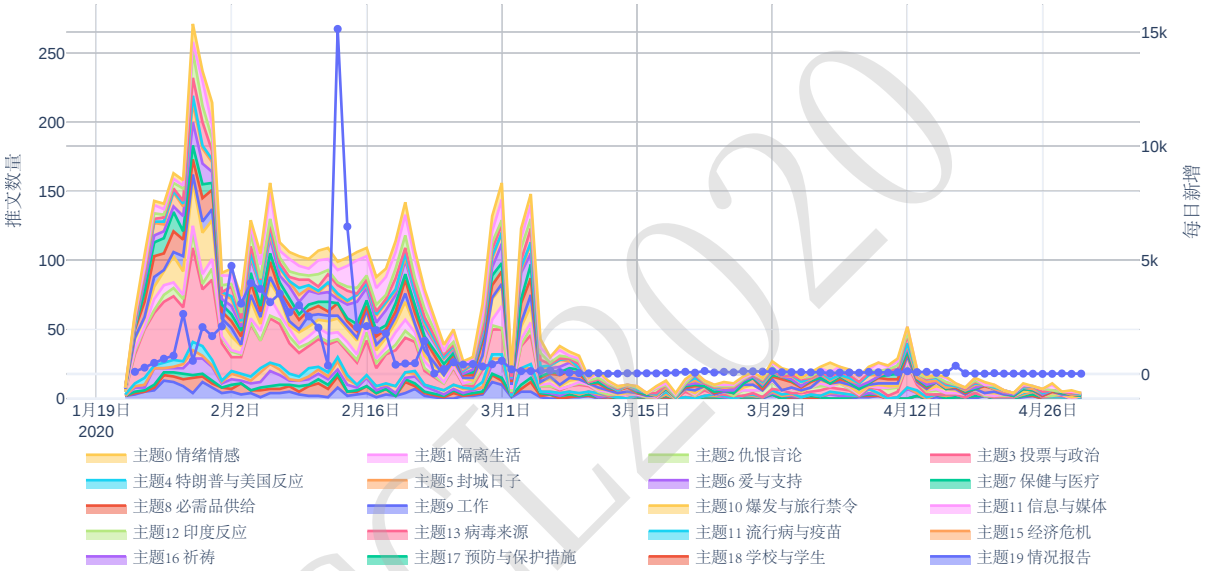

Figure 6: 中国推文话题流与每日新增确诊病例。

态”，两家美国航空公司宣布，在大流行期间取消往返中国的所有航班。1月30日最主要的话题是主题10爆发与旅行禁令，主题2 仇恨言论和主题0情绪情感感。在这个时间点，已经发生了几次爆发，许多国家禁止主要爆发国家的旅行者。人民对此表现出消极情绪。

从2月29日到3月4日，与新冠肺炎相关的推文总数达到了一万六千条，是我们整个研究时期中最高的。钻石公主号发生第一例死亡病例(Martin and Henriques-Gomes, 2020)，美国发生第一次死亡病例(Baker and Crowley, 2020)。美国扩大旅行限制到伊朗、意大利和韩国。最显著的主题是主题4特朗普与美国回应和主题2仇恨言论，这些主题和上述新闻事件紧密相关。

3月中旬之后，与新冠肺炎相关的总推文稳定地增加到将近六千，并且一直保持到研究阶段结束。在此期间，大流行逐渐蔓延到了欧美，确诊的美国病例开始超过中国(Dong et al., 2020)。

3.3 中国话题流与每日新增确诊病例

图6显示了中国各个话题的推文数量的动态变化和每日新增确诊病例数量的联系。由于中国国家防火墙的原因，以及本次研究只考虑英文推特，包含中国位置信息的英文推文数量总体较少。但是从趋势上看，推文数量的变化表现出和新增确诊病例的较强关联，尤其表现在3月1日以后，中国每日新增确诊病例归零之后，相关的推文也逐渐减少。值得注意的是，从1月下旬开始，话题13病毒起源一直占据讨论的话题中心，直到3月初。这也与中国是病毒的最初爆发国有

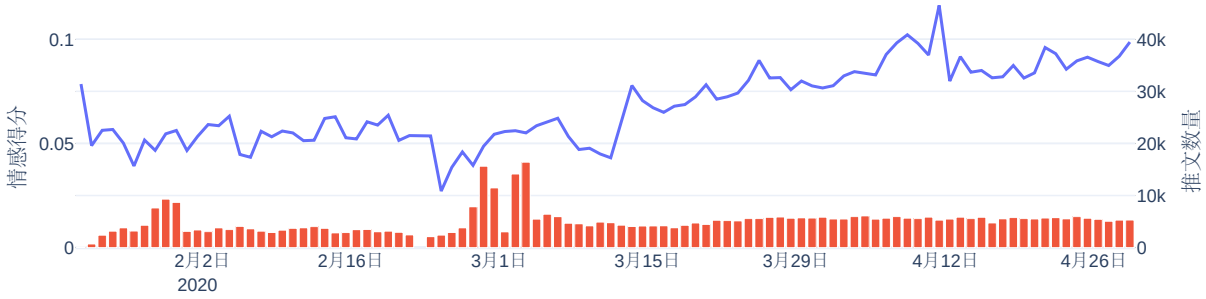

Figure 7: 整体情绪评分。折线显示情感评分，条形图显示推文数量。

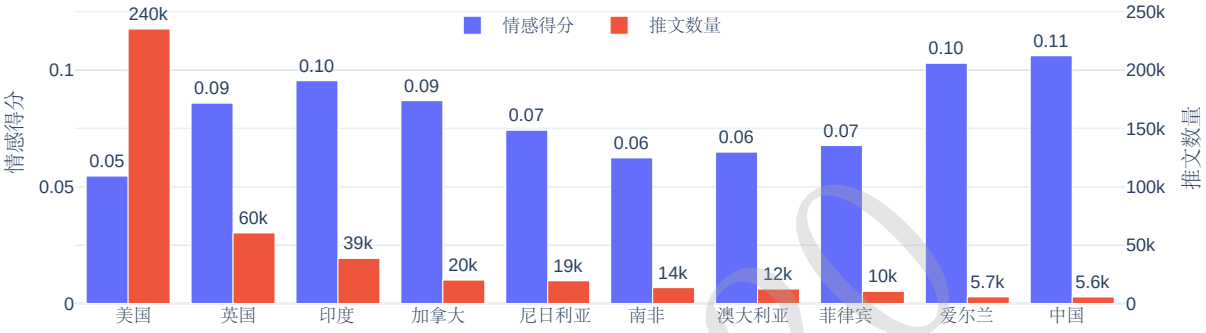

Figure 8: 推文总数前十的国家的综合情感评分和推文数量。左侧的条形图（蓝色）是该国家所有推文的平均情感得分，右侧的条形图（红色）是该国家的推文数量。

关。

3.4 特定国家的主题情绪变化

我们首先讨论与新冠肺炎相关的所有推文的情绪变化，不仅限于特定的国家或主题，然后针对某个主题的情感对特定国家进行分析。

总体情绪发展 图7显示了整体情绪的发展。我们可以看到，总体情绪略微乐观，并在2月底左右达到最低值，此时推文的数量达到了最大值。在最低点之后，情绪值逐渐增加。总的来看，人们对新冠肺炎的讨论持稍微积极的态度，并且随着时间的流逝，这些积极的态度也在增强。

图8中显示了推文数量前十的国家的综合情感评分。美国的推文数量最多，情感评分最低。爱尔兰和中国的推文绝对数量虽然不多，但是情感评分最高。印度，加拿大和英国的情感得分也比较高。

特定主题和国家的情感发展 我们选择了美国，英国，印度和加拿大作为研究对象，选择了主题0情绪情感、主题1隔离生活和主题13病毒起源进行了详细分析，如图9所示。在整个研究时期，这四个国家的推文数量最多。从图9中我们可以看出，这四个国家的情绪总体上是积极的，印度和加拿大的波动要比美国和英国的波动大。对于主题1隔离生活，在3月之前，情绪波动很大，随后逐渐稳定并转为积极状态。在上半年，印度和加拿大的波动最大。有趣的是主题13病毒起源的情绪发展具有相反的趋势：在下半年，情绪波动变得更加强烈，这一趋势在美国和加拿大尤为明显。

对美国特定主题的情感分析 我们选择了主题0情绪情感、主题1隔离生活、主题3投票与政治、主题4特朗普与美国反应、主题6爱与支持和主题13病毒来源，并在图10中说明了人民对各个主题的情感发展动态变化过程。对于主题0情绪情感，情感极性略微正面。对于主题1隔离生活，情绪最低值发生在3月初，然后逐渐上升，表明人民对城市封锁和隔离生活的逐步接受和适应。主题3投票与政治的情绪大多是负面的，表明大流行对美国大选的巨大影响，以及政治领域的指责与博弈。在关于主题13的讨论中，推文的数量在3月初达到顶峰，当时世卫组织宣布新冠肺炎为国际关注的突发卫生事件。人们的情绪在整个时期内都在波动。毫不奇怪，主题6爱与支持的情绪非常积极，并且呈上升趋势。最后，关于主题13病毒起源的推文情绪在4月份出现了较

Figure 9: 特定主题下推文数量前四的国家的推文数量和针对某个主题的情感波动。

大分歧，这可能表明人们对此问题的看法不同。

4 相关工作

最近有很多关于新冠肺炎的社交媒体分析的文献。我们将它们分为三类：数据集获取、社会和心理研究以及以自然语言处理方法为主的研究。

新冠肺炎相关的推特数据集 Yu (2020)创建了一个专用于机构和新闻媒体帐户的推特数据集。Chen et al. (2020)是一项持续的推特收集工作，其追踪可追溯至2020年1月22日的推文。他们利用推特的搜索和流应用程序接口来跟踪特定帐户，并实时收集涉及特定关键词的推文，例如**Coronavirus**（冠状病毒），**covid19**（新冠肺炎）和**social distancing**（社交距离）。GeoCov19(Qazi et al., 2020)是另一个推特数据集，包含超过五亿条多语言推文，可回溯至2020年2月1日。他们使用基于地名词典的方法来推断推文的地理位置。MegaCov(Abdul-Mageed et al., 2020)是一个十亿规模的多语言推特数据集，涵盖234个国家和地区、65种语言，具有超过3200万条带有地理位置标签的推文。此外，还有针对特定语言构建的数据集，例如，Alqurashi et al. (2020)和Haouari et al. (2020)是两个阿拉伯语推特数据集。

社会与心理学研究 Liang et al. (2019)从三个主流媒体的新闻报道中研究了在2008-2010年美国金融危机期间谁指责了谁。 Li et al. (2020)通过词频、情绪指标得分和情绪分析研究新冠肺炎对微博用户的心理后果。 Thelwall and Thelwall (2020)专注于转发次数最多的87条推文（这87条推文产生了1400万条转发），发现推特内容的主要主题包括隔离的生活、安全防护措施、人们对社交限制的态度、政治，以及对与新冠肺炎相关的工作者的关注。他们的工作以定性研究为主，我们的工作则以定量分析为主。 Rajput et al. (2020)在推特数据上使用了词频和情感分析。

关于推特的主题建模 主题建模已被研究人员广泛用于社交媒体分析和信息检索(Sun et al., 2017; Xu et al., 2017)。Wang et al. (2016)使用隐含狄利克雷分布来推断美国总统唐纳德·特朗普的追随者在推特上的话题题偏好，发现关于攻击奥巴马和希拉里·克林顿等民主党人的推特获得了最多的赞。最近，在社交媒体上有很多关于新冠肺炎分析的工作。 Boberg et al. (2020)分析了2020年1月至2020年3月下半月冠状病毒初期脸书上德国地区的帖子。 Wicke and Bolognesi (2020)根据在2020年3月和4月期间以关键词提取的20万条推文的语料库，分析了围绕新冠肺炎的论述。我们的工作和他们的区别在于：1）我们涵盖了更长的时间（2020年1月22日至2020年4月30日）2）我们侧重研究不同国家和地区（包括中国）的主题分布3）我们研究了人们对于相关主题的情感态度随着时间的动态变化。

Figure 10: 美国特定主题的情感得分。折线代表情感得分，条形图代表特定主题的推文数量

5 结论

我们分析了从1月22日（武汉封城前一天）到2020年4月30日的50万条包含了地理位置信息的推文，发现推特上最主要的话题是对新冠肺炎的看法和感受，包括仇恨言论和爱与支持两个极端。总体而言，讨论的气氛较为积极，并且随着时间的演变积极的程度逐渐增强。另外，中国、爱尔兰和印度比美国、澳大利亚和南非更积极。人们对特定主题的态度也随着时间而改变，例如，对隔离生活情感逐渐变得积极，但是不同国家的波动程度有所不同。我们希望我们的研究可以促进对社交媒体平台上的舆论的进一步理解。

致谢

感谢闵庆凯为本文提供许多帮助和建议。崔乐阳和何奇也参与了本文讨论。梁帅龙由新加坡科技设计大学校长奖学金资助。黄辉由澳门特别行政区科学技术发展基金资助（档案编号：0101/2019/A2）。张岳在工作中受到西湖大学融汇金信（www.rxhui.com）联合研究项目资助。张岳为通讯作者。

参考文献

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2020. Mega-cov: A billion-scale dataset of 65 languages for covid-19. *arXiv preprint arXiv:2005.06012*.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Salman Aslam. 2020. Twitter by the numbers: Stats, demographics & fun facts.

Mike Baker and Michael Crowley. 2020. Trump calls for calm on virus and expands travel restrictions. The New York Times.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Svenja Boberg, Thorsten Quandt, Tim Schatto-Eckrodt, and Lena Frischlich. 2020. Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis. *arXiv preprint arXiv:2004.02566*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Kevin Drum. 2017. Twitter is a cesspool, but it's our cesspool. Mother Jones.

Isaac Chun-Hai Fung, Zion Tsz Ho Tse, Chi-Ngai Cheung, Adriana S Miu, and King-Wa Fu. 2014. Ebola and the social media. *The Lancet, 384(9961), 2207*.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. {ArCOV-19}: The first arabic covid-19 twitter dataset with propagation networks. *arXiv preprint arXiv:2004.05861*.

Graham Hillard. 2018. Stop complaining about twitter - just leave it. National Review.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language misc (ALW2)*, pages 26–32, Brussels, Belgium, October. Association for Computational Linguistics.

The Lancet. 2014. The medium and the message of ebola.

Sijia Li, Yilin Wang, Jia Xue, Nan Zhao, and Tingshao Zhu. 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(6):2032.

Shuailong Liang, Olivia Nicol, and Yue Zhang. 2019. Who blames whom in a crisis? detecting blame ties from news articles using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 655–662.

Dianbo Liu, Leonardo Clemente, Canelle Poirier, Xiyu Ding, Matteo Chinazzi, Jessica T Davis, Alessandro Vespignani, and Mauricio Santillana. 2020. A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019*.

Sarah Martin and Luke Henriques-Gomes. 2020. Coronavirus: man evacuated from diamond princess becomes first australian to die of covid-19. The Guardian.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information. *ACM SIGSPATIAL Special*, 12(1).

Nikhil Kumar Rajput, Bhavya Ahuja Grover, and Vipin Kumar Rathi. 2020. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. *arXiv preprint arXiv:2004.03925*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Bernd Skiera, Lukas Jürgensmeier, Kevin Stowe, and Iryna Gurevych. 2020. How to best predict the daily number of new infections of covid-19. *arXiv preprint arXiv:2004.03937*.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25.

Mike Thelwall and Saheeda Thelwall. 2020. Retweeting for covid-19: Consensus building, information sharing, dissent, and lockdown life. *arXiv preprint arXiv:2004.02793*.

Jeremy Turiel and Tomaso Aste. 2020. Wisdom of the crowds in forecasting covid-19 spreading severity. *arXiv preprint arXiv:2004.04125*.

Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, and Tianran Hu. 2016. Catching fire via" likes": Inferring topic preferences of trump followers on twitter. In *Tenth International AAAI Conference on Web and Social Media*.

WHO. 2020. Public health emergency of international concern declared. World Health Organization.

Philipp Wicke and Marianna M Bolognesi. 2020. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. *arXiv preprint arXiv:2004.06986*.

Xinhua. 2020. China to extend spring festival holiday to contain coronavirus outbreak. Xinhua Net.

Zheng Xu, Yunhuai Liu, Junyu Xuan, Haiyan Chen, and Lin Mei. 2017. Crowdsourcing based social media data analysis of urban emergency events. *Multimedia Tools and Applications*, 76(9):11567–11584.

Jingyuan Yu. 2020. Open access institutional and news media tweet dataset for covid-19 social science research. *arXiv preprint arXiv:2004.01791*.