A Twitter narrative of the COVID-19 pandemic in Australia

Rabindra Lamsal*
The University of Melbourne
rlamsal@student.unimelb.edu.au

Maria Rodriguez Read
The University of Melbourne
maria.read@unimelb.edu.au

Shanika Karunasekera
The University of Melbourne
karus@unimelb.edu.au

ABSTRACT

Social media platforms contain abundant data that can provide comprehensive knowledge of historical and real-time events. During crisis events, the use of social media peaks, as people discuss what they have seen, heard, or felt. Previous studies confirm the usefulness of such socially generated discussions for the public, first responders, and decision-makers to gain a better understanding of events as they unfold at the ground level. This study performs an extensive analysis of COVID-19-related Twitter discussions generated in Australia between January 2020, and October 2022. We explore the Australian Twitterverse by employing state-of-the-art approaches from both supervised and unsupervised domains to perform network analysis, topic modeling, sentiment analysis, and causality analysis. As the presented results provide a comprehensive understanding of the Australian Twitterverse during the COVID-19 pandemic, this study aims to explore the discussion dynamics to aid the development of future automated information systems for epidemic/pandemic management.

Keywords

Crisis informatics, Situational awareness, Topic modeling, Granger causality, Network analysis

INTRODUCTION

The gravity of the COVID-19 pandemic made people more vocal on social media platforms, especially on microblogs such as Twitter. Twitter discussions, i.e. tweets, specific to the pandemic, collected by researchers and laboratories globally, have been reported to be in billions (Chen et al. 2020; Lamsal 2020; Imran, Qazi, et al. 2022; Lamsal, Read, et al. 2023), with the United States, United Kingdom, Canada, and India generating the most discussions for both English-only and multilingual discourse. The Australian Twitterverse also seemed significantly vocal towards the pandemic; (Lamsal 2020; Imran, Qazi, et al. 2022) report Australia's presence within the top 10 countries generating the most discussions on Twitter regarding the pandemic. The pandemic introduced numerous events within Australia in 2020–22 — notable ones include over a hundred film and TV show productions getting halted, official interest rates cut to a record low, the introduction of economic stimulus packages from federal and state governments, a recession for the first time in nearly three decades, agriculture workers shortage, local council elections getting disrupted, protests against lockdown restrictions and the national vaccine program, and the suspension of sporting events. Discussions related to these events and numerous global topics specific to the pandemic have been trending on the Australian Twitterverse since the COVID-19 outbreak.

Social media and situational awareness

The main reasons for the growth of social media are speed, transparency, and ubiquity, assisted by rapid developments in mobile technology. The events that would have remained obscure for a long course are now being reported and shared worldwide within seconds (Mayfield III 2011). Social media platforms provide an edge over traditional

*corresponding author

ways, such as individual cellular communications, by providing a public broadcast platform for individuals. Such broadcast platforms help engage people with the exchange of status updates, stories, and media items, which have been reported to have a significant advantage for extracting "situational awareness" (Imran, Castillo, et al. 2015; Lamsal, Harwood, et al. 2022a). During mass emergencies, compared to normal hours, people tend to make use of social media excessively to get situational updates (Vieweg 2012) and such socially generated discussions accumulate to hundreds of thousands and even millions in cases such as the COVID-19 pandemic (Lamsal, Harwood, et al. 2022a). Those discussions, if timely monitored, processed, and analyzed, can contain actionable information relating to the event (Hughes and Palen 2009; Vieweg et al. 2010; Vieweg 2012) that can assist first responders and decision-makers to come up with efficient plans for effective disaster management. Literature in the crisis computing discipline (Lamsal, Harwood, et al. 2022a; Imran, Castillo, et al. 2015) summarizes the effectiveness of social media discussions for numerous humanitarian aid-related tasks, with "situational awareness" as one of the most evident.

This paper presents an extensive retrospective Twitter narrative of the Australian Twitterverse during the COVID-19 pandemic by employing state-of-the-art supervised and unsupervised machine learning approaches. We perform (i) network analysis on hashtags and mentions, (ii) topic modeling with neural embeddings, (ii) sentiment analysis with a transformer-based language model, and (iv) causal analysis with Granger causality tests. Through the study of hashtags and mentions networks, we seek to distinguish the significance of country-level and state-specific hashtags and identify the classifications of Twitter accounts that generate the most engagement during a pandemic. With topic modeling and sentiment analysis, we aim to extract topics (events) that generate the highest tweet interest and investigate the sentiment trends during different pandemic phases. And with causal analysis, we seek to study the causality behavior of discussion-based time series on the confirmed cases and death cases time series. These analyses, combinedly, contribute to the during-disaster and post-disaster phases of disaster management, and the results assist in understanding the conversational dynamics of a pandemic to aid the development of future information systems for epidemic/pandemic management.

Contributions: As per our knowledge, this study, with a timeline of almost 145 weeks (January 2020 to October 2022), provides the most comprehensive social media narrative of the COVID-19 pandemic in Australia through multiple analyses. In doing so, presented results aid in identifying areas with room for improvements for designing robust automated information systems for epidemic/pandemic management. Further, we release a large-scale geotagged tweets dataset[1], curated as a part of this study using the *Full-archive search endpoint*[2] (this endpoint returns the entire volume of historical tweets).

LITERATURE REVIEW

Information systems researchers and practitioners have been formulating frameworks and tools to monitor, collect, analyze, summarize, and visualize social media data to assist in making timely and effective decisions during crisis events (Imran, Castillo, et al. 2015). The current literature heavily relies on microblog platforms, especially Twitter and Weibo, as the primary data source for designing such frameworks and tools — these platforms have large active user bases, their contents carry real-time attributes, and they provide multiple API endpoints for easy access to their public feed. Previous studies have corroborated the relevance of Twitter discussions in the management and analysis of emergency situations during all three phases of a disaster (Martinez-Rojas et al. 2018) — especially, designing emergency monitoring, event detection and decision support systems, identifying relevant contents, performing rapid assessments, and visualizing the spatial and temporal contexts of disasters. Refer to these studies (Vieweg 2012; Imran, Castillo, et al. 2015; Lamsal, Harwood, et al. 2022a) for a thorough review of the literature related to the use of social media data for "situational awareness" and associated methods, data sets, and algorithms.

Researchers have been collecting and sharing large-scale Twitter datasets to enable further research in better understanding the COVID-19 discourse. Tweets in (Chen et al. 2020; Lamsal 2020; Banda et al. 2021; Imran, Qazi, et al. 2022; Lamsal, Read, et al. 2023) are in hundreds of millions and these datasets are some of the largest COVID-19-specific tweets collections at present. These datasets are based on *streaming endpoint* whose payload returns 1% of the entire Twitter data at a particular time. The dataset used in this study is based on *Full-archive search endpoint*, which returns the entire volume of historical tweets. Our dataset also complements *MegaGeoCOV* (Lamsal, Harwood, et al. 2022b) with the use of additional keywords and an extended collection period.

Topic modeling has been rigorously performed to identify latent topics across multiple thematic areas of Twitter discourse, including public health (Ghosh and Guha 2013), sporting events (Steinskog et al. 2017), and disasters (Alam et al. 2018). Multiple studies (Boon-Itt, Skunkan, et al. 2020; Xue et al. 2020; Abd-Alrazaq et al. 2020;

[1] https://dx.doi.org/10.21227/42h1-ge40
[2] https://developer.twitter.com/en/docs/twitter-api/tweets/search/quick-start/full-archive-search

Lamsal, Harwood, et al. 2022b) have performed topic modeling on COVID-19-specific Twitter discussions to explore the public perception of the pandemic and uncover the trends and themes of concerns tweeted by individuals. The majority of the existing studies use *bag-of-words*-based techniques (Lamsal, Harwood, et al. 2022a), which consider documents as bag-of-words and model individual documents as a mixture of latent topics. The bag-of-words representation fails to capture the true semantics of words, leading to a possibly imprecise representation of documents. Recent progress in natural language processing has come through the use of *contextual embeddings*, such as from ELMo, BERT, and GPT-3, which handle the issues associated with semantic similarity and polysemous. As a result, neural embeddings have been used in topic modeling giving rise to neural topic models (Angelov 2020; Zhao et al. 2021; Grootendorst 2022). This study employs neural embeddings-based topic modeling.

Sentiment analysis is one of the most explored research fields for analyzing people's opinions, and attitudes toward factors such as situations, individuals, products, and organizations. Sentiment analysis techniques, in general, are machine learning-based, lexicon-based, and hybrid (Medhat et al. 2014). More recently, transformer-based (Vaswani et al. 2017) models, such as BERT, RoBERTa, and XLNet, are being used for designing high-performing sentiment analyzers. In the case of tweets, BERTweet (Nguyen et al. 2020), a RoBERTa-based (Liu et al. 2019) language model pre-trained on millions of tweets, seems to produce state-of-the-art results in part-of-speech tagging, named-entity recognition, and text classification tasks.

Granger causality (Granger 1969) analysis provides a powerful approach for performing causal inference, by testing whether a time series helps forecast another time series. The current literature employs Granger causality tests on time series data across fields such as neuroscience (Seth et al. 2015), tourism and economy (Dritsakis 2004; Akinboade and Braimoh 2010), stock markets (Bollen et al. 2011), and foreign direct investments (Hoffmann et al. 2005). Twitter discussions have also been studied for their causal behavior towards stock markets (Bollen et al. 2011), cryptocurrencies (Shen et al. 2019), elections (Bovet and Makse 2019), and public health (Lamsal, Harwood, et al. 2022b). In this study, we perform Granger causality tests on time series data generated during topic modeling and sentiment analysis.

For network analysis of social media data — exploring and understanding graphs formed by socially generated data, such as tweets — tools such as Networkx, Gephi, Pajek, and IGraph are available (Akhtar 2014). Twitter discussion-based networks are majorly studied to develop an understanding of online communities (Cheong and Christopher 2011; Abascal-Mena et al. 2015; Hagen et al. 2018; Ahmed et al. 2020; Lamsal 2021). In this study, we perform network analysis of [state→hashtag] and [state→mention] relationships to identify state-specific concerns and highly engaging Twitter accounts, and we use Gephi (Bastian et al. 2009) for graph-based visual analytics.

OVERVIEW OF THE PANDEMIC IN AUSTRALIA

Australia, one of the few countries around the world to adopt the zero-covid "suppression with a goal of no community transmission" public health policy during the COVID-19 pandemic, implemented controls on international travel and response to local outbreaks with stringent lockdowns and thorough contact tracing of local COVID-19 clusters. The country closed its international borders to the outside world for almost two years while imposing strict limits on local movements across its states and territories, thus often being referred to as "Fortress Australia".

Australia's mitigation strategies included early interventions to international travel to reduce transmissions from other countries, suppressing the growth of local COVID-19 clusters with exhaustive contact tracing, early recruitment of contact tracing workers, and use of intense lockdowns. The country had its first public COVID-19 vaccination on February 21, 2021, and by the early last quarter of 2021, 80% of the eligible population (i.e. age≥16) was administered at least a single dose of the COVID-19 vaccine. By the end of March 2022, this percentage increased to 95.0%. The country opened its borders on February 21, 2022, for all fully vaccinated people, and further restrictions on international travel under the Biosecurity Act were lifted on April 18, 2022, thus effectively opening up the country to the world. Although the country's mitigation strategies were in contrast to the ones implemented by other countries and territories worldwide, compared to the United States, the United Kingdom, and European countries, the COVID-19 numbers in Australia have been significantly lower until 2022.

DATA COLLECTION

In this study, we used Twitter's Full-archive search endpoint to collect global COVID-19-specific geotagged tweets created between January 1, 2020, and October 9, 2022. More than 90 keywords and hashtags were used with `has:geo` and `lang:en` operators while querying the endpoint to collect tweets that are geotagged and written in English. The hashtags and keywords are listed in Table 1. We collected 17,826,615 tweets originating from 245 countries and territories worldwide. These tweets are geotagged with either point location or place information. Since the geo attributes for retweets are NULL, the collected data do not include retweets.

Table 1. Keywords and hashtags used for data collection.

coronavirus, #coronavirus, covid, #covid, covid19, #covid19, covid-19, #covid-19, corona, #corona, sarscov2, #sarscov2, sars cov2, sars cov 2, covid_19, #covid_19, #ncov, ncov, #ncov2019, ncov2019, 2019-ncov, #2019-ncov, #2019ncov, 2019ncov, pandemic, #pandemic, quarantine, #quarantine, #lockdown, lockdown, ppe, n95, #ppe, #n95, pneumonia, #pneumonia, virus, #virus, mask, #mask, vaccine, vaccines, #vaccine, #vaccines, lungs, flu, flatten the curve, flattening the curve, #flatteningthecurve, #flattenthecurve, hand sanitizer, #handsanitizer, social distancing, #socialdistancing, work from home, #workfromhome, working from home, #workingfromhome, #covidiots, covidiots, herd immunity, #herdimmunity, chinese virus, #chinesevirus, wuhan virus, #wuhanvirus, kung flu, #kungflu, wearamask, #wearamask, wear a mask, corona vaccine, corona vaccines, #coronavaccine, #coronavaccines, face shield, #faceshield, face shields, #faceshields, health worker, #healthworker, health workers, #healthworkers, #stayhomestaysafe, #coronaupdate, #frontlineheroes, #coronawarriors, #homeschool, #homeschooling, #hometasking, #masks4all, #wfh, wash ur hands, wash your hands, #washurhands, #washyourhands, #stayathome, #stayhome, #selfisolating, self isolating

Table 2. Distributions of tweets based on country, city, and source (global). We list only the top 10 entries.

Country	tweets
United States	8,792,388
United Kingdom	2,840,423
India	1,494,866
Canada	915,229
Australia	481,944
South Africa	406,316
Nigeria	336,515
Ireland	264,582
Philippines	200,404
Pakistan	116,879

City	tweets
Los Angeles, CA	286,393
Manhattan, NY	218,571
New Delhi, India	174,107
Toronto, Ontario	171,569
Mumbai, India	166,001
Chicago, IL	148,641
Florida, USA (state)	144,245
Melbourne, Victoria	142,895
Houston, TX	133,062
Brooklyn, NY	132,873

Source	tweets
Twitter for iPhone	9,568,155
Twitter for Android	6,791,426
Instagram	830,623
Twitter for iPad	313,718
dlvr.it	81,657
Tweetbot for iOS	62,372
Twitter Web App	25,832
Twitter Web Client	15,483
Hootsuite Inc.	15,297
Twitter for Mac	12,516

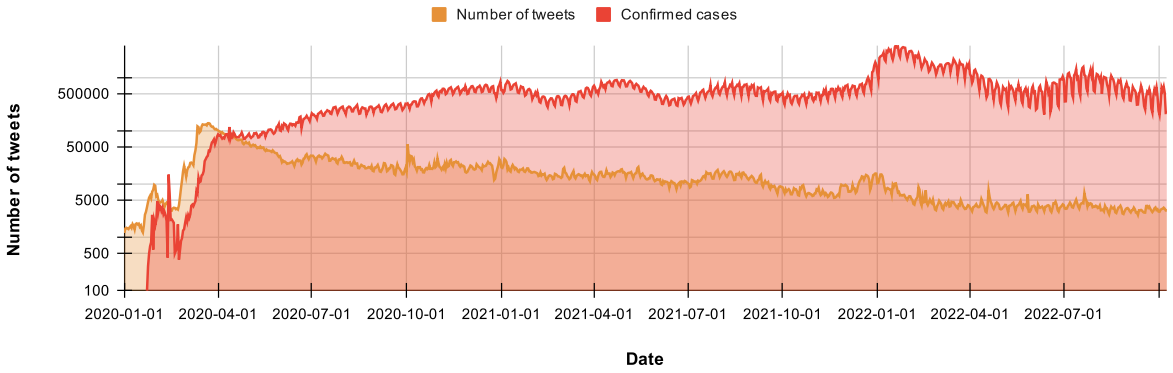


Figure 1. Daily distribution of collected tweets and confirmed cases (worldwide). *Y*-axis is in log scale.

Table 2 gives the distributions (top 10) of tweets based on country, city, and source. Since we collected only English tweets, the dominance of native English-speaking nations is evident. Tweet distribution across countries ranks Australia fifth for generating the most tweets. Melbourne, Victoria is ranked eighth in terms of cities. Twitter native apps for iPhone, Android, and iPad, Instagram, and dlvr.it are the top sources of geotagged tweets. We filtered Australian tweets from the global corpus by conditioning the `geo.country` tweet object. In total, 481,944 tweets in the corpus were identified as originating from Australia.

We performed some basic exploratory data analysis on the collected tweets: Figure 1 and Figure 2 present the daily distribution of tweets alongside the daily confirmed cases in the world and Australia, respectively, Table 3 lists most tagged Australian geolocations, and Figure 3 is a geographical plot of the spatial distribution of tweets in Australia. The state-wise volume of tweets had the following order: Victoria, New South Wales, Queensland, Western Australia, Southern Australia, Australian Capital Territory, Tasmania, and Northern Territory. Melbourne (Victoria) and Sydney (New South Wales) seem to be participating in the discourse significantly compared to other major cities of Australia.

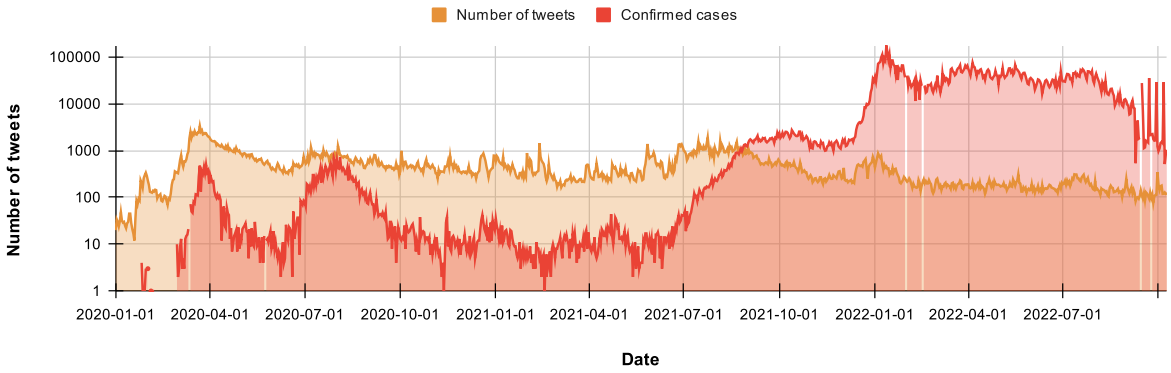


Figure 2. Daily distribution of collected tweets and confirmed cases (Australia). Y-axis is in log scale. Vertical white lines represent adjusted (in negative) values.

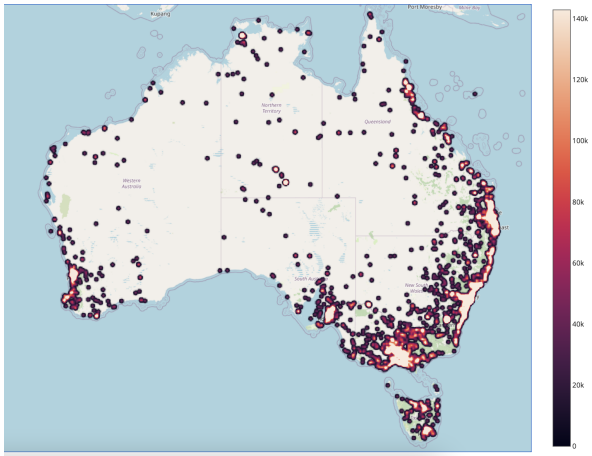


Figure 3. Geoplot showing spatial distribution of COVID-19-specific tweets within Australia. Color scale represents number of tweets.

Table 3. Top Australian regions in the COVID-19 discourse. Based on `geo.full_name` tweet object.

Place	tweets
Melbourne, Victoria	142,895
Sydney, New South Wales	108,511
Brisbane, Queensland	35,176
Perth, Western Australia	29,588
Adelaide, South Australia	21,990
Canberra, Australian Capital Territory	12,283
Gold Coast, Queensland	11,302
Victoria, Australia	7,829
New South Wales, Australia	6,583
Newcastle, New South Wales	5,650
Sunshine Coast, Queensland	4,280
Central Coast, New South Wales	3,736
Tasmania, Australia	3,635
Hobart, Tasmania	3,469

EXPERIMENTS, RESULTS, AND DISCUSSIONS

Data pre-processing

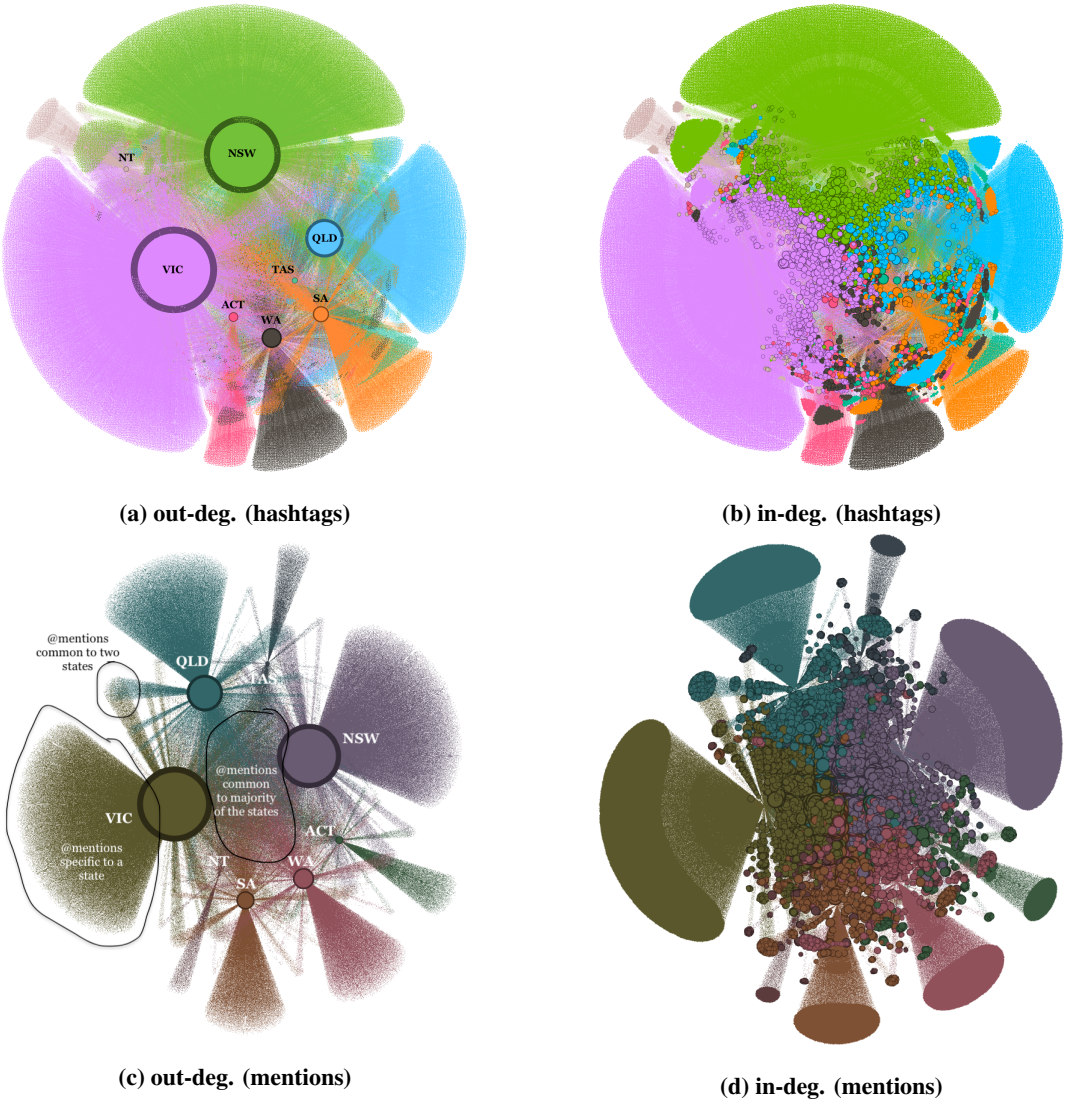
We performed basic text pre-processing tasks on the collected tweets: replacing (i) URLs with the <HTTPURL> token, (ii) HTML entities with their character representation, (iii) emojis with the <EMOJI> token, and removing unnecessary spaces, indentations, and section breaks.

Analyses

Exploring hashtags and mentions usage at the geo-level

Hashtags — keywords or phrases prefaced by the hash sign (#) — have always been a go-to method for people to categorize, search and join discussions related to a particular event. Since the outbreak, many COVID-19-specific keywords and phrases evolved and were in use while referencing the pandemic. The Twitterverse received hundreds of hashtags related to the pandemic, some notable and globe-specific ones include #coronavirus, #covid19, #sarscov2, #pandemic, #quarantine, #flattenthecurve, #handsanitizer, #workfromhome, #herdimmunity, #stayhomestaysafe, #frontlineheroes, #homeschooling, and #faceshields, among others. Country/territory-specific hashtags were also in use. Therefore, with a network analysis, we seek to identify hashtags that were specifically used in each of the Australian states to discuss pandemic-related situations.

The [state→hashtag] relationships form a dense block of nodes representing the hashtags common to multiple states, while state-specific hashtags are represented by sparsely connected blocks of nodes. Timely construction of similar `[county→hashtag]` and `[district→hashtag]` networks and mining of tweets based on these



(a) out-deg. (hashtags) **(b) in-deg. (hashtags)**

(c) out-deg. (mentions) **(d) in-deg. (mentions)**

Figure 4. Visualization of hashtags and mentions networks.

relationships can assist in understanding region-specific concerns. Similarly, we also perform network analysis of `[state→mention]` relationships to identify the most notable Twitter accounts in the Australian COVID-19 discourse. Such highly-engaging accounts can assist in the timely dissemination of factual information and aid in fighting misinformation, to a significant extent.

We deployed a planet-level geocoding endpoint powered by OpenStreetMap data[3] to extract state information for each place name in `geo.full_name` tweet object. This step was necessary to normalize place names such as "Melbourne, Victoria" and "Victoria, Australia". In total, 317,158 `[state→hashtag]` relationships were generated for the hashtags network and 523,857 `[state→mention]` relationships for the mentions network. The hashtags network had 65,022 nodes and 86,352 edges, while the mentions network had 123,220 nodes and 164,978 edges. The visual representations of the networks, in terms of out-degree and in-degree relationships, are shown in Figure 4. State-specific out-degree information is provided in Table 4. Results from the out-degree analysis show that Victoria used the highest number of (unique) hashtags and mentions, followed by New South Wales and Queensland. The in-degree analyses on hashtags and mentions are summarized in Table 5, Table 6, and Table 7. For a detailed outlook on the most common hashtags, Table 5 provides in-degree (suggesting the number of states that used the hashtag) and weighted in-degree (suggesting the overall usage) information. Hashtags such as #covid-19, #lockdown, #stayhome, #morrison are common to multiple states; therefore, such hashtags are assigned to states where they were prominent. The mentions network also follows the same notion.

[3]https://www.openstreetmap.org/

Table 4. Out-degree information of Australian states in the hashtags and mentions networks.

	#hashtags network		@mentions network	
State	**out-degree**	**Wgt. out-degree**	**out-degree**	**Wgt. out-degree**
Victoria (*VIC*)	29,387	122,745	53,809	173,948
New South Wales (*NSW*)	26,071	101,770	46,001	138,221
Queensland (*QLD*)	12,912	41,335	26,343	78,578
Western Australia (*WA*)	6,641	20,001	15,263	37,694
South Australia (*SA*)	5,247	14,514	13,072	31,291
Australian Capital Territory (*ACT*)	2,954	8,827	5,556	12,336
Tasmania (*TAS*)	1,662	4,777	3,183	5,971
Northern Territory (*NT*)	1,478	2922	1,751	3,307

Table 5. Most common hashtags in Australia during the COVID-19 pandemic. In-degree suggests how many states used the hashtag, weighted in-degree suggests the overall volume, and a hashtag assigned to a particular state suggests that the hashtag was used prominently in that state compared to the rest.

State	**#hashtags** with (in-degree:weighted in-degree) information
VIC	**covid19** (8:29,646), **lockdown** (8:5,294), **covid19aus** (8:3,702), **covid19vic** (8:3,479), **stayhome** (8:2,842), **melbourne** (8:2,548), **stayathome** (8:1,627), **covidiots** (8:1,411), **staysafe** (8:1,376), **covid–19** (8:1,370)
NSW	**coronavirus** (8:11,033), **covid** (8:5,348), **australia** (8:2,996), **covid19australia** (8:2,378), **socialdistancing** (8:2,191), **covid19nsw** (8:1,967), **pandemic** (8:1,561), **sydney** (8:1,504), **quarantine** (8:1,146), **wfh** (8:1,114)
QLD	**auspol** (8:7,870), **workfromhome** (8:590), **brisbane**(8:391), **qldpol** (8:346), **trump** (8:334), **queensland** (7:325), **sarscov2** (6:311), **usa** (8:301), **covid19qld** (7:285), **love** (7:284)
WA	**perth** (6:301), **wapol** (6:220), **delta** (7:217), **perthnews** (3:198), **morrison** (7:196), **wanews** (3:158), **covid19wa** (5:125), **wa** (8:118), **4corners** (8:105), **australians** (7:104)
SA	**adelaide** (8:391), **7news** (6:256), **southaustralia** (5:196), **covid19sa** (6:136), **saparli** (4:94), **covidsa** (4:69), **sa** (7:65), **children** (6:62), **bullshitboy** (4:58), **salockdown** (2:42)
ACT	**breaking** (8:623), **canberra** (7:310), **trumpvirus** (6:170), **wtf** (8:121), **qt** (7:111), **cbr** (2:53), **canberralockdown** (4:52), **actlockdown** (3:47), **thailand** (7:46), **zerocovid** (6:40)
TAS	**travel** (8:316), **politas** (6:192), **covid19tas** (4:156), **tourism** (8:123), **tasmania** (6:96), **hobart** (8:73), **aviation** (6:57), **smhr22** (1:32), **flying** (4:32), **emergency** (6:29)
NT	**covid19vaccine** (8:149), **darwin** (6:53), **coronapocalypse** (7:45), **closetheschools** (7:43), **territorytogether** (1:34), **palmerstonnt** (1:32), **nt** (5:32), **shuttheschools** (6:31), **northernterritory** (6:25), **thoughtoftheday** (3:23)

Table 6. Hashtags that were state-specific, i.e. in-degree=1. Notes: ^{a}represents total number of hashtags with in-degree=1, and ^{b}represents weighted in-degree.

State	**#hashtags**	**#**^{a}	**Wgt.**^{b}
VIC	railphotography, railfansofinstagram, yarratrams, tramsoninstagram	20,604	29,438
NSW	shanghailockdown, mydogposts, oliverscampaign, greatersydneylockdown	17,567	24,677
QLD	surfphotography, surflife, wavesfordays, surfinglife, surfrider	7,316	9,574
WA	infreo, perthtwins, rollupforwa, canon5d, keepthebordersclosed	3,290	4,057
SA	adlfest, fiveaa, justbekind, fostercare, kinshipcare	2,641	3,396
ACT	sideastart, lockdownvinyl, cbrlockdown, canberragardener, politicslive	1,240	1,618
TAS	smhr22, proudtobepublic, covid19pakistan, mona, ecodye	603	734
NT	territorytogether, palmerstonnt, sleevesupnt, darwinaustralia	624	808

Results from hashtags network show a significant presence of state-specific hashtags (i.e., in-degree=1); for instance, Victoria had 70%, New South Wales had 67%, and Queensland had 56% state-specific hashtags. Although the respective weighted in-degrees of the state-specific hashtags are lower, their combined presence in the network is significant. Therefore, information systems for epidemic/pandemic management can benefit by starting with a small set of prominent hashtags such as the ones in Table 1 and adding state-specific hashtags incrementally for comprehensive data collection and timely identification of region-specific concerns. Similarly, results from the

Table 7. Most mentions (top 50) in Australia during the COVID-19 pandemic. Notes: [a]represents the state where the mention was prominent, [b]represents in-degree of the mention as a node, [c]represents weighted in-degree.

@mention	Sig.[a]	in-deg.[b]	Wgt.[c]	@mention	Sig.[a]	in-deg[b]	Wgt.[c]
scottmorrisonmp	NSW	8	6,997	9newsaus	NSW	8	951
danielandrewsmp	VIC	8	5,900	patskarvelas	VIC	8	882
gladysb	NSW	8	4,109	prguy17	VIC	8	854
abcnews	NSW	8	3,605	australian	VIC	8	839
realdonaldtrump	VIC	8	3,470	who	NSW	8	833
albomp	NSW	8	2,151	lesstonehouse	QLD	8	783
nswhealth	NSW	8	1,954	noplaceforsheep	NSW	8	652
skynewsaust	VIC	8	1,916	9newsmelb	VIC	7	652
covid_australia	QLD	8	1,892	mikecarlton01	NSW	8	640
annastaciamp	QLD	8	1,796	drtedros	NSW	6	639
greghuntmp	VIC	8	1,753	peterfitz	NSW	8	636
victoriancho	VIC	8	1,529	vicgovdhhs	VIC	6	635
markmcgowanmp	WA	8	1,347	abcmelbourne	VIC	7	623
theage	VIC	8	1,249	bradhazzard	NSW	8	608
newscomauhq	NSW	8	1,231	rafepstein	VIC	8	568
normanswan	NSW	8	1,182	afl	VIC	8	557
smh	NSW	8	1,152	abc730	VIC	8	552
vicgovdh	VIC	7	1,150	domperrottet	NSW	7	550
sbsnews	WA	8	1,137	7newsmelbourne	VIC	7	535
youtube	NSW	8	1,075	samanthamaiden	NSW	8	530
breakfastnews	VIC	8	1,025	sophieelsworth	VIC	7	530
joshfrydenberg	VIC	7	1,018	timsmithmp	VIC	8	524
mjrowland68	VIC	8	980	3aw693	VIC	6	506
theheraldsun	VIC	8	965	billbowtell	NSW	8	502
vanonselenp	NSW	8	955	leighsales	VIC	8	490

mentions network show that accounts belonging to politicians, government bodies, news channels, journalists, radio stations, social activists, public health officials, and health agencies generate the most engagement. Replies to tweets from and tweets with mentions of such engaging accounts seem to include statements of approval, criticism, and request for aid/volunteering, which after filtration of irrelevant content are advantageous for sketching first-hand reports of a situation as it unfolds. Such accounts can also play a vital role during a pandemic in the dissemination of factual information and diminish the flow of misinformation.

Tracking of Australian events and their sentiments

Tracking of Australian events. There were numerous additional topics discussed by the Australian public besides the film industry, economy, finance, workforce, and protests. The timely identification of such topics through social media discussions can be useful in acquiring a better picture of a situation, such that first responders and decision-makers can formulate actionable plans accordingly. In textual data mining, topic modeling is a powerful technique for discovering a set of abstract "topics" from a collection of textual documents where each topic represents an interpretable semantic concept. In this study, the abstract "topics" are the "events" we seek to extract. Topic models can also assist in the screening of tweets specific to humanitarian assistance tasks, such as identifying the demands of a crisis-hit community — the discussions related to "demands" can be further analyzed for planning the timely distribution of relief supplies. We also perform dynamic topic modeling to analyze the evolution of selected topics over time. Topic modeling in a near-real-time scenario can be effectively used as an event detection technique, and tracking the evolution of a topic helps to understand the trend of word usage related to an event.

For the topic modeling task, we used BERTopic (Grootendorst 2022) to take advantage of contextual embeddings from sentence transformers (Reimers and Gurevych 2019) and create clusters of topics through a class-based term frequency-inverse document frequency approach (Grootendorst 2022). We experimented with two settings for *minimum cluster size*: (i) size of 10 for generating highly-specific clusters, and (ii) size of 1000 for generating generalized clusters.

Results from highly-specific clusters show that discussions regarding quarantine, isolation, social distancing, illness, Australian politics, the COVIDSafe app, media, vaccines, sports, TV shows and movies, the second wave, restrictions, and shopping, among others, generated the highest interests. More than 2500 topics were identified, out of which

topics with at least 450 tweets are listed in Table 8. The listed topics are highly-specific—for instance, there were seven clusters related to discussions on vaccines, namely "Vaccine rollout", "Opinions about vaccine", "Vaccines and vaccinations", "Vaccines in NSW", "Vaccine distribution", "Discussion about the AstraZeneca vaccine", and "The Pfizer/BioNTech COVID-19 vaccine". Similarly, results from generalized clusters deduced 42 topics, which are listed in Table 9. Some of the topics in the generalized clusters seem to be highly correlated with each other. We computed cosine similarities of topic embeddings to create a similarity matrix (refer to Figure 5). While generating the similarity matrix, we cluster the topics such that the highly correlated ones form dense-colored blocks in the matrix. The volumetric patterns in topics for the Australian states and territories in the COVID-19 discourse are identical, with a handful of irregularities. For instance, discussions related to the "Ruby Princess cruise" were insignificant in Victoria, the issues on "hairdressing during Lockdown" were insignificant to Queensland, Western Australia, and Tasmania, while discussions around "Gladys and New South Wales" were negligible in Northern Territory. The state-topic-based tweets distribution is provided in Figure 6.

Table 8. Highly-specific clusters-based most discussed topics in Australia during the COVID-19 pandemic. Topics are sorted based on the volume of tweets, and we list topics with at least 450 tweets.

Topic Name ↓(0–33)	**Topic Name** ↓(34–67)
(Self) Quarantine	Vaccines in NSW
Global pandemic	Pets and other animals during the pandemic
The Virus	PPE
(Seasonal) Flu	Productivity, labor force, unemployment
COVID-19 general discussions	Deaths due to COVID-19
Australian politics	Vaccine distribution
Schools and education	Chinese Virus
Social distancing	COVID-19 in NSW
Prime Minister Scott Morrison	Opinions on COVID-19
COVID-19 and live music	Self-isolation
COVID-19 testing	Unproven treatment for COVID-19
COVID-19 and covidsafe app	Mental health issues
Lockdown	Masks
Work from home	Opinions about Australian Open 2021
Lockdown in Melbourne	Coughing and other COVID-19 symptoms
Protests	Italian people
Vaccine rollout	Discussion about the Astrazeneca vaccine
Wearing masks	Coronavirus in India
Hotel quarantine	Tenant rights
Opinions about vaccines	Rat kits and test performance
Sporting events	Caroline Flack
Lockdown in Sydney	Universities and COVID-19
Astrazeneca and Pfizer	TV and film recommendations
Gladys	Lockdown cooking
China	About Aged care facilities and COVID-19
TV shows and movies	The Pfizer/BioNTech COVID-19 vaccine
COVID-19 variants	Health workers and frontline workers
Victorians	World War I and the Spanish flu
Shopping and availability of products	Cruise ships and COVID-19
Hashtags related to Scott Morrison	COVID-19 and children
Got COVID-19	Dan Andrews and Coronavirus
Flu shots and immunizations	State by state latest updates
Vaccines and vaccinations	Easter
Ruby Princess	Ashes, Cricket

We also studied the evolution of keywords through dynamic topic modeling. The study timeline was split into 33 sub-timelines, each representing months between January 1, 2020, and October 9, 2022. We summarize the results from the analysis for selected topics (due to space limitations), namely "Face masks", "Lockdown in Australia", "Sporting events", "Vaccines in Australian context", "Jobs", and "Toilet paper and panic buying", in Table 10. The listed sets of keywords are influential terms describing the respective topics during the mentioned timeline.

Table 9. Generalized clusters-based most discussed topics in Australia during the COVID-19 pandemic. Topics are sorted based on the volume of tweets.

Topic Name	Topic Name
(0) Face mask	(21) Astrazeneca and Pfizer
(1) NSW and Victoria	(22) Flu
(2) Hotel Quarantine	(23) Self-promotion, links, spam
(3) COVID-19 general discussions	(24) Music, Albums
(4) Lockdown	(25) Food, shopping
(5) Vaccines	(26) Dinner and cooking in lockdown
(6) Lockdown in Melbourne and Sydney	(27) Referring to women figures
(7) The Virus	(28) Gladys and New South Wales
(8) Sporting events	(29) Corona
(9) COVID-19 as a pandemic	(30) Lockdown in Victoria
(10) Vaccines in Australian context	(31) COVIDsafe app and tracing
(11) China and Wuhan	(32) Hygiene
(12) Work from home	(33) COVID-19 and India
(13) Scott Morrison	(34) Wear a mask
(14) Vaccinated	(35) Cruise ships, infection, and outbreak
(15) COVID-19 testing	(36) TV shows and movies
(16) COVID-19 and Trump	(37) Haircuts, hairdressing in Lockdown
(17) Schools and education	(38) Protests
(18) Coronavirus	(39) Toilet paper and panic buying
(19) Social distancing	(40) Herd Immunity
(20) Jobs	(41) Vax, vaxxed and anti

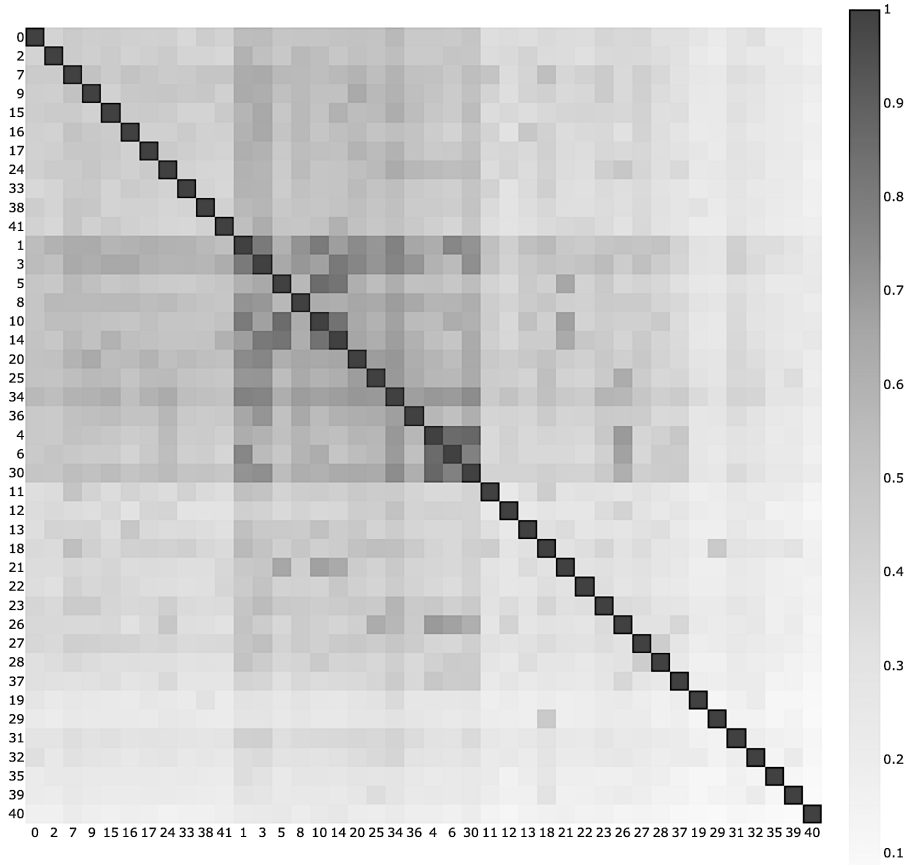


Figure 5. Similarity matrix based on the Cosine similarity. X- and Y-axis represent topics. Highly correlated topics appear near each other in the matrix forming dark-colored blocks. Color scale represents similarity score. Refer to Table 9 for topic names.

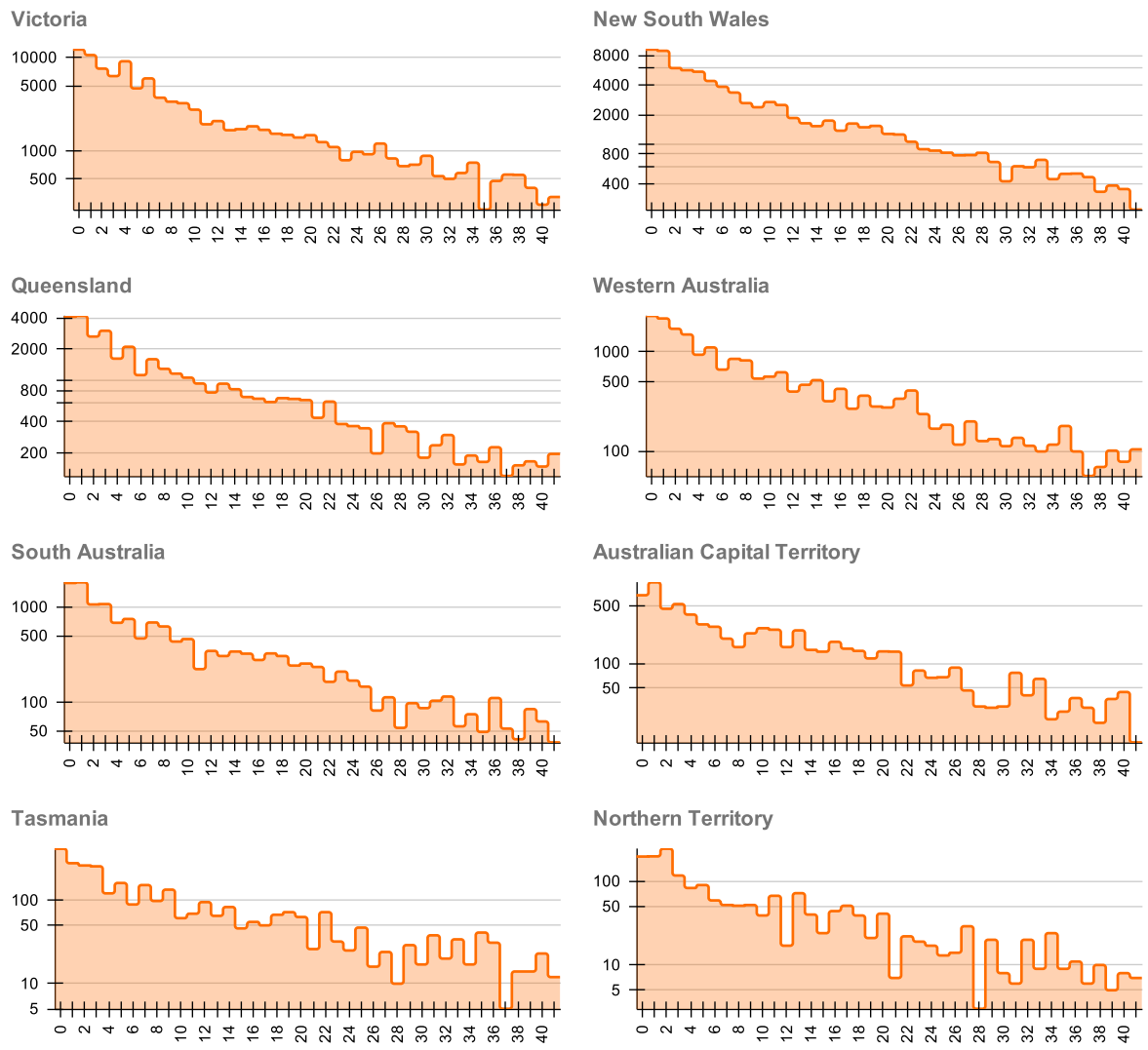


Figure 6. State-topic-based distribution of tweets. For all sub-figures, Y-axis represents the number of tweets and is in log scale, and X-axis represents topics. For topic names refer to Table 9.

Learning the sentiments of people. People share their opinions and feelings regarding various dynamics of a crisis event. The outbreak followed by lockdowns, curfews, travel restrictions, social distancing, quarantine, and a cumulative rise in confirmed cases and deaths between 2020–2022 affected people both in terms of physical and mental health. Studies related to the pandemic and sentiments have reported a rise in negative feelings and pessimism. Therefore, we investigate the overall sentiment trend of the Australian Twitterverse during different phases of the pandemic. Australia experienced four major COVID-19 waves[4]: (i) March–May 2020, (ii) June–November 2020, (iii) July–December 2021 (Delta wave), and (iv) during 2022–until the end of September 2022 (Omicron wave).

For the sentiment analysis task, we finetuned BERTweet (covid19-base-cased) on SemEval-2017 Task 4 dataset (Rosenthal et al. 2017). The results from sentiment analysis are summarized in Figure 7 and Figure 8. The neutral, negative, and positive sentiment brackets for each state and territory were as follows: Australian Capital Territory [50.98%, 34.80%, 14.22%], New South Wales [49.36%, 39.01%, 11.63%], Northern Territory [48.36%, 39.95%, 11.69%], Queensland [46.29%, 42.25%, 11.46%], South Australia [49.12%, 38.61%, 12.27%], Tasmania [46.91%, 41.19%, 11.90%], Victoria [47.45%, 39.82%, 12.73%], Western Australia [47.71%, 42.20%, 10.10%]. The daily distribution of tweets across states based on their sentiments is shown in Figure 7. There is the presence of significant peaks in tweet interest across all states during the first three waves. The interest, however, does not seem associated with the fourth wave, except for Victoria, where neutral and negative tweets form a peak for a restricted timeline. Overall, the Australian Twitterverse seemed more inclined toward neutral and negative sentiments.

[4]https://www.abs.gov.au/articles/covid-19-mortality-wave

Table 10. Evolution of keywords for selected topics. We list only selected timelines due to space limitations. Keywords are influential terms describing the respective topics during the mentioned timeline.

Topic Name	Keywords
Hotel Quarantine	**January 2020**: quarantine, evacuees, 14, days **March 2020**: quarantine, self **June 2020**: quarantine, hotel **April 2022**: quarantine, travellers, inbound, redundant
Lockdown in Australia	**March 2020**: stayhome, australia, lockdown, sydney, melbourne **July 2020**: melbourne, lockdown, victoria, stage **November 2020**: lockdown, melbourne, adelaide, australia, south **Janurary 2021**: lockdown, brisbane, perth, 6pm **March 2021**: brisbane, lockdown, queensland **April 2021**: perth, lockdown, perthlockdown **November 2021**: lockdown, melbourne, back **April 2022**: lockdown, melbourne, longest **August 2022**: rent, lockdown, nsw, longest, vic
Sporting events	**January 2020**: coronavirus, olympics, tokyo **March 2020**: coronavirus, afl, nrl, season **June 2020**: afl, essendon, game **December 2020**: cricket, scg, test, players **January 2021**: tennis, players, quarantine, cricket **July 2021**: olympics, nrl, players, afl **December 2021**: ashes, djokovic, novak, tennis, covid
Vaccines in Australian context	**March 2020**: vaccine, australia, testing **April 2020**: vaccine, australia, until, vahs **January 2021**: vaccine, australia, vaccines, pfizer **February 2021**: vaccine, australia, vaccines, rollout **August 2021**: nsw, vaccine, vaccines, vaccinated, vaccination **December 2021**: vaccine, booster, vaccinated, vaccines, australia **March 2022**: fourth, vaccine, australia, rolled, updates
Jobs	**January 2020**: coronavirus, recession, global, leads **March 2020**: workers, coronavirus, health **May 2020**: economy, workers, nurses **June 2020**: economy, unemployment, recession, jobs **December 2020**: bill, covid, workers, relief **January 2021**: covid, jobkeeper, money, workers **April 2021**: jobseeker, keeper, supplement, covid, payments **July 2021**: covid, workers, health, pay **January 2022**: workers, economy, health, covid, supply **April 2022**: covid, unemployment, labor, pandemic
Panic buying	**January 2020**: paper, toilet, tootpapercrisis2020 **March 2020**: toilet, paper, coronavirus, toiletpaper, buying **June 2020**: toilet, paper, hoarding, panic, buying

To have a better perspective on the pandemic sentiments, we selected a few topics (due to space limitations) and explored their interests over time. Figure 8 presents sentiment analysis on the selected topics, namely "Hotel Quarantine", "Lockdown in Melbourne and Sydney", "Vaccines in Australian context", "Scott Morrison", "COVID-19 testing", "Schools and education", "Jobs", "COVIDsafe app and tracing", and "Wear a mask". Tweet interests in Figure 8 are computed as relative to the highest point in the plots, similar to the search trends analogy of Google trends. Results show that negative sentiments majorly dominated the discourse. Lockdown-related discussions had more positive sentiments during the first wave; as negative sentiments started to become significant during the early month of the second wave, the topic recorded the highest tweet interest during the third wave. Discussions on hotel quarantine had negative sentiments throughout 2020–early 2022, with statistically significant positive sentiments during the first wave. Discussions on vaccines started gaining tweet interest in early 2021 and attained the second most tweet interest (with negative sentiments) during the third wave. Similarly, discussions related to Scott Morrison with negative sentiments achieved significantly high tweet interest during the third wave. Tweets related to COVID-19 testing also inclined significantly towards negative sentiments recording its highest tweet interest in early 2022. After February 2022, the discourse around the selected subjects shows descending tweet interests.

Our study suggests that the number of topics during a pandemic can be in the thousands as country-, state-, city-, county-, and district-level large-scale and small-scale events accumulate over time. Although the topics from highly-specific clustering were based on a minimum topic size of 10, the tweet corpus had only geotagged tweets, and today <1% of tweets are geotagged. A comparative analysis done by (Lamsal, Harwood, et al. 2022b) reported the daily distributions of full-volume (based on Twitter's *counts API*) and geotagged tweets to have significantly identical patterns; therefore, the identified topics are near-true representations of the events discussed during the pandemic in Australia. Hence, a minimum topic size of 10 helps identify small-scale events. The identification of small-scale events is necessary as they include the concerns of a district or a county. Overall, topical analysis shows that information systems can largely assist management authorities in obtaining a comprehensive situational view of an epidemic or pandemic through the mining of highly-specific topics and performing further analyses — tracking evolutions of keywords, exploring sentiment trends, and studying tweeting interests and causality behavior — inside the topic clusters.

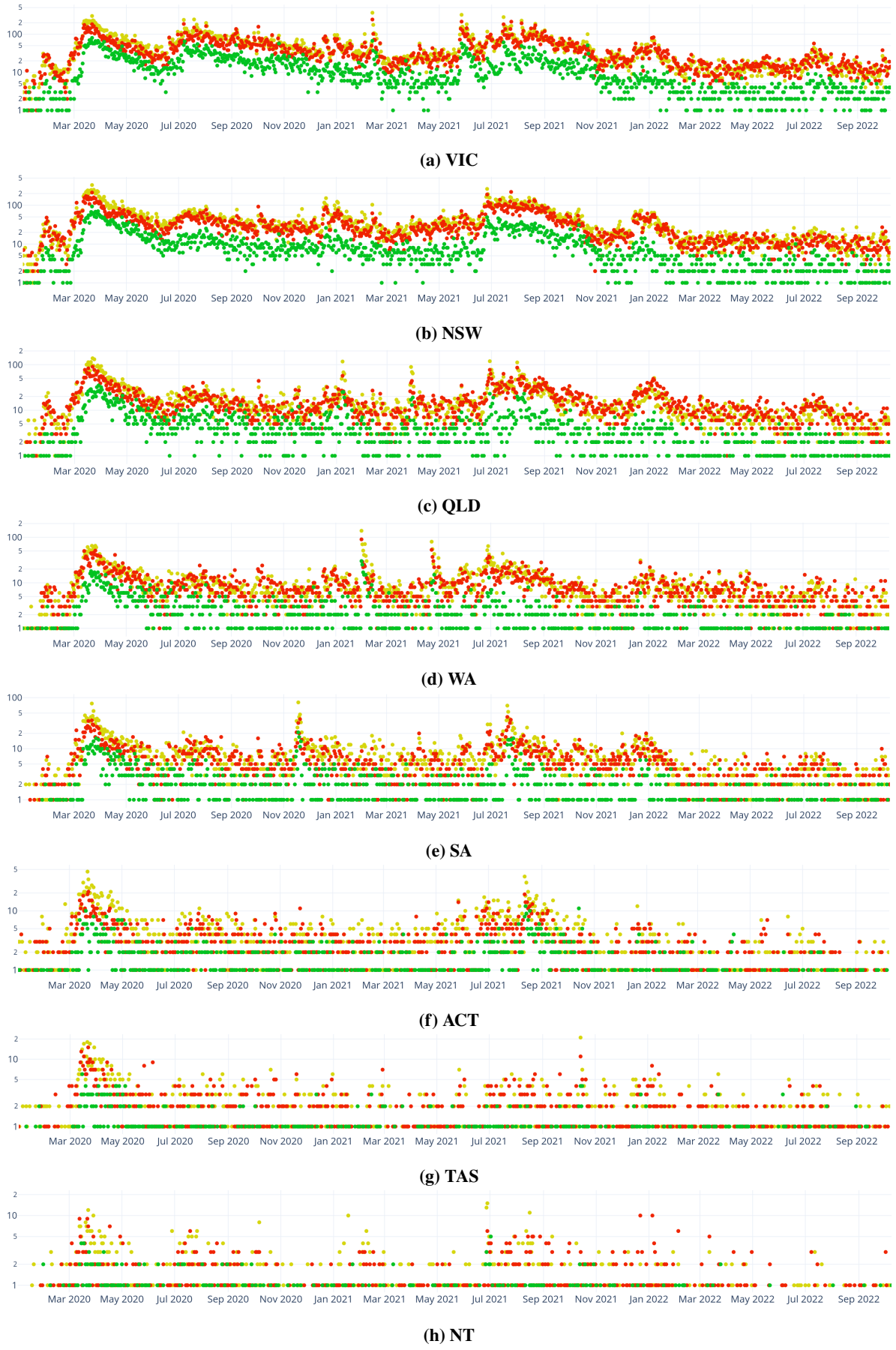


Figure 7. Daily distribution of tweets based on sentiments. *Y*-**axis represents the number of tweets and is in log scale. Yellow, Red, and Green dots represent neutral, negative, and positive sentiments, respectively.**

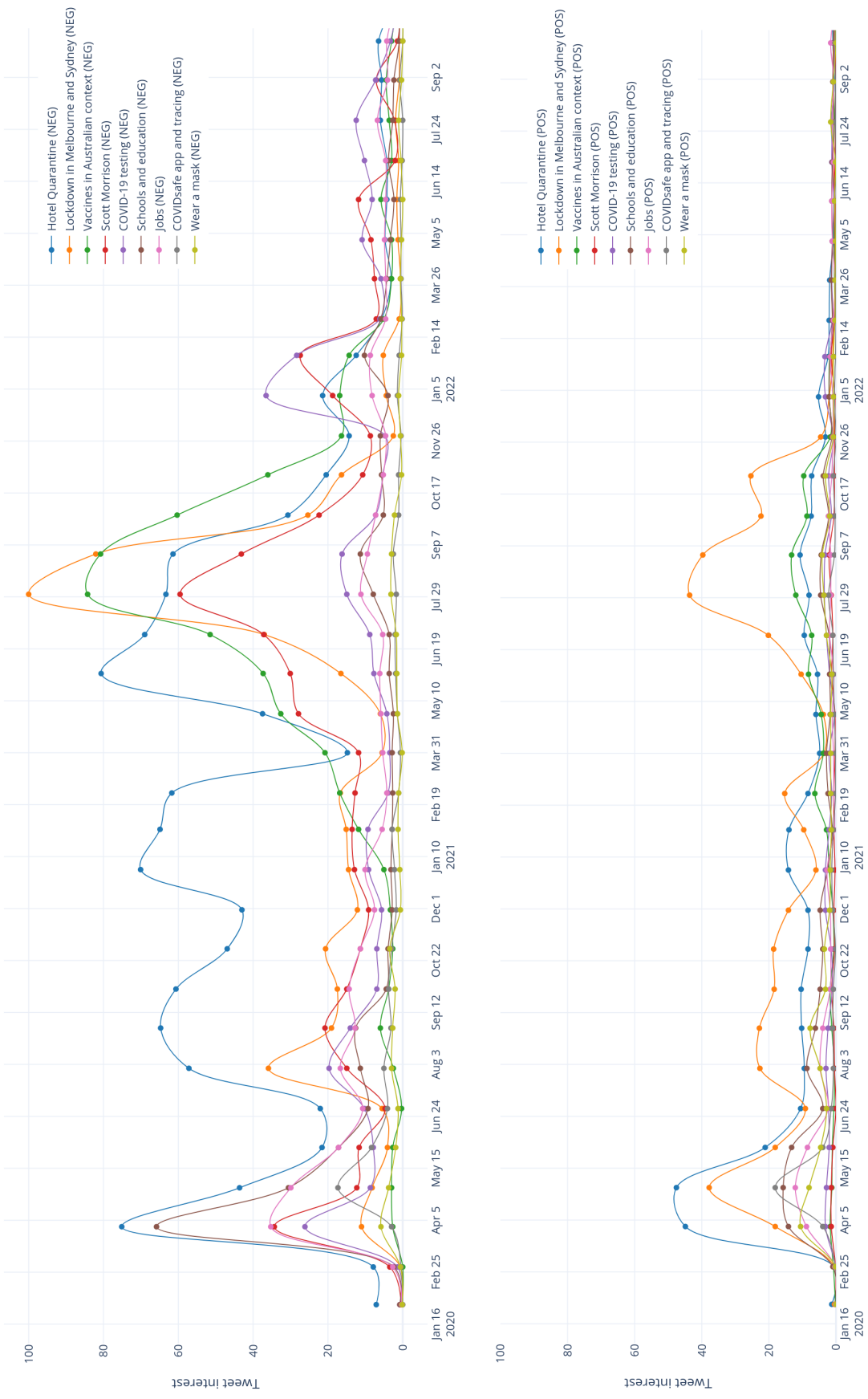


Figure 8. Sentiment analysis on selected topics. The top plot refers to the distribution of negative sentiment tweets, and the bottom plot refers to the distribution of positive sentiment tweets. *Y*-axis is tweet interest — numbers represent interest relative to the highest point on the first plot; therefore, the two plots are comparable.

Table 11. Time series that Granger-cause (at 5% significance) the Australian confirmed COVID-19 cases.

Time series	Discussion cluster	Signif. p-values	Signif. lags
tp^3sn^{NEG}	COVID-19 general discussions	30	4–33
tp^8sn^{NEG}	Sporting events	89	1, 3–90
$tp^{13}sn^{NEG}$	Scott Morrison	34	2, 9–38, 40–42
$tp^{14}sn^{NEG}$	Vaccinated	67	18, 25–90
$tp^{15}sn^{NEG}$	COVID-19 testing	84	7–90
$tp^{25}sn^{NEG}$	Food, shopping	3	9–11
$tp^{36}sn^{NEG}$	TV shows and movies	82	1–3, 5, 13–90
$tp^{41}sn^{NEG}$	Vax, vaxxed and anti	75	8, 16–18, 20–90
tp^3sn^{NEU}	COVID-19 general discussions	27	8–29, 40–43, 46
tp^8sn^{NEU}	Sporting events	85	6–90
$tp^{14}sn^{NEU}$	Vaccinated	8	8–13, 15, 29
$tp^{15}sn^{NEU}$	COVID-19 testing	88	3–90
$tp^{27}sn^{NEU}$	Referring to women figures	15	9–16, 28, 30–35
$tp^{36}sn^{NEU}$	TV shows and movies	76	15–90
$tp^{41}sn^{NEU}$	Vax, vaxxed and anti	72	19–90
$tp^{15}sn^{POS}$	COVID-19 testing	49	5–53
$tp^{36}sn^{POS}$	TV shows and movies	76	15–90
$tp^{41}sn^{POS}$	Vax, vaxxed and anti	86	5–90

Studying the causality behavior of discussion-based time series

Geotagged Twitter discussions have been reported to have latent variables (time series) that Granger-cause the daily confirmed COVID-19 cases (Lamsal, Harwood, et al. 2022b). With a causality analysis of discussion-based time series, we seek to identify a set of time series that Granger-cause the Australian confirmed cases and death cases. For this task, we generated multiple time series based on the volume of tweets over different topics and sentiments. The generated time series dataset took the following form:

$$Time\ series : T^{t^i}_{tp^j sn^k}$$

Where, t^i represents the date component, tp^j represents the topic component, and sn^k represents the sentiment component.

Consider a time series y. Its autoregressive model y_t is:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_n y_{t-n} + e_t \tag{1}$$

Consider another time series x. Now we include the lagged values of x into Equation 1:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_n y_{t-n} + b_s x_{t-s} + ... + b_l x_{t-l} + e_t \tag{2}$$

According to Granger-causality (Granger 1969), x Granger-causes y if lagged values of x in Equation 2 are significant (in F-test). We performed causality tests between y, i.e., confirmed cases and death cases, and each $tp^j sn^k$ for the maximum lags of 90 at a 5% significance level. To identify the variables that Granger-cause the confirmed cases, we considered every $tp^j sn^k$ as independent variables, while for the death cases we also included the daily confirmed cases as an independent variable besides $tp^j sn^k$. The results from causality tests are summarized in Table 11 and Table 12.

Results show the presence of 18 variables (out of 126) that Granger-cause the daily confirmed COVID-19 cases (refer to Table 11). Variables related to Sporting events [negative sentiments], COVID-19 testing [neutral sentiments], Vax, vaxxed and anti [positive sentiments], Sporting events [neutral sentiments], COVID-19 testing [negative sentiments], TV shows and movies [negative sentiments] were observed Granger-causing the confirmed cases for more than 80 lags. Similarly, we identified 9 variables that Granger-cause the Australian COVID-19 death cases (refer to Table 12). Results show that the daily confirmed COVID-19 cases Granger-cause the death cases for all 90 lags, and is followed by COVID-19 testing [negative sentiments] with 55 significant lags, Vax, vaxxed and anti [positive sentiments] with 48 significant lags, COVID-19 testing [neutral sentiments] with 17 significant lags.

Table 12. Time series that Granger-cause (at 5% significance) the Australian COVID-19 death cases.

Time series	Discussion cluster	Signif. p-values	Signif. lags
$tp^{14}sn^{NEG}$	Vaccinated	1	33
$tp^{15}sn^{NEG}$	COVID-19 testing	55	36–90
$tp^{41}sn^{NEG}$	Vax, vaxxed and anti	2	89,90
$tp^{15}sn^{NEU}$	COVID-19 testing	17	44–46, 48–50, 52, 53, 78, 81, 82, 85–90
$tp^{22}sn^{NEU}$	Flu	1	9
$tp^{41}sn^{NEU}$	Vax, vaxxed and anti	2	88,89
$tp^{13}sn^{POS}$	Scott Morrison	2	72, 73
$tp^{41}sn^{POS}$	Vax, vaxxed and anti	48	29–56, 62–66, 76–90
$cases$	-	90	1–90

The variables that Granger-cause the confirmed cases and death cases reveal additional information about their forecasting properties. Most of the variables that Granger-cause the confirmed cases start to provide forecasting power within their 1–2 weeks lag. However, for the variables that Granger-cause the death cases, except for flu and confirmed cases, the explanatory power is evident only after a couple of weeks.

The causality analysis in our study used "volumetric" features of topics discussed during a pandemic. Using "volume" as a feature reduces the computational complexity since we rely only on the volume of tweets based on their topical and sentimental characteristics. Inclusion of the Granger-causing variables, such as the ones listed in Table 11 and Table 12, into forecasting models, have shown improved performance on forecasts compared to models fitted on just the lagged values of the dependent variable (Lamsal, Harwood, et al. 2022b). Pandemic (confirmed and deaths) cases forecasting models fitted on such time series data can be deployed on small-scale infrastructures. Early predictions of the cases help authorities and decision-making bodies to make early estimates of resources to cope with the consequences of future waves of an ongoing epidemic or pandemic.

CONCLUSION

During an ongoing crisis, people use social media as a broadcast platform for disseminating situational updates through exchanges of statuses, stories, and media items regarding what they have seen, felt, or heard. Such conversations, if timely monitored and analyzed, can contain actionable information that can assist first responders and decision-makers in formulating plans for effective disaster management. In this study, we performed an extensive analysis of COVID-19-related Twitter discussions generated in Australia between January 2020, and October 2022, and discussed the significance of such analysis towards the extraction of "situational awareness" concerning a crisis event. We analyzed hashtags and mentions at the state level with in-depth network analysis and performed topic modeling to discover highly-specific topics and generalized topics discussed by the Australian Twitterverse during the pandemic. Next, we explored the conversation dynamics of the Twitterverse across topics and sentiments over temporal and spatial dimensions. Finally, we utilized the knowledge gathered during topic modeling and sentiment analysis to generate numerous discussion-based time series to study the causality behavior of each time series on the Australian COVID-19 confirmed cases and death cases. Overall, we studied the discussion dynamics of the COVID-19 pandemic in Australia to also explore areas that can aid in designing future automated information systems for effective epidemic/pandemic management.

ACKNOWLEDGEMENTS

This study was supported by the *Melbourne Research Scholarship* from the University of Melbourne, Australia. We (the authors) are thankful to *Nectar Research Cloud* (a service of the *Australian Research Data Commons*) for supporting this study with a large-volume compute instance (24 VCPUs, 216GB memory, 20TB volume).

DATA AVAILABILITY

The data collected as a part of this study is available at `https://dx.doi.org/10.21227/42h1-ge40` as an open-access item. We name the dataset `MegaGeoCOV Extended`. A free *IEEE* account is sufficient to access the dataset. The shared tweet identifiers need to be hydrated to re-create the dataset locally. Note that, after hydration, the number of tweets can vary as deleted or private tweets are not retrievable. The dataset includes the following tweet objects for filtering the tweet identifiers: `created_at`, `id`, `author.verified`, `author_id`, `geo.country`, and `source`.

REFERENCES

Abascal-Mena, R., Rose, L., and Sedes, F. (2015). "Detecting sociosemantic communities by applying social network analysis in tweets". In: *Social Network Analysis and Mining* 5.38.

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., Shah, Z., et al. (2020). "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study". In: *Journal of medical Internet research* 22.4, e19016.

Ahmed, W., Vidal-Alaball, J., Downing, J., and López Seguí, F. (May 2020). "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data". In: *J Med Internet Res* 22.5, e19458.

Akhtar, N. (2014). "Social Network Analysis Tools". In: *2014 Fourth International Conference on Communication Systems and Network Technologies*, pp. 388–392.

Akinboade, O. A. and Braimoh, L. A. (2010). "International tourism and economic development in South Africa: a Granger causality test". In: *International Journal of Tourism Research* 12.2, pp. 149–163.

Alam, F., Ofli, F., Imran, M., and Aupetit, M. (2018). "A twitter tale of three hurricanes: Harvey, irma, and maria". In: *arXiv preprint arXiv:1805.05144*.

Angelov, D. (2020). "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470*.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., and Chowell, G. (2021). "A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration". In: *Epidemiologia* 2.3, pp. 315–324.

Bastian, M., Heymann, S., and Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*.

Bollen, J., Mao, H., and Zeng, X. (2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8.

Boon-Itt, S., Skunkan, Y., et al. (2020). "Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study". In: *JMIR Public Health and Surveillance* 6.4, e21978.

Bovet, A. and Makse, H. (Jan. 2019). "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature Communications* 10.

Chen, E., Lerman, K., Ferrara, E., et al. (2020). "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set". In: *JMIR public health and surveillance* 6.2, e19273.

Cheong, F. and Christopher, C. (2011). "Social media data mining: A social network analysis of tweets during the 2010-2011 Australian floods". In: *Proceedings of PACIS*.

Dritsakis, N. (2004). "Tourism as a Long-Run Economic Growth Factor: An Empirical Investigation for Greece Using Causality Analysis". In: *Tourism Economics* 10.3, pp. 305–316.

Ghosh, D. and Guha, R. (2013). "What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System". In: *Cartography and geographic information science* 40.2, pp. 90–102.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438.

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794*.

Hagen, L., Keller, T., Neely, S., DePaula, N., and Robert-Cooperman, C. (2018). "Crisis Communications in the Age of Social Media: A Network Analysis of Zika-Related Tweets". In: *Social Science Computer Review* 36.5, pp. 523–541.

Hoffmann, R., Lee, C.-G., Ramasamy, B., and Yeung, M. (2005). "FDI and pollution: a granger causality test using panel data". In: *Journal of International Development* 17.3, pp. 311–317.

Hughes, A. L. and Palen, L. (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International journal of emergency management* 6.3-4, pp. 248–260.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Computing Surveys (CSUR)* 47.4, pp. 1–38.

Imran, M., Qazi, U., and Ofli, F. (2022). "Tbcov: two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels". In: *Data* 7.1, p. 8.

Lamsal, R. (2020). *Coronavirus (COVID-19) Tweets Dataset*.

Lamsal, R. (2021). "Design and analysis of a large-scale COVID-19 tweets dataset". In: *applied intelligence* 51.5, pp. 2790–2804.

Lamsal, R., Harwood, A., and Read, M. R. (2022a). "Socially enhanced situation awareness from microblogs using artificial intelligence: A survey". In: *ACM Computing Surveys (CSUR)*.

Lamsal, R., Harwood, A., and Read, M. R. (2022b). "Twitter conversations predict the daily confirmed COVID-19 cases". In: *Applied Soft Computing* 129, p. 109603.

Lamsal, R., Read, M. R., and Karunasekera, S. (2023). "BillionCOV: An Enriched Billion-scale Collection of COVID-19 tweets for Efficient Hydration". In: *arXiv preprint arXiv:2301.11284*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Martinez-Rojas, M., Carmen Pardo-Ferreira, M. del, and Rubio-Romero, J. C. (2018). "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review". In: *International Journal of Information Management* 43, pp. 196–208.

Mayfield III, T. D. (2011). "A commander's strategy for social media". In: *Joint Force Quarterly* (60), pp. 79–83.

Medhat, W., Hassan, A., and Korashy, H. (2014). "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4, pp. 1093–1113.

Nguyen, D. Q., Vu, T., and Nguyen, A. T. (2020). "BERTweet: A pre-trained language model for English Tweets". In: *arXiv preprint arXiv:2005.10200*.

Reimers, N. and Gurevych, I. (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084*.

Rosenthal, S., Farra, N., and Nakov, P. (Aug. 2017). "SemEval-2017 Task 4: Sentiment Analysis in Twitter". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 502–518.

Seth, A. K., Barrett, A. B., and Barnett, L. (2015). "Granger Causality Analysis in Neuroscience and Neuroimaging". In: *Journal of Neuroscience* 35.8, pp. 3293–3297.

Shen, D., Urquhart, A., and Wang, P. (2019). "Does twitter predict Bitcoin?" In: *Economics Letters* 174, pp. 118–122.

Steinskog, A., Therkelsen, J., and Gambäck, B. (2017). "Twitter topic modeling by tweet aggregation". In: *Proceedings of the 21st nordic conference on computational linguistics*, pp. 77–86.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vieweg, S. (2012). "Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications". PhD thesis. University of Colorado at Boulder.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1079–1088.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., and Zhu, T. (2020). "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter". In: *PloS one* 15.9, e0239441.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. (2021). "Topic modelling meets deep neural networks: A survey". In: *arXiv preprint arXiv:2103.00498*.