基于事件不确定性的 Twitter 主题上下文向量关联

Vaibhav Khatavkar 浦那工程学院

Swapnil Mane 浦那工程学院

Parag Kulkarni 浦那工程学院

arXiv:2304.01423v1 [cs.CL] 4 Apr 2023

摘要—**关键词提取是文本挖掘中的关键过程。从 Twitter 数据中提取关键词及其对应的上下文事件是一项重大挑战。主要难点在于所用语言的非正式性。拼写错误词、首字母缩略词和歧义术语的使用导致了语言的非正式性。当前系统对非正式语言的关键词提取采用基于模式或基于事件的方法。本文借助数据关联，利用主题事件提取情境关键词。所提出的系统依据不确定性原理识别事件的主题上下文。系统借助称为主题上下文向量的向量对主题上下文进行加权，这些向量用于表示事件是确定的还是不确定的。系统在 Twitter COVID-19 数据集上进行了测试，结果证明其有效。系统从测试数据集中提取特定于事件的主题上下文向量并对其排序。提取出的主题上下文向量用于聚类情境主题向量，使轮廓系数相比词频和词频-逆文档频率等最先进方法提高了 0.5%。主题上下文向量还可用于网络欺凌、反讽检测、比喻性语言检测等其他应用。**

一、引言

对于文本数据，通常会运用自然语言处理来开发命名实体识别、产品排名、句子排名、推荐系统等系统。文献 [29] 讨论了利用数据挖掘方法从社交媒体内容中提取数据。Dylan Grosz 和 Patricia Conde-Cespedes 在其研究 [13] 中采用深度学习方法进行仇恨言论检测。研究人员指出，规模、噪声和动态性是社交媒体内容的重要特征。Wahab Khan 等人在其研究 [17] 中介绍了迄今为止采用的自然语言处理方法。研究人员还阐述了如何通过词性标注、词义消歧和话语分析来解决文本处理中的歧义问题。此外，研究人员也在自然语言推理方面开展了一些探索，试图从文本数据中推断含义 [18]。

通过社交媒体、博客等渠道，互联网上有海量文本格式的数据可供使用。 [2]。“每天发送 5 亿条推文、294 封电子邮件，以及 650 亿条 WhatsApp 消息。社交媒体平台上约有 3 亿名被识别为新用户的用户。” [1] 文本挖掘涉及多个领域，这些数据可在其中转化为有用的信息。潜在狄利克雷分配等主题建模算法被应用于 Twitter 数据，以提取主题 [25]。David 和 Martin 还开发了一种用于 Twitter 数据的主题模型[5] ，该模型基于作者主题模型。David Alvarez-Melis 和 Martin Saveski 在 [5] 中通过使用两种主题建模技术，即潜在狄利克雷分配和作者主题模型，开发了一种用于推文及其配方的主题模型。作者主题模型被证明比潜在狄利克雷分配更有效，但耗时更长。此外，数据聚类能够将相似的推文归为数据簇 [30]。，Twitter 数据也可以分为多个类别 [34]。Namugera 等人在其研究 [22]中，利用推文的正面和负面情感对主题进行关联。他们还指出，负面推文在 Twitter 上的信息传播中发挥着重要作用。将 Twitter 视为时间序列并进行事件识别。

Richard J. Povinille 在其论文 [23] 中将时间序列数据挖掘应用于时间金融数据，以发现其中隐藏的模式。他定义了一个目标函数，用于从金融数据中的股票寻找交易优势。该系统在纽约证券交易所等股票市场的投资交易数据上进行了测试。Masoud Makrehchi 等人将社交媒体文本中的市场涨跌作为事件，并据此制定营销策略 [3]。Selene Yue Xu [32] 利用 Yahoo Finance 和 Google Trends 数据开发了股价预测系统。他们进行了 ARIMA 时间序列分析，并分析了新闻与事件之间的相关性。研究人员得出结论，新闻会影响股价预测。Nikan Chavoshi 等人 [10] 识别 Twitter 上的异常用户账户；他们采用滞后敏感哈希技术和扭曲不变相关性度量来识别异常用户并分析其相关性。Saeed 等人借助动态心跳图，利用 Twitter 时间序列数据的动态特性进行事件检测 [27]。

II. 相关工作

人们尝试将 Twitter 数据映射为时间序列并进行分析。Salvatore 等人 [28] 在其研究中以 Twitter 为基础示例，对社交媒体内容进行了定性和定量分析。他们构建了一个及时反映社交媒体情况的社会经济指标。Marta 等人 [7] 在其研究中使用 Twitter 时间序列分析提取了社会、经济和商业指标。Alshaabi 等人[4]分析了与医疗、政治及 COVID 相关操作有关的推文。这些推文涵盖 24 种语言，收集时间为一年，即从 2019年3月至2020年3月。他们观察到，在 1月和2月，「virus」一词在所有语言中的出现值都达到峰值。Cakit 等人 [9] 研究了监督式软计算技术的应用，包括模糊时间序列（FTS）、基于人工神经网络（ANN）的模糊时间序列以及自适应

神经模糊推理系统（自适应神经模糊推理系统），用于预测 Twitter 数据中表达的情绪状态。Gonzalo 等人[26]使用贝叶斯网络识别关键事件，这些事件包括自然灾害和社会运动。研究人员利用贝叶斯网络寻找词语之间的相关性。推文中的模式对分类很有用。Malayak 等人[21]在其研究中使用 Twitter 时间序列，通过分析 2015年至2019年期间 Twitter 上的推文、转推、提及、话题标签和收藏来预测奥斯卡奖得主。他们使用逻辑回归和时间序列分解来预测奥斯卡奖得主。

Pratap 等人 [24]利用分类和聚类后的推文及其上下文进行检测。该系统可用于监控高速公路等大型区域的交通状况。Arapostathis 和 Karantzia [6]利用志愿地理信息（VGI）和 Twitter 数据进行灾害管理。该系统识别了火灾事件，并对其进行跟踪和分类。Chen 和 Terejanu [11] 研究了子事件识别。在他们的方法中，他们使用卡尔曼滤波器、高斯过程和概率主成分分析来构建异常值检测问题。Behzadan 等人 [8] 检测了社交网络（尤其是 Twitter）中的网络安全事件。检测到的事件使用卷积神经网络（CNN）进行分类。

Hossny 等人 [15]使用朴素贝叶斯分类器，根据推文中出现的历史事件和抗议活动等实时事件来预测未来事件。Hasan 等人[14] 将基于随机索引的词项向量模型与局部敏感哈希相结合，以便在固定时间范围内对与各类事件相关的推文进行聚类。这有助于从数据中识别更多热门、活跃且有意义的新闻。Katragadda 等人 [16] 通过评估推文对话模式中的主题，找出了可信事件。Kolchyna 等人 [19]研究了 Twitter 数据时间序列中最优事件窗口的自动识别。在识别事件后，他们进行了事件聚类（聚类类型）和量化。Grindrod 等人 [12] 在其研究中使用了 Twitter 时间序列中的事件峰值。Eric Lai 等人 [20] 通过构建主题之间的关系，在时空域中对 Twitter 数据中的主题进行建模。Yadav 等人[33] 使用时间序列对与印度板球超级联赛 T20 相关的推文进行建模，以识别事件。Arias 等人 [4]利用 Twitter 时间序列中的事件分析票房收入趋势并进行预测。

通常，事件相关数据采集通过将关键词用作可用数据中的搜索模式来完成。数据规模是一个主要问题。一些研究人员报告称，使用大型 Twitter 数据集成功进行了事件检测。 [31] 提出了事件识别方法，以检索更多与特定内容相关的事件。该方法监测事件，并使用词频-逆文档频率（TF-IDF）根据事件与预定义事件的相关性对其进行评分。更多相关的突发事件会实时作为过滤模型插入。

在所提出的方法中，主题上下文基于可用数据构建，这些数据可以很少，也可以很多。主题上下文由关键词及其与事件的关联构成。事件可以从搜索模式中推导出来，而搜索模式是针对数据的查询。

III. 所提方法

时间序列通常是由时间和数据构成的元组，即 TS = T_i, D_i。根据给定的时间序列，可以对事件和数据关联进行解读，以供后续处理。时间序列数据如图 1 所示。该图展示了发生在不同时间点的各种事件 E_1， E_2、En。事件是来自 Twitter 时间序列中推文的关键词，表示为 D_1， $D_2 \ldots D_n$。

时间序列中的事件可以按主题关联，形成主题上下文向量。可以计算时间序列中上下文事件的主题向量的不确定性，并进行知识发现，以便对时间序列数据中的确定主题上下文向量进行排序。

所提出的系统如图 2 所示。该系统将文本数据转换为时间序列数据，用于识别主题上下文。主题上下文通过数据与事件的关联来识别。

主题上下文的不确定性根据下方的公式（1）推导得出。

$$Uncertainty\ UN_j(i) = \ 1 \ - IG_j(i) \tag{1}$$

$$RankedWeight = UN_j(i) + Count(i)/Totalcountofallwords \tag{2}$$

其中， $IG_j(i) = C(j) * H_j(i)$ 表示 j事件的信息增益，

$H_j(i) = -(P(j) * log_2(1/P(j)) + P(i) * log_2(1/P(i)) + P(i|j) * log_2 P(i|j))$ 表示给定单词 i时事件 j的上下文熵，

$C\ (j) \ = count\ of\ event\ j\ in\ time\ series$。

计算主题上下文的不确定性后，会生成其对应的上下文向量。主题上下文向量可以排序和聚类。下面给出了用于查找主题上下文的算法。

查询是需要从 Twitter 数据中获取的答案。查询对应的主题上下文向量将被生成。排序后的主题上下文向量形式如下：$CV_j =< word_{1j},\ uncertainty_{1j},\ word_{2j},\ uncertainty_{2j} \ldots$。主题上下文向量 CV_j 将确定事件 j 的主题。

IV. 实验

实验使用了公开可用的 COVID-19 Twitter 数据集，如表 I 所示。排除所有转推和话题标签后，推文总数为 3116（表 II）。

数据集中的推文以时间序列表示为 TS = $Ti,\ Di$ 其中 T = 2020-05-02 00:59,2020-05-02 00:03，D = 表示推文中的关键词。图 3 展示了该时间序列，其中 X 轴表示推文时间，Y 轴表示该时间点的关键词数量。

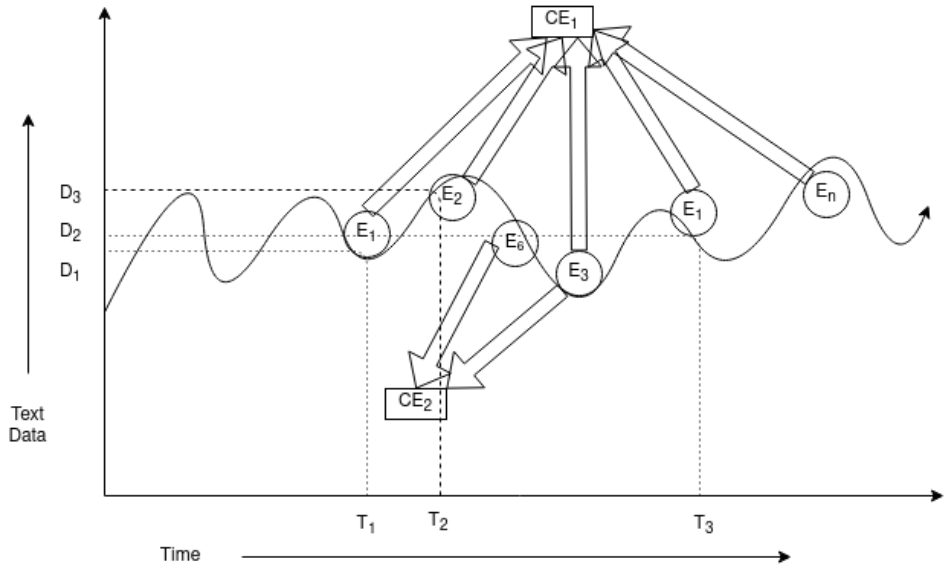

图 1. 情境事件发生时的 Y 轴需更改

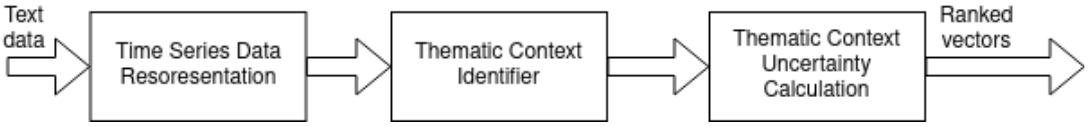

图 2. 基于不确定性的主题上下文向量关联系统

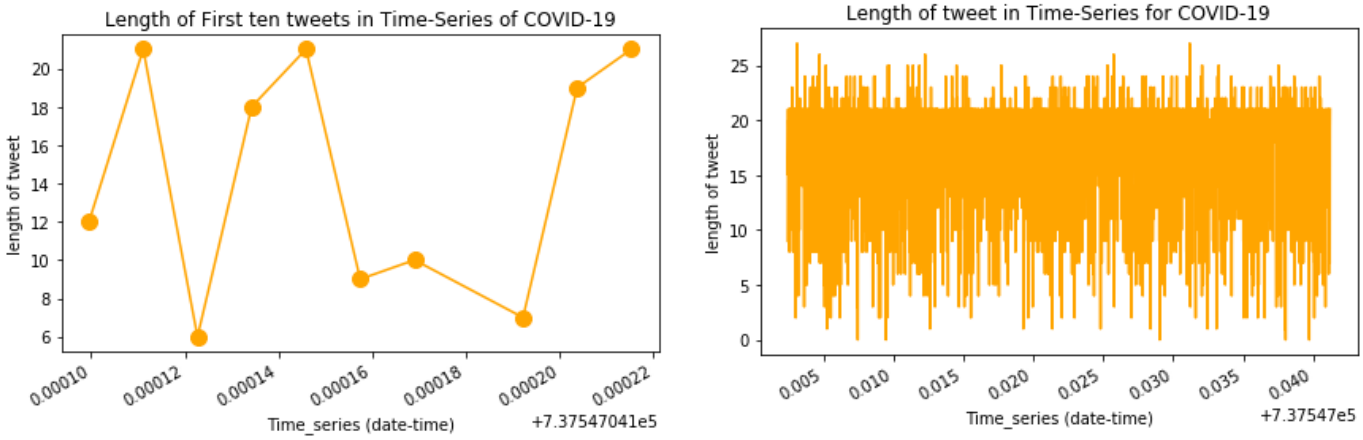

图 3. COVID-19 时间序列中每条推文的长度

图 4 展示了事件时间序列，其中 X 轴表示发布推文的时间，Y 轴表示该时间点发生的事件数量。事件数量随时间变化。该数据的事件时间序列在数值 1 处出现峰值，表示事件的存在。连续发生的事件以粗线表示。表 4 展示了 COVID-19 数据集中各类事件对应的前 10 个确定的主题上下文向量。

从事件时间序列和关键词时间序列中，任务是寻找它们之间的主题上下文。对于主题上下文向量识别，考虑了各种查询。找出了与事件相关的主题上下文关键词并进行了排序。

使用 K-means 聚类来提取主题事件。为选择聚类数量，采用肘部法。该方法给出每个聚类的惯性。图 5 显示了按簇惯性；在主题事件聚类中，肘部出现在聚类数量接近 5 时。

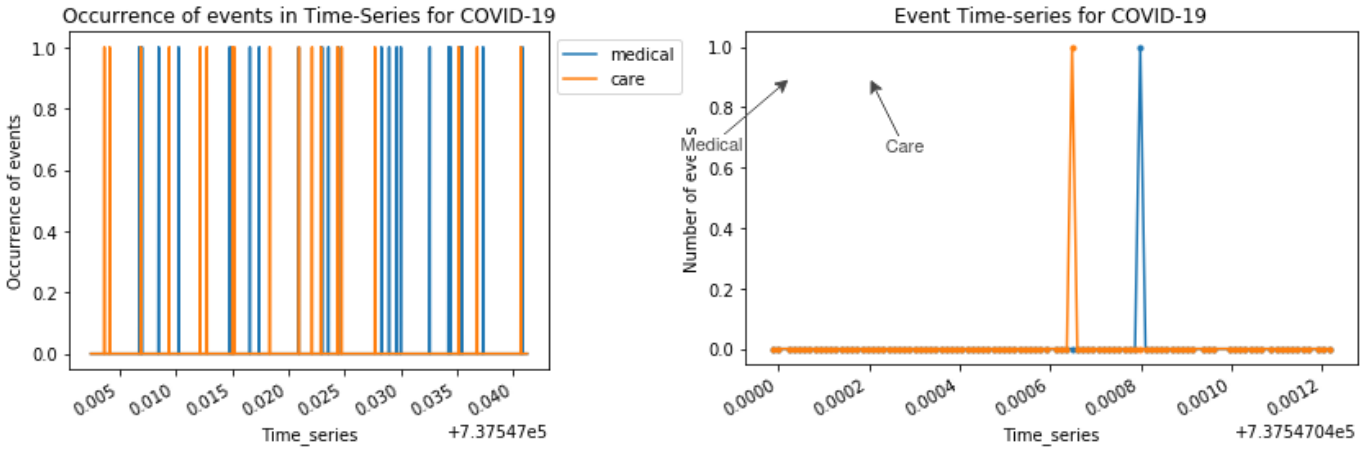

图 4. 2019 冠状病毒病时间序列中各事件的出现情况

算法 1： 主题排序算法

输入推文和查询
输出 排序后的主题上下文向量及其对应的事件
当 $t\ in\ tweets$ **时执行**
    $TS_i = t$
/* 提取数据集中存在的那些事件 */
$E \leftarrow extract_event(query)$
$E_k \leftarrow extract_keywords(TS_i)$
/*将 Ulist 定义为上下文事件的不确定性列表事件 */
当 $e\ in\ E_k$ **时执行**
    当 $k\ in\ E_k$ **执行时**
        如果 $cos\ (k,\ e)\ !\ =0$ **那么**
            计算 $P(k|e)$ /* 追加以下内容的不确定性：Ulist 中包含关键词 k 的情境事件 e */
            $Ulist.append(Un(k|e))$
/*获取以下内容中情境事件的主题排名：Ulist 及其权重*/
获取_rank(Ulist)

Twitter 数据集详情表 I

	数据集详情
推文数量	3116
平均每条推文词数	10
数据集中的词语总数	4729
开始日期时间	2020-05-02 00:59
结束日期时间	2020-05-02 00:03

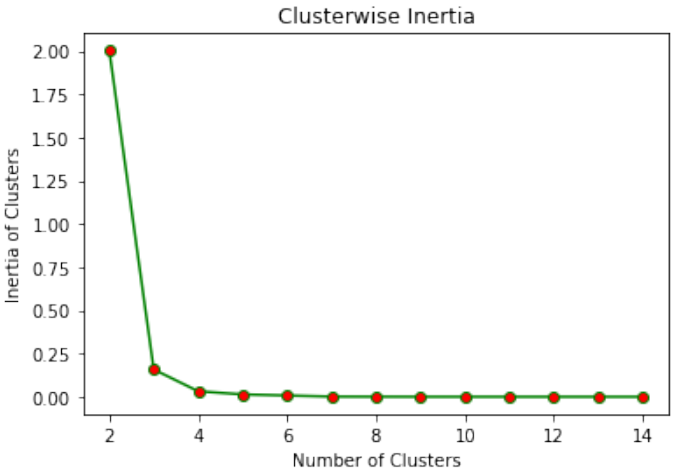

图 5. 按簇惯性

V. 结果讨论

对于推文，我们考虑使用查询来提取主题上下文向量。例如，我们考虑了示例查询“关于医疗护理有哪些已发表的内容？”。在此查询中，“医疗”和“护理”是事件。上下文信息将根据推文中的相关词语得出。下表列出了针对上述两个事件识别出的关键词及其不确定性值。这些事件分为两类：确定和不确定。我们将借助不确定性值接近 0 的关键词来识别确定事件。

“医疗”和“护理”事件的主题上下文向量为：

- CV_1 = {(病毒,0.007), (紧急情况,0.004), (封锁, 0.005), (药物,0.001), (护理,0.001), (设施,0.003)}
- CV_2 = {(通常，1.327308), (不再，1.32914), (恐慌, 1.32757), (政府，1.246714), (私人的，1.238603), (政治的, 1.12092)}

表 I I 两个事件的已识别关键词及其不确定性值

确定事件				不确定事件			
医疗		Care		医疗		Care	
关键词	排名权重	关键词	排名权重	关键词	排名权重	关键词	排名权重
病毒	0.007	意外事件	0.000523	通常	1.327308	提供商	1.321089
紧急情况	0.004	大流行病	0.029304	不再	1.32914	服务	1.322659
封锁	0.005	恐慌	0.000785	恐慌	1.32757	大约	1.321089
药物	0.001	医疗	0.006018	政府	1.246714	rest	1.237301
护理	0.001	研究人员	0.002093	私人	1.238603	疏忽	1.321089
设施	0.003	news	0.014129	政治的	1.12092	heal	1.235993

主题上下文向量 CV_1 显然比 CV_2更能定义医疗事件的上下文。类似地，对于护理事件，我们得到：

- CV_3 = {(意外事件,0.000523), (大流行病,0.029304), (恐慌, 0.000785), (新闻,0.006018), (医疗,0.002093), (研究人员, 0.014129)}
- CV_4 = {(提供商, 1.321089), (服务, 1.322659), (约,1.321089), (其余, 1.237301), (疏忽, 1.321089),(治愈, 1.235993)}

主题上下文向量 CV_3 比 CV_4更能确定事件护理的上下文。表 III 展示了 COVID-19 数据集中各种事件的前 10 个确定性主题上下文向量。

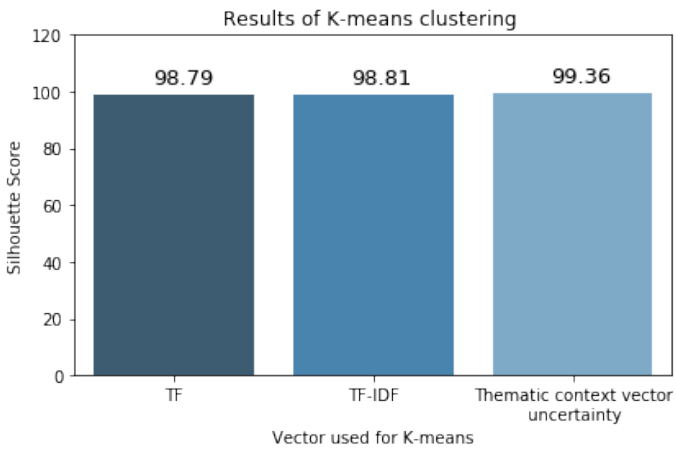

图 6. K 均值聚类结果

图 V 展示了使用词频、词频-逆文档频率和主题上下文向量进行聚类的结果。使用主题上下文向量进行聚类的效果较两种最先进的方法平均提升了 0.5%。

VI. 结论

Twitter 数据时间序列中的关键信息是一组与情境事件相关的关键词。主题上下文向量是从 Twitter 数据中提取的，可提供事件的上下文。查询用于查找事件的上下文。不确定性原理用于对事件中的主题向量进行排序。对于数据中彼此相关的确定事件，主题上下文向量的不确定性值为 0。使用所提出的方法进行事件聚类时，轮廓系数的平均提升幅度比使用词频和词频-逆文档频率方法进行聚类高 0.5%。在上下文关系方面，使用关键词与相应事件之间的条件概率，事实证明这种方法有效。该方法可用于自动生成推文中的事件。在网络欺凌、反讽检测、语言模型等应用中，也可使用该系统识别出的主题上下文向量。

参考文献

[1] 统计数据见。 https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day. [2] 互联网使用情况统计见。 https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read.[3] Ghaith Abdulsattar A Jabbar Alkubaisi、Siti Sakira Kamaruddin 和 Husniza Husni。基于混合朴素贝叶斯分类器，对 Twitter 上的情感进行分析的股票市场分类模型。计算机与信息科学，11(1)：52–64，2018年。[4] Thayer Alshaabi、Joshua R Minot、Michael Vincent Arnold、Jane Lydia Adams、David Rushing Dewhurst、Andrew J Reagan、Roby Muhamad、Christopher M Danforth 和 Peter Sheridan Dodds。世界集体注意力如何投向一场大流行病：Twitter 上 24 种*12614*语言的 COVID-19 相关一元词时间序列，2020年*2003*。*arXiv* 预印本 *arXiv:* 。[5] David Alvarez-Melis 和 Martin Saveski。Twitter 主题建模：按对话聚合推文。载于第十届 *AAAI* 网络与社交媒体国际会议，2016年。[6] Stathis G Arapostathis 和 Marianthi Karantzia。利用志愿地理信息来源（Twitter）绘制火灾事件信息，以有效开展灾害管理（希腊）；阿提卡东北部火灾案例研究，2017年8月，（希腊）。载于阿拉伯地球科学杂志会议，第257–260页。Springer，2018年。[7] `Angels Mart´ınez Arias、Manuel Bernal-Sprekelsen、Elena Rioja、Joaquim Ense˜nat、Alberto Prats-Galino 和 Isam Alobid。经鼻翼突内镜入路及蝶窦脑膜脑膨出切除后的颅底修复：我们的经验。*Acta Otorrinolaringol´ogica Espa˜nola*，66(1)：1–7，2015年。[8] Vahid Behzadan、Carlos Aguirre、Avishek Bose 和 William Hsu。用于收集 Twitter 流中网络威胁指标的语料库与深度学习分类器。载于*2018*年 *IEEE* 国际大数据会议（*Big Data*），第5002–5007页。IEEE，2018年。[9] Erman C.ak、Waldemar Karwowski 和 Les Servi。应用软计算技术估计 Twitter® 时间序列数据所表达的情绪状态。《神经计算与应用》，32(8)：3535–3548，2020年。[10] Nikan Chavoshi、Hossein Hamooni 和 Abdullah Mueen。识别 Twitter 上的相关机器人账号。载于社交信息学国际会议，第14–21页。Springer，2016年。

表 III 各类事件中排名前 10 的确定主题上下文向量

	不确定事件									
Rank	**杰出的**	**科学家**	**宣布**	**医疗**	**care**	**存活**	**行动**	**COVID**	**冠状病毒**	**冠状病毒-病毒**
0	公告 水泥	韩国人	宣布了	病毒	意外事件	ask	房屋	COVID	期货	gone
1	long	人们	公告	紧急情况	大流行病	time	特朗普	黑人	艺术家	借口
2	想法	结论	这里是	封锁	恐慌	决定	amp	阻塞	预期	six
3	擒抱	感染的	保障措施	药物	医疗	总统	已采取	已屏蔽	借口	加利福尼亚
4	科学家	不能	重新开放	护理	研究人员	guys	决定	区块	证据	地球
5	放弃了	南方	满意	die	服务	omg	故意地	漂白剂	服务过	生日
6	国家的	两次	决定	dean	see	母亲	new	指责	set	证据
7	made	爆发	may	恐慌	安全地	仁慈	设计的	Blackhorse	地球	bad
8	今天	本地	推出	通常	节省	jose	seem	婊子	Espiritu	nice
9	倡议	导致	最新的	说	跳跃了	Jones	共和党人	啤酒厂	eric	指责

[11] Chao Chen 和 Gabriel Terejanu。Twitter 网络中的子事件检测。载于 *IFIP*人工智能应用与创新国际会议，第50–60页。Springer，2018年。[12] Peter Grindrod。分析学的数学基础：理论与应用。牛津大学出版社，2014年。[13] Dylan Grosz 和 Patricia Conde-Cespedes。自动检测工作场所中常见的性别歧视言论。载于亚太知识发现与数据挖掘会议（*PAKDD*），研讨会（用于聚类的学习数据表示）*LDRC*，2020年。[14] Mahmud Hasan、Mehmet A Orgun 和 Rolf Schwitter。Twitternews：从 Twitter 数据流中实时检测事件。*PeerJ*预印本，4:e2297v1，2016年。[15] Ahmad Hany Hossny、Lewis Mitchell、Nick Lothian 和 Grant Osborne。Twitter 事件检测的特征选择方法：一种文本挖掘方法。社会网络分析与挖掘，10(1)：1–15，2020年。[16] Satya Katragadda、Shahid Virani、Ryan Benton 和 Vijay Raghavan。利用 Twitter 检测事件开始。载于*2016*年国际神经网络联合会议（*IJCNN*），第1539–1546页。IEEE，2016年。[17] Wahab Khan、Ali Daud、Jamal A Nasir 和 Tehmina Amjad。自然语言处理领域最先进机器学习模型综述。《科威特科学杂志》，43(4)，2016年。[18] Simran Khanuja、Sandipan Dandapat、Sunayana Sitaram 和 Monojit Choudhury。一个用于混合代码对话自然语言推理的新数据集。*arXiv* 预印本 *arXiv:2004.05051*，2020年。[19] Olga Kolchyna、Th´arsis TP Souza、Philip C Treleaven 和 Tomaso Aste。Twitter 事件检测与区分框架及其在零售品牌中的应用。载于*2016*年未来技术会议（*FTC*），第323–331页。IEEE，2016年。[20] Eric L Lai、Daniel Moyer、Baichuan Yuan、Eric Fox、Blake Hunter、An- drea L Bertozzi 和 P Jeffrey Brantingham。微博的主题时间序列分析。《*IMA*应用数学杂志》，81 431，2016 3 409( ): – 。[21] Colette Therese Malyack、Krystal M Hunter 和 Starr Roxanne Hiltz。Twitter 与奥斯卡奖得主预测。2020年。[22] Frank Namugera、Ronald Wesonga 和 Peter Jehopio。情感的文本挖掘与决定因素：乌干达传统媒体机构对 Twitter 社交媒体的使用。《计算社会网络》，6(1)：3，2019年。[23] Richard J Povinelli。用于特征描述的时间模式识别。载于时间、空间与时空数据挖掘：第一届研讨会 *TSDM 2000*，法国里昂，*2000*年*9*月*12*日，修订论文集，第46页。Springer，2003年。[24] A Raghuvira Pratap、JVD Prasad、K Praveen Kumar 和 S Babu。基于 Twitter 数据分析优化交通流的研究。载于*2018*年第二届创新通信与计算技术国际会议（*ICICCT*），第320–325页。IEEE，2018年。[25] Vala Ali Rohani、Shahid Shayaa 和 Ghazaleh Babanejaddehaki。社交媒体内容的主题建模：一种实用方法。载于*2016*年第三届计算机与信息科学国际会议（*ICCOINS*），第397–402页。IEEE，2016年。[26] Gonzalo A Ruz、Pablo A Henr´ıquez 和 Aldo Mascare˜no。通过贝叶斯网络分类器分析关键事件期间的 Twitter 数据情感。《未来一代计算机系统》，106：92–104，2020年。

[27] Zafar Saeed、Rabeeh Ayaz Abbasi、Abida Sadaf、Muhammad Imran Razzak 和 Guandong Xu。利用文本流构建时序网络——一种用于检测 Twitter 新兴事件的动态心跳图。载于 亚太知识发现与数据挖掘会议，第 534–545 页。Springer，2018年。[28] Camilla Salvatore、Silvia Biffignandi 和 Annamaria Bianchi。用于构建新社会指标的社交媒体和 Twitter 数据质量。《社会指标研究》，第 1–30 页，2020年。[29] Frederic Stahl、Mohamed Medhat Gaber 和 Mariam Adedoyin-Olowe。社交媒体分析数据挖掘技术综述。《数据挖掘与数字人文期刊》，2014年，2014年。[30] Steven Van Canneyt、Nathan Claeys 和 Bart Dhoedt。Twitter 上主题相关的情感分类。载于 欧洲信息检索会议，第 441–446 页。Springer，2015年。[31] Xinyue Wang、Laurissa Tokarchuk 和 Stefan Poslad。识别用于实时事件检测的相关事件内容。载于 *2014*年*IEEE/ACM*社会网络分析与挖掘进展国际会议 *(ASONAM 2014)*，第 395–398 页。IEEE，2014年。[32] Selene Yue Xu 和加州大学伯克利分校。利用 Yahoo Finance 和 Google Trends 信息预测股价。加州大学伯克利分校，2014年。[33] Garima Yadav、Mansi Joshi 和 R Sasikala。Twitter 数据分析：结合时间与词频分析进行实时事件检测。载于 *IOP*会议系列：材料科学与工程，第 263 卷，文章编号 042081，2017年。[34] Madani Youness、Erritali Mohammed 和 Bengourram Jamaa。利用大数据技术进行 Twitter 数据分类。载于 *2018*年互联网与电子商务国际会议论文集，第 124–129 页，2018年。