

试析美国主流媒体在推特平台关于涉华疫情的发文特征

王丽娜^{1,2}, 耿雪芹²

(中国人民大学 信息资源管理学院, 北京 100872; 2. 湖南蚁坊软件股份有限公司, 湖南 长沙 430000)

摘要: 在中国新冠疫情面临新一轮攻坚战的当下, 美国媒体对涉华疫情话题的关注也呈持续态势。以美国主流媒体在推特平台发布的涉华疫情推文及其下评论转发等数据为研究对象, 媒体涵盖美国有线电视新闻网、美联社等 10 家美国主流媒体, 时间跨度为 2022 年 8 月 24 日至 2022 年 11 月 23 日, 共采集 163 条涉华疫情推文以及约 8 788 条评论转发等数据。通过定量分析方法总结美国主流媒体对涉华疫情的推特发文特征, 进一步明晰其报道偏向与手法, 并就中国涉疫信息的对外传播提出建议。

关键词: 中国; 新冠肺炎疫情; 推特; 西方媒体

中图分类号: G203

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2023.02.014

引用格式: 王丽娜, 耿雪芹. 试析美国主流媒体在推特平台关于涉华疫情的发文特征[J]. 网络安全与数据治理, 2023, 42(2): 89-95.

Experimental analysis of the characteristics of U.S. and Western media's tweets on the China-related epidemic

Wang Lina^{1,2}, Geng Xueqin²

(1. School of Information Resource Management, Renmin University of China, Beijing 100872, China;

2. Hunan Eefung Software Co., Ltd., Changsha 430000, China)

Abstract: As China's novel COVID-19 outbreak launched a new round of offensive, U.S. media attention on China-related outbreak issues has also continued. The article's research object takes 163 China-related tweets posted by 10 U.S. mainstream media, including CNN and Associated Press from August 24, 2022 on November 23, 2022, and about 8 788 comments and retweets under each tweet. Meanwhile, the article tries to summarize the characteristics of U.S. and Western media's tweets on China-related epidemic through quantitative analysis, further clarify their reporting bias and methods and make suggestions on the foreign communication of China's epidemic-related information.

Key words: China; COVID-19; twitter; Western media

0 引言

在中美关系日趋紧张的大背景下, 中国涉疫话题仍为部分美国媒体关注的重要议程。特别是美国部分主流媒体基于政治偏见, 针对中国抗疫举措炮制各种论调, 行政治化、标签化和污名化之实, 对中国的国际形象造成一定冲击和影响^[1]。自 2020 年新冠肺炎疫情全面暴发以来, 各领域学者相继对西方媒体涉华疫情新闻报道文本进行话语特征、报道偏向以及背后逻辑、建构动因等多维度分析, 而对推特平台上的涉华疫情文本及其评论转发内容却相对少有专门研究与分析。但实际上, 伴随网络技

术的发展, 社交媒体已经成为疫情期间获取相关信息、舆情表达以及信息共享的最重要途经之一^[2]。根据推特官方 2022 年 9 月 12 日信息发布显示, “推特用户是狂热的新闻消费者”, 有“94% 的推特用户表示对时事感兴趣”^[3]。

本文基于推特平台上美国主流媒体账号发布的关于中国新冠疫情相关话题数据, 基于深度学习的自然语言处理模型和语义网络等量化分析方法, 提出了话题分布、提及热词、情感倾向、文本结构、传播反响等分析模型, 探究美国主流媒体推特账号关于涉华疫情话题的发文特征, 以期进一步把握当

下美国主流媒体对中国涉疫话题的报道取向、反响情况,以针对性地提出相关应对策略,促进对外传播的良性构建。

本文选取美国有线电视新闻网(@CNN)、美联社(@AP)、《纽约邮报》(@nypost)、《纽约时报》(@nytimes)、福克斯新闻网(@FoxNews)、《华尔街日报》(@WSJ)、美国广播公司(@ABC)、美国全国公共广播电台(@NPR)、《洛杉矶时报》(@latimes)、《华盛顿邮报》(@washingtonpost)10家美国主流媒体所发推文及其传播情况为研究材料,具体研究和分析了2022年8月24日至2022年11月23日的涉华疫情相关内容,整理并研究主题相关的163条推文以及各推文下的约8788条评论转发等数据为分析样本。

1 话题提取及聚类分析

目前网络话题仍然以文本为主要的表达方式,而现阶段从文本中提取话题的技术手段仍然局限在词法级别,即依靠关键词、热词、共现词、敏感词、情感倾向词、实体词识别等寻找与话题有关的信息。然而,仅在词的级别上来分析,获取到的往往是局部的信息,无法获取到完整的语义信息。

鉴于此,本文给出一种基于依存句法分析(Dependency Parsing)的话题分析^[4]的方法,能够从海量文本中发现各种有意义的话题,以及从特定舆情事件相关的信息文本中发现该事件的子话题和其他相关话题。本技术设计思路为先从文本数据流中提取话题标签,描述文本所谈论的内容,将相似的话题标签聚类,并将文本分配到聚类后的话题标签下。

1.1 话题标签提取

本文中使用的话题标签借鉴了百度、新浪微博等热搜榜单对话题的描述形式,使用依存句法分析算法实现话题标签的抽取。依存句法分析算法中常见的中文句法语料主要有哈工大信息检索研究室汉语依存树库(HIT IR-Lab Chinese Dependency Treebank)和清华大学信息检索研究室汉语依存树库(THU Chinese Dependency Treebank);主流的模型主要有三种:基于图的分析方法、基于转移的分析方法或基于深度学习的分析方法。另外,目前已实现了依存句法分析算法的工具包有:斯坦福开发的Stanford NLP、复旦大学开发的FuDanNLP、hankcs开发的Hanlp以及哈工大开发的LTP等。

本文中的话题标签提取的实现过程如下:

首先,基于依存句法分析模型或工具将文本句

中的词分成主语、谓语、宾语、定语、状语、补语等中文句法成分,并确定词与词之间句法依存关系(如图1所示)。词与词之间的句法关系构成依存对,其中包含两个词(核心词、修饰词)和一个依存弧(用来表示两个词的句法依存关系),依存关系包括主谓、动宾、定中、状中等。



图1 词之间的句法关系

其次,根据词之间的句法关系,按照常用句法表达组合生成简洁通顺的短语作为话题,并从文本中提取出话题集。句法组合包括<主谓关系+动宾关系>的词组合成主谓宾结构短语、<定中关系+主谓关系>的词组合成名词短语等。

1.2 话题标签聚类

话题标签聚类指的是将语义相似的话题标签聚类。一方面,将抽取得到的话题标签进行相似性合并,如通过聚类、编辑距离、语义词向量等算法进行相似性判定;另一方面,根据话题标签所在的文档分布进行合并。若两个话题标签所在的文档列表中重复的文档个数超过预设占比,则认为这两个话题标签来自于语义相同的文本,即认为两个文本语义相似。根据语义相似将话题标签聚类后,按照类的大小进行排序。

1.3 话题提取及聚类分析算法流程及结果

话题提取及聚类分析算法的主要流程包括如下四个步骤(如图2所示):

(1)数据预处理:将数据进行分句、分词、去除停用词处理,然后计算分词结果中每一分词的文档频率,并进行降序排列,获取排列在前的top词作为关键词,从句子集合中将不包含话题关键词的句子去除,得到包含话题的句子。

(2)句法分析:将步骤(1)结果输入到句法分析模型,输出句子的句法依存对。

(3)话题标签提取:根据步骤(2)结果确定与核心词依存关系为主谓或动宾关系的分词,按照常用句式表达进行组合,如主谓宾顺序<主谓关系词+核心动词+动宾关系词>等,获得话题标签。



图2 话题分析算法流程

(4)话题标签聚类:对步骤(3)的结果,保留在预设长度区间内的话题,计算话题标签语义相似度,将语义相似的话题聚类排序,选择 top 话题输出作为最终的话题。

基于上述算法,对选取的数据集进行处理,得到如图3所示话题分析结果。通过对文本内容进行主题聚类分析可知,在观察时间段内,10家美国主流媒体所发涉华疫情推文主要集中于防疫政策话题。其中,近三月期间,《华尔街日报》(37条)、《纽约时报》(31条)、美联社(25条)等媒体累积发推131条,在总体发文中占比高达81%,具体报道内容多涉及关注中国坚持严格的防疫政策。

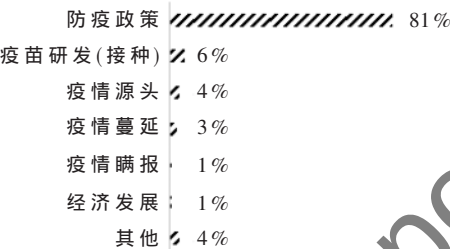


图3 美国主流媒体所发涉华疫情推文话题总体分布

2 情感分析

情感分析(Sentiment Analysis)是经典的自然语言处理技术,常用于分析文本中表达的情感倾向,属

于文本分类中的一种。目前情感分析任务在实现上主要是分为基于词典的情感分析方法和基于机器学习/深度学习算法构建情感分析模型这两大类主流方法。其中,情感词典和匹配规则是核心,情感词典的标注质量直接决定情感分析的结果,但情感词典的构建耗时耗力,另外在识别时还要考虑转折句、否定句以及情感词程度,难以把握。鉴于此,本文采用学习情感语义特征的TextCNN模型^[5]来构建情感分类模型。

2.1 TextCNN 模型

与图像处理中的CNN网络相比,TextCNN在网络结构上没有任何变化,其通过卷积和池化获取局部语义信息并提取文本特征,局部感知和权值共享克服传统神经网络的缺点。在情感分析中,卷积神经网络的结构无需复杂就能取得很好的结果,主要包括输入层、卷积层、池化层、全连接层、softmax层,如图4所示。

在输入层,通过使用word2vec将文本序列信息转为数值型向量,每个词对应相同维度的向量,维度大小可人为设定。就此输入文本序列转换为一个行大小为序列长度,列大小为词向量长度的向量矩阵。可表示为:

$$x_{1:n}=x_1\oplus x_2\oplus\cdots\oplus x_n \tag{1}$$

其中, x_i 表示第*i*个词所表示的词向量,而 $x_{1:n}$ 则

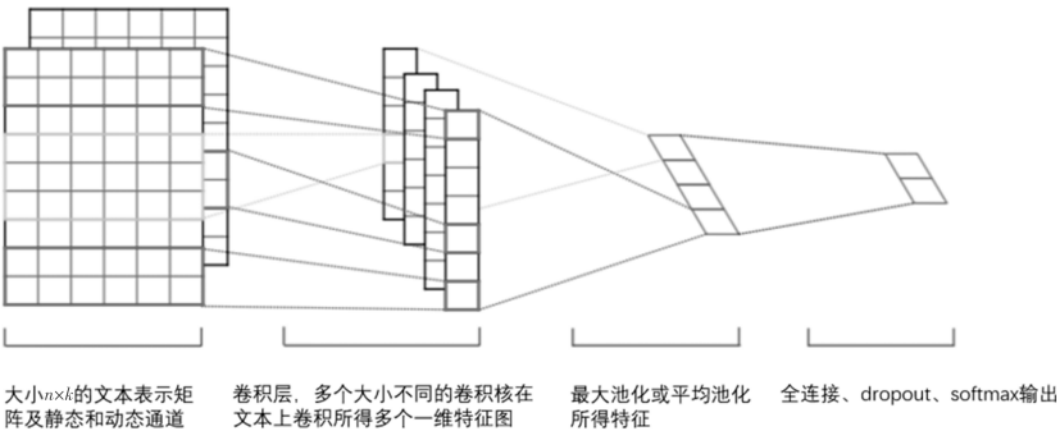


图4 基于情感分析的卷积神经网络基本结构

表示使用 word2vec 转化所得该输入序列的向量矩阵。

在卷积层,在循环神经网络处理文本时,卷积核的宽度必须保持与词向量的维度一致。在特征抽取时,如果卷积核的宽度小于词向量维度,抽取信息也失去了词特征。卷积核的高度可自由设定(一般取值为 3,4,5),不同大小的多个卷积核在输入矩阵上卷积,可得到多个特征向量,卷积过程不仅考虑到了语义而且考虑了其上下文关系。卷积操作表示为:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad (2)$$

式中 w 为卷积核的权重矩阵, n 为输入文本序列长度, h 为卷积核窗口大小, b 为偏置值, f 函数是激活函数。经过一次卷积,可得一个 $n-h+1$ 维的向量 c :

$$c = [c_1, c_2, \dots, c_{n-h+1}] \quad (3)$$

在卷积过程中,使用不同大小的卷积核,所得卷积特征向量的维度也不相同,需要紧接池化层处理。池化可以降低特征维度,防止过拟合,保留显著性特征。在池化层,一般使用最大池化或均值池化。

$$C_{\max} = \max[c_1, c_2, \dots, c_{n-h+1}] \quad (4)$$

$$C_{\text{mean}} = \text{mean}[c_1, c_2, \dots, c_{n-h+1}] \quad (5)$$

最大池化,即保留每个卷积核卷积后向量中最大值来表示重要特征,结合每个卷积核卷积后特征向量中的最大特征,拼接起来得最终的特征向量。均值池化就是保留均值作为特征,将其连接为特征向量。

卷积和池化提取的局部特征不能以偏概全地表示全局特征,而若将所有的局部特征全连接很可能导致模型过拟合,即在训练集中分类准确率很高,但在训练集文本以外的文本上效果差强人意。所以在池化与全连接之间需要 dropout,一般设置 dropout=0.5,意思是将池化得到的特征向量随机脱落一半,也就是将一个 n 维的向量变成 $n/2$ 维向量。

对于研究情感分类问题,全连接层可分为两个部分,一是先激活,常用 ReLU 作为激活函数($f(x) = \max(x, 0)$);二是使用分类激活函数得每个类的概率,取概率最大的类为该文本所预测的类。

基于 TextCNN 进行情感分析的数据主要包括两个部分:一部分是准备已标注情感倾向性的文本数据,通过 TextCNN 对其进行训练,从而得到情感分类模型;另一部分是基于分类模型准备的测试数据,通过模型输出相应的情感倾向。

2.2 情感分析模型训练过程

情感分类模型的训练步骤包括:

(1)准备已标注情感标签的文本数据。首先搜集一些公开语料库提供的情感标注数据,将其提炼成正面、负面和中立三大情感类别,然后通过一些开源的情感分析工具进行情感标签比对,分析标签和原标签不一致的数据,再随机抽取这些情感标注数据,人工分析其标注准确性,对不合理、不确定的情感标注数据进行标签调整或删除,最终得到一批质量较好的情感标注语料。

(2)进行文本预处理,准备模型的训练数据。训练数据为标注好情感标签的预处理文本,其中预处理文本主要为分词和去噪。分词是按词义将句子分割成词;去噪是去掉不含情感语义特征的词,如一些口语化的虚词、量词、标点等。

(3)训练 TextCNN 情感分类模型。基于准备好的训练语料,输入到 TextCNN 神经网络,进行调参迭代,最终得到训练完成的情感分类模型。

(4)对训练完成的情感分类模型进行评估。若输出结果达到预设指标(一般为模型的 F1 值达到 80%),则判定该模型合格。

2.3 情感分析算法实现流程及结果

本文基于 TextCNN 对数据集进行粗粒度的情感分析,即将群体言论划分为正面、负面、中立 3 类,具体流程如图 5 所示。

分析结果显示,在情感倾向方面,美国主流媒体关于涉华疫情的推文主要呈现负面态度(如图 6 所示)。在观察时间段内,10 家美国主流媒体所发涉华疫情推文皆以负面情感为主,占比达 92%,主要是抨击与抹黑中国防疫政策、关注中国疫情持续蔓延态势、造谣新冠肺炎疫情可能为中国实验室泄露等;中立情感推文次之,占比达 7%,主要是关注中国疫苗研发与接种新进展、传播中国防疫政策调整动态等;正面情感推文最少,占比仅为 1%,主要是美联社声称的中国首创可吸入式新冠疫苗以及《洛杉矶时报》称“没有证据表明新冠病毒是从中国实验室泄露的”。

3 其他特征分析

本文基于语义网络分析思想^[6]及直观统计模型,对数据集进行了多方面的特征分析探索,包括提及热词分析、文本结构分析、传播反响分析及重复发文分析等。

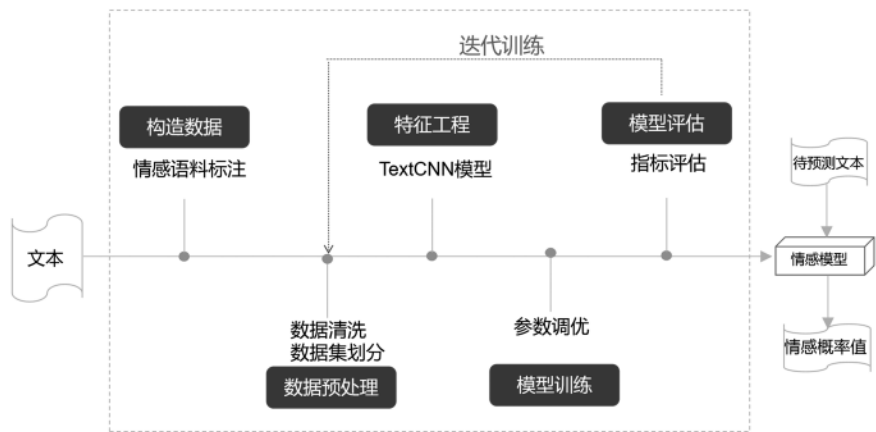


图 5 情感分析算法实现流程

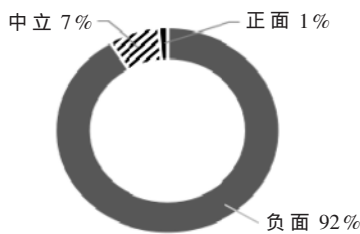


图 6 美国主流媒体所发涉华疫情主题推文情感倾向分布

最多,达 35 次,相关推文内容主要是关于中国防疫坚持的“动态清零”“零容忍”政策;“iPhone”(16 次)、“Apple”(11 次)、“factory”(10 次)等词汇的词频突出,则主要是《纽约时报》《纽约邮报》《华尔街日报》、美国广播公司、美国有线电视新闻网等媒体相继发文关注,受郑州疫情与防疫政策影响,苹果公司最新系列 iPhone 的出货量将受到影响。

3.2 文本结构分析

如表 2 所示,使用“文字摘要+新闻卡片”结构的推文共计 109 条,约占所有样本的 67%。

表 2 美国主流媒体所发涉华疫情主题推文文本结构分布

参数	发布数量/条
文字摘要+新闻卡片	109
新闻标题+新闻卡片	16
新闻标题+新闻链接+图片	9
文字摘要+新闻链接+视频	7
文字摘要+新闻链接	5
文字摘要+视频	2
新闻标题+新闻链接+视频	1
新闻标题+新闻链接	1

3.1 提及热词分析

基于 NLP 文本分词技术和词频统计方法,分析结果如下:

在提及热词方面,美国主流媒体关于涉华疫情呈现出对“清零”政策的高度关注(如表 1 所示)。通过对文本内容进行词频分析可知,在观察时间段内,除去无意义词汇以及“新冠”“中国”词汇后,10 家美国主流媒体所发涉华疫情推文提及“zero”的次数

表 1 美国主流媒体所发涉华疫情主题推文高频提及热词(Top20)

排名	高频词汇	词频	排名	高频词汇	词频
1	zero	35	11	Apple	11
2	policy	27	12	government	10
3	Beijing	18	13	people	10
4	new	17	14	strict	10
5	iPhone	16	15	factory	10
6	down	15	16	Guangzhou	9
7	country	14	17	city	9
8	restrictions	14	18	death	8
9	outbreak	12	19	measures	7
10	cases	11	20	production	7

“新闻卡片”一般包含新闻标题、新闻图片与新闻链接,用户可通过点击卡片直接进入媒体官网中的对应新闻页面(如图 7 所示)。这种“文字摘要+新闻卡片”的文本结构,一方面相较新闻报道全文更加简洁明了、重点突出,另一方面又包含新闻图片,醒目吸睛,还可以直接跳转新闻原文,在助力信息传播的同时,还可增加媒体网站自身流量。



图 7 《纽约时报》、福克斯新闻网所发包含“新闻卡片”的推文样例

3.3 传播反响分析

在传播反响方面,美国主流媒体关于涉华疫情的推文单条传播反响相对有限(如图 8 所示),主题相关推文单条转发量多居于 0 至 100 次区间、点赞量多居于 0 至 200 次区间,所有推文的单条转发量均未超过 500 次。而在观察时间段内,仅《华尔街日报》账号所发推文,就出现 26 条转发量达 500 次以上的内容,主要关涉美国国内总统选举、疫情应对、经济金融等,最高单条转发量达 5 867 次、最高单条点赞量达 32 361 次。

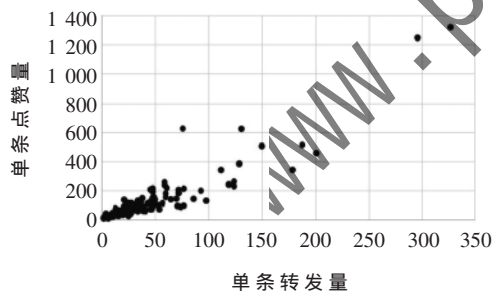


图 8 美国主流媒体所发涉华疫情主题推文单条转发量与点赞量分布

同时,在观察时间段内,美国主流媒体关于涉华疫情的推文中,转发量与点赞量最高的两条推文为福克斯新闻网于 10 月 29 日 00:15 与 4:45 两个时间点分别发布的同一内容,即“COVID-19 pandemic ‘likely’ originated with Chinese lab leak, Senate Republican report finds”(译文:参议院共和党报告发现,COVID-19 大流行病“可能”源于中国实验室的泄

漏),而通过查看上述两条推文的评论区内容,又发现网民讨论大多显现出对“共和党”的不满与抨击。

3.4 重复发文情况

此外,值得注意的是,21 条文本被发布 2 次以上,涵盖 8 家美国主流媒体推特账号。一方面,《华尔街日报》对内容“Want to fly to China? Prepare to pay thousands of dollars for a seat in coach as Beijing maintains its zero-tolerance policy toward Covid-19”(译文:想飞往中国?准备好为经济舱支付数千美元,因为北京坚持对 Covid-19 的零容忍政策)在 10 月 18 日至 20 日期间连续发布 4 次;福克斯新闻网对内容“Millions in China under COVID lockdown restrictions”(译文:中国有数百万人受到 COVID 的封锁限制),在 9 月 5 日至 6 日期间连续发布 3 次。

另一方面,在 9 月 7 日,美联社与美国广播公司对同一文本内容“India and China have cleared a new approach in COVID-19 vaccination—two needle-free options, one a squirt in the nose and the other inhaled through the mouth”(译文:印度和中国已经批准了 COVID-19 疫苗接种的新方法——两种无针方案,一种是在鼻子里喷洒,另一种是通过嘴吸入)进行传播,不过美联社文本后附带的是美联社官网新闻卡片,而美国广播公司附带的是美国广播公司官网新闻链接。

4 结论

总体而言,当前美国主流媒体在推特平台内关于涉华疫情的发文内容仍多以负面情感倾向为主,

并重点抨击中国严格的防疫政策,尤以“动态清零”“零容忍”“封城举措”为甚。在其建构的话语空间中,“动态清零”“封城举措”往往与所谓的“政府强制”“民众不满”“剥夺自由”等消极话题有关,直接导致相当部分网民在评论区发表涉华负面观点,甚至质疑中国疫苗有效性。预计后续“中国严格的防疫政策”仍为美西方主流媒体的重点报道与方向,中国官方与主流媒体对外宣传可对应做好进一步传播与反制工作。

一是,妥善应对与处置防疫政策衍生的负面舆情。在各地杜绝“一刀切”现象的同时,政府方面与主流媒体亦需进一步做好相关政策解读、民众问题与建议听取工作。

二是,进一步释放有关中国科技抗疫、防疫的进展信息,形塑中国海外积极形象。鉴于美国主流媒体涉华疫情话题的正面与中立信息多集中于中国疫苗研发与接种新进展,中国官方与主流媒体可加大对我抗疫领域科研新进展的消息释放与新闻报道。

三是,持续发布有关新冠溯源的纠偏信息,反制美西方对中国的污名化攻击。考虑到美国主流媒体对所谓新冠肺炎疫情为“中国实验室泄露”“起源于中国”等谣言的反复炒作,中国媒体也可持续对新冠溯源话题进行关注报道,无论是科研成果、调查报告,抑或是谣言澄清、观点驳斥等,都可作为语料素材加以应用。

四是,结合美西方本土动态资讯,优化涉华疫情信息的海外传播。自2008年开始,中国就已经开始打造具有较强国际影响力的外宣旗舰媒体,加强国际社交媒体能力建设^[7]。相对涉华疫情,美国国内网民更关注于本土资讯动态,尤其涉及两党议题与总统选举的话题。故中国媒体在海外舆论场进行涉华疫情话题信息传播时,可因地制宜,借力美国本土舆论热点主题。

后续将进一步在以下方面开展工作:

(1)基于网络议题设置理论及相关研究,探索西方媒体针对中国的议题设置策略,及时发现涉中新议题,并实时跟踪分析其进展演变,做到未雨绸缪。

(2)在本文热词分析及话题分析的基础上,建立不同媒体关于同一话题的热词共现网络关系,以发现美国主流媒体作为群体针对中国话题的协同炒作及传播发展脉络。

参考文献

- [1] 孙明,孔祥龙.国际舆论视域下的中国抗疫行动与公共外交[J].当代世界,2020(5):11-17.
- [2] DEPOUX A, MARTIN S, KARAFILLAKIS E, et al. The pandemic of social media panic travels faster than the COVID-19 outbreak [J]. Ravel. Med., 2020, 27(3): taa031. DOI: 10.1093/jim/taaa031.
- [3] How many people come to Twitter for news? As it turns out, a LOT[EB/OL].[2022-12-09].https://blog.twitter.com/en_us/topics/insights/2022/how-many-people-come-twitter-for-news.
- [4] 耿雪芹,王晓斌,焦梦姝,等.一种话题分析方法,装置和存储介质,CN110874531A[P].2020-03-10.
- [5] KIM Y. Convolutional neural networks for sentence classification[D]. Ithaca: Cornell University, 2014.
- [6] 谷羽.语义网络分析方法在传播学中的应用及批判[J].《现代传播》,2019(4):155-159.
- [7] 相德宝,杨月清.百年未有之大变局背景下国际涉华舆论斗争的新形势与新特点[J].统一战线学研究,2022(2):101-108.

(收稿日期:2022-12-09)

作者简介:

王丽娜(1993-),女,本科,主要研究方向:网络舆情、开源情报。

耿雪芹(1992-),女,硕士研究生,工程师,主要研究方向:自然语言处理、深度学习、知识图谱。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.cn