Twitter Speaks: An Analysis of Australian Twitter Users' Topics And Sentiments About COVID-19 Vaccination Using Machine Learning

Stephen Wai Hang Kwok, Sai Kumar Vadde, Guanjin Wang

Submitted to: Journal of Medical Internet Research
on: January 05, 2021

Table of Contents

Twitter Speaks: An Analysis of Australian Twitter Users' Topics And Sentiments About COVID-19 Vaccination Using Machine Learning

Stephen Wai Hang Kwok[1] PhD, RN; Sai Kumar Vadde[1] BSc; Guanjin Wang[2] PhD

[1]Murdoch University Perth AU
[2]Discipline of Information Technology, Media and Communications Murdoch University Perth AU

Corresponding Author:
Guanjin Wang PhD
Discipline of Information Technology, Media and Communications
Murdoch University
90 South St
Perth
AU

Abstract

Background: The novel coronavirus disease (COVID-19) is one of the greatest threats to human beings in terms of healthcare, economy and society in recent history. Up to this moment, there are no signs of remission and there is no proven effective cure. The vaccine is the primary biomedical preventive measure against the novel coronavirus. However, the public bias or sentiments, as reflected on social media, may have a significant impact on the progress to achieve the herd immunity needed principally.

Objective: This study aims to use machine learning methods to extract public topics and sentiments on the COVID-19 vaccination on Twitter.

Methods: We collected 31,100 English tweets containing COVID-19 vaccine-related keywords between January and October 2020 from Australian Twitter users. Specifically, we analyzed the tweets by visualizing the high-frequency word clouds and correlations between word tokens. We built the Latent Dirichlet Allocation (LDA) topic model to identify the commonly discussed topics from massive tweets. We also performed sentiment analysis to understand the overall sentiments and emotions on COVID-19 vaccination in Australian society.

Results: Our analysis identified three LDA topics, including "Attitudes towards COVID-19 and its vaccination", "Advocating infection control measures against COVID-19", and "Misconceptions and complaints about COVID-19 control". In all tweets, nearly two-thirds of the sentiments were positive, and around one-third were negative in the public opinion about the COVID-19 vaccine. Among the eight basic emotions, "trust" and "anticipation" were the two prominent positive emotions, while "fear" was the top negative emotion in the tweets.

Conclusions: Our new findings indicate that some Australian Twitter users supported infection control measures against COVID-19 and would refute misinformation. However, the others who underestimated the risks and severity of COVID-19 would probably rationalize their position on the COVID-19 vaccine with certain conspiracy theories. It is also noticed that the level of positive sentiment in the public may not be enough to further a vaccination coverage which would be sufficient to achieve vaccination-induced herd immunity. Governments should explore the public opinion and sentiments towards COVID-19 and its vaccination and implement an effective vaccination promotion scheme besides supporting the development and clinical administration of COVID-19 vaccines.

(JMIR Preprints 05/01/2021:26953)
DOI: https://doi.org/10.2196/preprints.26953

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users. Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

JOURNAL OF MEDICAL INTERNET RESEARCH

Original Paper

Twitter Speaks: An Analysis of Australian Twitter Users' Topics And Sentiments About COVID-19 Vaccination Using Machine Learning

Stephen Wai Hang Kwok [1], RN, PhD; Sai Kumar Vadde, BSc [2]; Guanjin Wang [3], PhD

[1] School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
[2] Murdoch University, Perth, Australia
[3] Discipline of Information Technology, Media and Communications, Murdoch University, Perth, Australia

Corresponding Author:
Guanjin Wang, PhD
Discipline of Information Technology, Media and Communications, Murdoch University, Perth, Australia
Phone: +61 08 9360 7351
Email: Guanjin.Wang@murdoch.edu.au

Jan 2021

Abstract

Background: The novel coronavirus disease (COVID-19) is one of the greatest threats to human beings in terms of healthcare, economy and society in recent history. Up to this moment, there are no signs of remission and there is no proven effective cure. The vaccine is the primary biomedical preventive measure against the novel coronavirus. However, the public bias or sentiments, as reflected on social media, may have a significant impact on the progress to achieve the herd immunity needed principally.

Objective: This study aims to use machine learning methods to extract public topics and sentiments on the COVID-19 vaccination on Twitter.

Methods: We collected 31,100 English tweets containing COVID-19 vaccine-related keywords between January and October 2020 from Australian Twitter users. Specifically, we analyzed the tweets by visualizing the high-frequency word clouds and correlations between word tokens. We built the Latent Dirichlet Allocation (LDA) topic model to identify the commonly discussed topics from massive tweets. We also performed sentiment analysis to understand the overall sentiments and emotions on COVID-19 vaccination in Australian society.

Results: Our analysis identified three LDA topics, including "Attitudes towards COVID-19 and its vaccination", "Advocating infection control measures against COVID-19", and "Misconceptions and complaints about COVID-19 control". In all tweets, nearly two-thirds of the sentiments were positive, and around one-third were negative in the public opinion about the COVID-19 vaccine. Among the eight basic emotions, "trust" and "anticipation" were the two prominent positive emotions, while "fear" was the top negative emotion in the tweets.

Conclusions: Our new findings indicate that some Australian Twitter users supported infection control measures against COVID-19 and would refute misinformation. However, the others who underestimated the risks and severity of COVID-19 would probably rationalize their position on the COVID-19 vaccine with certain conspiracy theories. It is also noticed that the level of positive sentiment in the public may not be enough to further a vaccination coverage which would be sufficient to achieve vaccination-induced herd immunity. Governments should explore the public opinion and sentiments towards COVID-19 and its vaccination and implement an effective vaccination promotion scheme besides supporting the development and clinical administration of COVID-19 vaccines.

KEYWORDS

COVID-19; vaccination; public topics; public sentiments; Twitter; social media; natural language processing; machine learning; Latent Dirichlet Allocation

Jan 2021

Introduction

Background of COVID-19

The novel coronavirus disease (COVID-19) is an infectious disease caused by a novel coronavirus first identified in Wuhan, China in December 2019 [1]. Till early January 2021, the cumulative number of confirmed cases was 83,862,300, while the number of deaths was 1,837,253, affecting 222 countries or regions globally [2]. In Australia, the total number of confirmed cases was 28,483 and the number of deaths was 909 in early January 2021 [3]. Both the incidence and the prevalence have been rising globally, though these rates differ between countries [4]. In 2020, the pandemic had significant negative impacts on individuals, governments and the global economy [5, 6].

Patients could have either no symptoms, common signs and symptoms of infection, respiratory distress or die from COVID-19. The proportion of asymptomatic patients was estimated at 16%, while the proportion in children was nearly double that of adults [7, 8]. However, over 80% of those asymptomatic had either unilateral or bilateral pulmonary involvement in computerized tomography scans [8]. Among those who were symptomatic, fever, cough and fatigue were the most common symptoms [9, 10]. Five percent of COVID-19 patients developed acute respiratory distress syndrome [11]. Among them, the death rate ranged between 13% and 69% across countries [12].

The viruses could be transmitted through close contact, or even droplets, between individuals, where the mucous membranes of healthy individuals are exposed to secretions produced by the carriers [13]. The reproductive number (R0) of COVID-19 was around three and varied from two to seven across countries [14, 15]. That means one carrier could infect three individuals on average. Under public infection control measures, social distancing seems not applicable to family households where the risk of transmission is high. A meta-analysis of 24 studies found that the intra-family transmission rate of SARS-CoV-2 (the novel coronavirus) was higher than the transmission rate of Severe Acute Respiratory Syndrome Coronavirus or Middle East Respiratory Syndrome Coronavirus in households [16] which may contain vulnerable groups such as elderly, those who are immunocompromised, or suffering from chronic diseases.

Background of Vaccination

Briefly, the purpose of vaccination is to allow the immune system to memorize the features of the pathogen and be able to initiate an immune response that is fast and

strong enough to defeat the live pathogen in the future. Over 115 vaccines for COVID-19 are under investigation and trials, and most of them are targeting the spike protein of SARS-CoV-2 [17]. Usually, the development of a vaccine would take years. The relatively fast development of the COVID-19 vaccine could be ascribed to the previous work on vaccines for SARS-CoV, which has 80% similarity to SARS-CoV-2, and the immense and urgent need for vaccinations [18].

Vaccination which is research evidence-based and officially approved by health authorities, is generally safe. The adverse effects, as well as their incidence rates, vary across types of vaccines. Previous studies have reported the incidence rates of severe adverse reactions after receiving vaccines which are found in general populations. For example, the incident rate of febrile seizures after receiving measles, mumps, and rubella (MMR) and varicella vaccine was 8.5 per 10,000 doses [19]. The rates attributable to influenza vaccines or 13-valent pneumococcal conjugate vaccines (PCV13) were 13 to 45 per 100,000 doses [20]. On the other hand, the incident rate of thrombocytopenic purpura after MMR injection was one per 20,000 doses [19]. Moreover, the incidence rates of some rare diseases such as intussusception after rotavirus vaccine injection were one to five per 100,000 doses [20]. There were insufficient statistics to conclude that vaccination was the direct cause of severe adverse effects compared with the vast majority of those who benefited from vaccinations.

Vaccination is a collective strategy that needs a high proportion of the population to be vaccinated in order to generate the protective effect. The proportion is calculated as (R0-1)/R0 [21]. If one patient could infect three individuals, then the proportion of the population which needs to be vaccinated would be two-thirds. The two-thirds should be individuals who have immune systems functioning normally. Those who are immunocompromised are contraindicated to certain types of vaccines such as live vaccine because of poor responses or severe adverse reactions [22, 23]. Severe allergic reaction to a vaccine is a contraindication though the risk is as small as one per 1,000,000 doses [19]. Hence, the higher the proportion of those who have normal immune systems receiving vaccinations, the better for achieving herd immunity to protect oneself and the others.

Explore Public Opinion On COVID-19 Vaccine

In the last two decades, a prominent wave of anti-vaccination movement which has drawn the most concern was the fall in MMR vaccination coverage and the rise in measles outbreaks in the United States, the United Kingdom and certain major

European countries [24]. A case study article which proposed an association between MMR vaccine and autism [25], though was disproved by several studies in the subsequent years [26-31], was considered as a fuel to the anti-vaccine movement and then was retracted [32]. Nevertheless, the dangerous factors of anti-vaccination might be ignoring high-level evidence such as the results of randomized controlled trials of vaccines [33-35] as well as selective adoption of unverified information in public.

Contemporarily, social media has become a frequently used platform for both authorized information and misinformation to disseminate. Authorized sources such as those from the World Health Organization [36], the U.S. Centers for Disease Control and Prevention [37], the U.S. Food and Drug Administration [38] and the U.K. Department of Health and Social Care [39] are available online. However, previous studies showed that around 30% to 60% of the information related to vaccination on social media were anti-vaccine content [24]. In the websites which provided vaccine information, over 50% contained inaccurate information [40]. Though anti-vaxxers proposed different rationales to oppose vaccination [41], the fact is that only vaccination has a history of successful eradication of viral diseases such as smallpox [42].

As several COVID-19 vaccine trials are progressing to or have nearly completed Phase 3 in the second half of 2020, it is expected that vaccines would be fully available to the markets by 2021 [43, 44]. However, negative news about the vaccine, and anti-vaccine sentiment, could be hurdles to achieving vaccination-induced herd immunity. For example, the University of Oxford and AstraZeneca [45] and Johnson & Johnson [46] had paused their vaccine trials in September and October 2020, respectively, to investigate adverse reactions in participants during the trials, which were resumed investigations. However, information associated with the adverse effects of vaccinations were commonly manipulated by anti-vaxxers to fuel their movements [47]. They had even started touting conspiracy theories against the purpose of developing COVID-19 vaccines as early as the time when the development had not yet begun [48-50]. Therefore, online public opinion and sentiments around COVID-19 vaccination need to be explored and reviewed so as to promote public vaccination scheme based on factors affecting vaccination acceptance.

This study aimed to explore major topics and sentiments on tweets about COVID-19 vaccination among Australian Twitter users. Findings from this study can help governments and health agencies plan, modify and implement a timely promotion of vaccination to achieve vaccination-induced herd immunity.

Jan 2021

Methods

Data Collection

The social media Twitter was chosen as the source of text. Twitter is currently one of the major social media with 187 million daily active users as of the third quarter of 2020 [51]. Twitter is a common source of text for sentiment analysis [52, 53] and analysis of sentiments towards vaccinations [54, 55]. We used the R library “rtweet” [56] to access the Twitter premium API (*Application Programming Interface*) service and collect COVID-19 vaccine-related tweets posted between January 22 and October 20, 2020. Retweets and all those tweets in a language other than English, and with geolocation outside Australia were excluded. The search terms "vacc OR vax OR vaccine OR vaccination" AND "corona OR covid" were used to search target tweets. Boolean operators “AND” and “OR” guaranteed that tweets which contained words belonging to the root of vaccine as well as the root of either coronavirus or COVID could be searched. As a result, 31,100 tweets were collected and used in this study. The number of tweets collected from January 22 to October 20, 2020, are shown in Error: Reference source not found.

Data Pre-processing

The R libraries of "qdapRegex" [57] and "tm" [58, 59] were used for the pre-processing of text. The procedures included (1) removing content which were not English words or were common words which would not provide insights into a specific topic, such as short function words; (2) case folding which changed words into lower case for stemming; and (3) stemming inflected words into roots and then underwent stem completion to return complete words (tokens) for the results visualizations. The custom stop words removed were "amp" (ampersands) and the inflected words derived from vaccine, coronavirus and COVID. Also, stop words in the R library "tm", Python libraries "spaCy" [60] and "gensim" [61], as well as stop words suggested by Sedgewick and Wayne [62] and the SAO/NASA Astrophysics Data System [63], were also removed. The dictionary used for stem completion was a corpus saved before the stemming procedure.

Associations Between Word Tokens

The word tokens were sorted by their counts in the corpus and plotted against their counts as shown in Error: Reference source not found. It was observed that the inflection point of the concave up decreasing curve was located at around 250 counts. Thus word tokens having counts greater than 250 were included in pairwise correlation tests. The R library "widyr" [64] was used to compute the correlations between word tokens. Then, the word pairs with Pearson correlation coefficients

larger than 0.1 were plotted in a network graph (). On the other hand, word pairs were also sorted by their counts and plotted against the counts as shown in Error: Reference source not found. Word pairs having counts larger than 150 were plotted in another network graph (Figure 2).

LDA Tuning and Model Building

Latent Dirichlet allocation (LDA) [65] is an unsupervised machine learning method that allows observations such as words or documents in a corpus to be explained by latent groups such as topics. LDA has been used in topic modeling of public opinions on certain vaccinations against h*uman papillomavirus* (*HPV*) [66] and influenza virus [67]. However, LDA topic modeling on COVID-19 vaccination was yet to be done. The corpus preprocessed was converted into a document-term matrix, and then terms that were sparse by less than 99.9% were retained for LDA modeling. The R library "ldatuning" [68] was used to estimate the optimal number of topics in the LDA model. Four different metrics were computed in a range of topics (2 to 50) to identify the optimal number (Error: Reference source not found). The lower the metrics of "Arun2010" [69] and "CaoJuan2009" [70], and the higher the metrics of "Griffiths2004" [71] and "Deveaud2014" [72], indicated better number of topics to fit a LDA model. In this study, the metric of "Deveaud2014" reached its highest level, and the metric of "CaoJuan2009" reached one of the lowest levels, at three topics which were adopted as the number of topics for LDA modeling. Another R library "topicmodels" [73] was used to estimate the two posterior Dirichlet distributions. They were distribution (theta) over the three topics within each tweet and distribution (beta) over all words within each topic. Only the top 100 words with the highest beta values were visualized in the word cloud for each topic. The words with higher beta values have larger font sizes and a higher level of opacity. In each topic, the top 20 tweets, except news, with the highest theta values larger than those of the other two topics for each tweet, were reported in Error: Reference source not found-8.

Sentiment Analysis

The R library "syuzhet" [74], which applies Stanford's CoreNLP [75] on text against *emotion dictionary*, was used to score each tweet based on the eight emotions and two sentiments defined in the Canadian National Research Council's Word-Emotion Association Lexicon [76, 77]. The eight emotions were anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, while the two sentiments were negative and positive. If a tweet is associated with a particular emotion or sentiment, it would score points that reflect the degree of valence with respect to that category. Otherwise, it would have no score for that category.

Jan 2021

Results

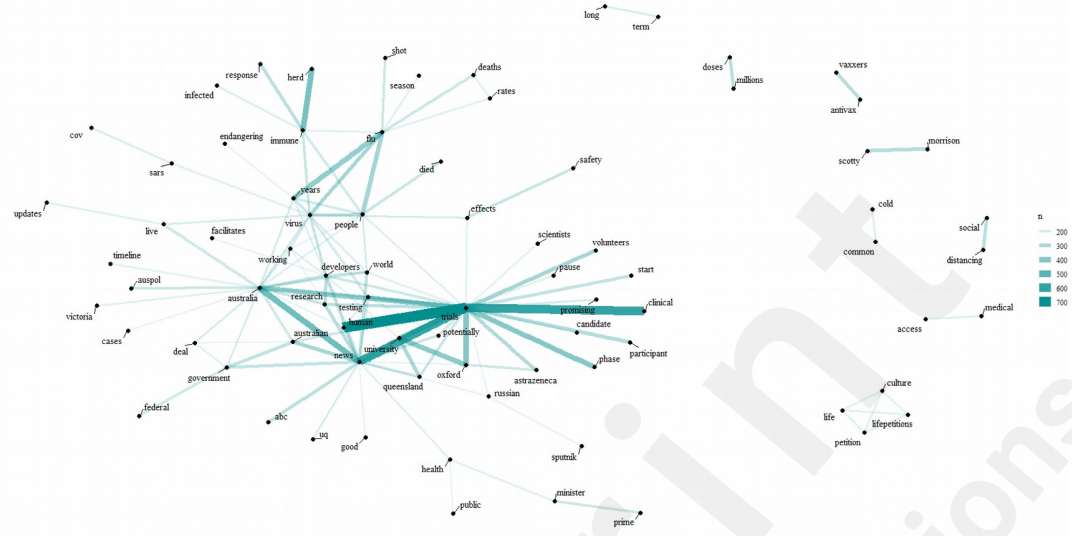
We first analyzed the preprocessed tweets by visualizing the word tokens with a count of over 250 in the corpus, as shown in a word cloud in Figure 1. The larger the word font size in the cloud, the higher number of counts in the corpus. The top ten high-frequency words were "trials", "australia", "virus", "news", "developers", "flu", "people", "years", "world", and "testing". Following that, some other words were also frequently used such as "research", "working", "timeline", "immune", "australian", "effects", "russian", "health", "human", and "government". Based on the descriptive statistics of word counts, the news about the pandemic, seasonal flu, and vaccine trials were among Australians' major discussion topics. Other topics, such as the effects of infection control strategies and immunity, situations overseas, and the government's responses were also relatively prominent.

Figure 1. Word cloud of word tokens with count over 250 in the corpus.

Figure 2 shows the network of word pairs with counts above 150 in the corpus. The word tokens linked with edges where thicker and more opaque lines indicated higher number of counts. From the graph, a group of words which were frequently used together were "trials", "human", "clinical", "news", and "australia". Moreover, the word "trials" was linking to a number of word tokens such as "phase", "australia", "testing", "volunteers", and "university" which was linked to "oxford" and "queensland". Another cluster of words which were commonly used together included "flu", "years", "virus", and "people". The bigrams such as "herd" and "immune" had some associations with "flu" and "virus". There were a few word pairs, such as "antivax" and "vaxeers", which did not connect with the main network and had relatively small number of counts at the periphery of the graph.

Jan 2021

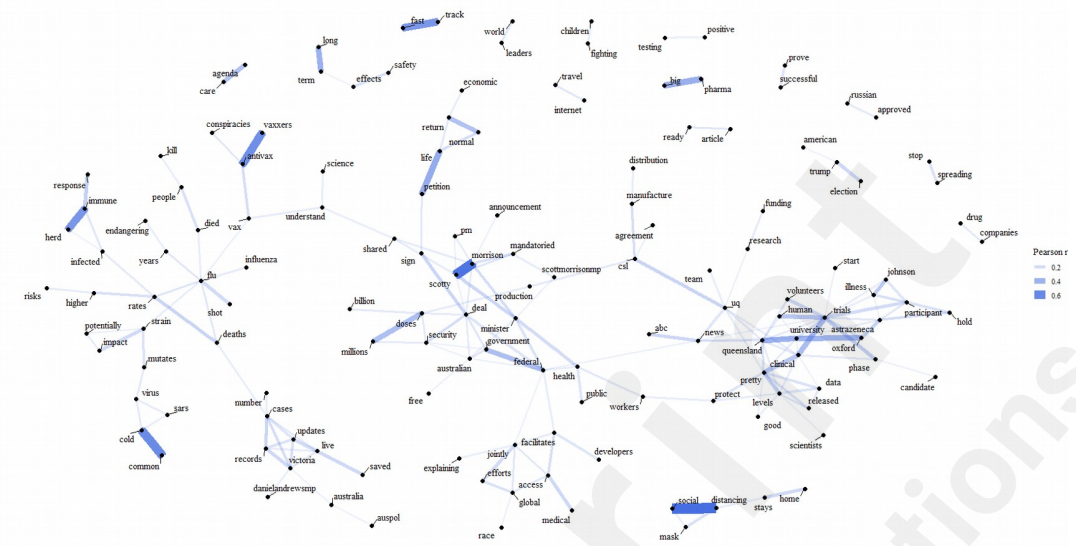
Figure 2. Network of word pairs with counts above 150 in the corpus.



We further examined the correlations between word tokens. The network of correlations (r > 0.1) between word tokens with count above 250 in the corpus was visualized in , where the edges with larger width and higher opacity mean stronger correlations between word tokens. A major network of words consisted of key words associated with the development and clinical trial of vaccines such as "trials", "clinical", "human", "phase", "volunteers", "participant", "astrazeneca", "university", "queensland", and "oxford". Another major word network which is worth noticing was composed of key words which were related to the Australian government's deal with vaccine manufacturers about providing doses for the public, such as "deal", "federal", "government", "scotty", "morrison", "millions", and "doses". On the other hand, "flu" was the center of another cluster associated with "influenza", "deaths", "rates", "vax", and "shot". Some word pairs such as "common" and "cold", "herd" and "immune", and "antivax" and "vaxxers", had distal associations with the main network. The pair "antivax" and "vaxxers" had some associations with "conspiracies" and "vax" linking with "flu" and "understand" which in turn correlated with "science" and "shared". Furthermore, "social" and "distancing" had strong correlation, but this bigram, along with a few words which had some associations with them, did not link with the large network of word tokens. Other similarly independent bigrams included "fast track" and "big pharma" for example.

Jan 2021

Figure 3. Network of correlations (r > 0.1) between word tokens with count above 250 in the corpus.



We built a three-topic LDA model and visualized the top 100 probability (beta) distributions of words for each topic in word clouds (Figure 4). The beta values were reported in Error: Reference source not found, and the top 20 probability (theta) distributions of topics in tweet samples were shown in Error: Reference source not found-8. Three topic themes were synthesized from the word clouds and tweets extracted.

Topic 1: Attitudes Towards COVID-19 and Its Vaccination

The latent topic 1 was centered on the public's attitudes or actions towards COVID-19 vaccination which were associated with personal values, theories, information received or personal experiences. Vaccine supporters would accept COVID-19 vaccination because they considered that measures should be taken to cope with the rising number of infections, deaths, healthcare burden and costs arisen from COVID-19. They would scorn those who pretended to be experts or posted misinformation such as claiming that deaths from COVID-19 were attributable to other diseases. Besides, they also supported public vaccinations by taking actions such as seeking funding sources and media to promote vaccine trials. Those who worried about the COVID-19 vaccination were skeptical about conspiracy theories such as "mark of the beast" and microchips in vaccines. The sudden pause of vaccine trials also triggered worries among them about the safety of vaccinations. Some of the Twitter users claimed that they would not get vaccinated further because of previous experience in vaccination's adverse effects. However, the stock markets had gone up when positive

news about vaccine developments were released. Some other Twitter users were those who disregarded COVID-19. They thought that COVID-19 had a much lower death rate than flu, so it is insignificant for vaccinations which would only benefit pharmaceutical firms or be politicized. Moreover, lockdown before mass vaccination was not considered efficient in the long run. They also thought that COVID-19 should not deserve much more attention than other global problems such as climate change, aged care, or other diseases.

Topic 2: Advocating Infection Control Measures Against COVID-19

The latent topic 2 was that some Twitter users were positive towards the development of COVID-19 vaccines and antivirals and recognized the needs of these products. Meanwhile, they would advocate following infection control measures and disprove misinformation or conspiracy theories. Some of the Twitter users rebutted tweets which were probably posted by anti-vaxxers or conspiracy theorists. For example, they refuted skepticism over the vaccines' safety, which were produced in fast-track; false claims about the association between the flu vaccine and COVID-19 infections and deaths; and inaccurate belief about vaccination coverage for achieving herd immunity depending on diseases. Some of their tweets emphasized the rising number of deaths of COVID-19 within a rather short period compared with other pandemics in history. They argued that though there were flu deaths, there were drugs, vaccines and promotion campaigns targeting flu. In comparison, the deaths from COVID-19 were soaring, and even worse than flu, without mass vaccinations or antivirals. However, COVID-19 deaths could have been preventable. With the previous experiences in developing the vaccine for other coronaviruses such as the MERS virus, they believed that the COVID-19 vaccine could be successfully developed to protect vulnerable groups such as patients. They thought that everyone was susceptible to COVID-19 after contracting the coronavirus without vaccination. In the future, antivirals could also be developed. Beyond vaccines and drugs, they thought physical measures such as wearing masks and social distancing should be followed, particularly when mass vaccination and antivirals were not yet available.

Topic 3: Misconceptions and Complaints About COVID-19 Control

The latent topic 3 generally shows the baseless claims and conspiracy theories that anti-vaxxers held against the COVID-19 vaccine and complaints and helplessness about the testing and lockdown, which would likely be ended with vaccination-induced herd immunity. Some Twitter users made claims which were unfounded or based on conspiracy theories against the COVID-19 vaccine. For example, one concluded that Australia suggested using stuff that has never been tested or certified to

fight the virus. Some others believed that hydroxychloroquine was an effective treatment so that banning it, but vaccine was politicized. They also thought that those rejecting hydroxychloroquine should take vaccines from Bill Gates who was falsely accused of planning to implant microchips into human bodies via vaccinations. However, other Twitter users pointed out vaccinations' limitations, such as not preventing viral transmissions and not treating COVID-19 and its complications. Even if vaccines are available, a high number of doses globally and tests for the virus or even antibodies are required if COVID-19 is not eradicated. Some complained that the tests led to the increase of known positive cases and in turn, the prolonged lockdown in which it was even more helpless when vaccines are not available. On the other hand, pro-vaxxers would celebrate success in vaccine development. They criticized anti-vaxxers, who should believe in science and accept vaccinations, of disregarding the disastrous consequences of COVID-19 and suggesting natural herd immunity which would be catastrophic. For example, allowing the rampant spread of coronavirus would lead to healthcare system breakdown and sacrifice others' lives.

Figure 4. Word distributions over three topics in the LDA model.

Figure 5 shows the change in sentiment scores of all tweets between Jan 2020 and Oct 2020. In each tweet, there could be both positive and negative sentiment with valences in opposite directions. The figure shows that the scores increased gradually between Jan and Mar 2020. The higher the sentiment score, regardless of directions, the more likely to have stronger sentiments in the tweet. However, mostly the tweets were inclining to having positive sentiment (67%) rather than the negative one (30%), while 3% of the tweets were neutral.

Jan 2021

Figure 5. Distributions of sentiment valences between Jan 2020 and Oct 2020.

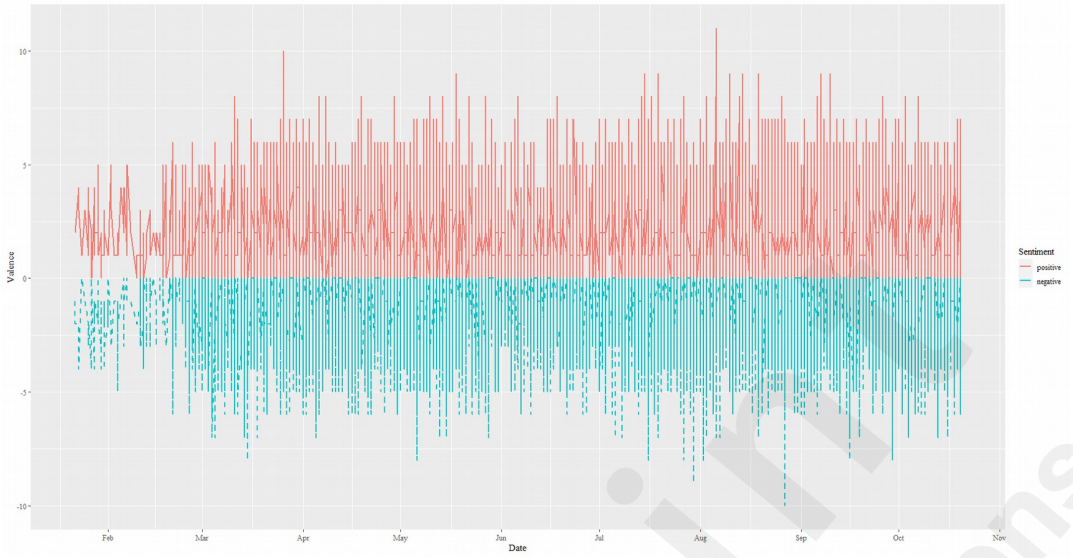


Figure 6 shows the emotion scores with respect to anticipation, joy, surprise and trust in all tweets. The scores also rose in the first quarter of 2020. Moreover, around 45% of the tweets were associated with these four emotions. Overall, the compositions of the emotions were trust (17%) and anticipation (14%). Some of the tweets were scored for the emotions of surprise (6%) and joy (8%).

Figure 6. Distributions of emotion valences regarding anticipation (A), trust (B), joy (C), and surprise (D) between Jan 2020 and Oct 2020.

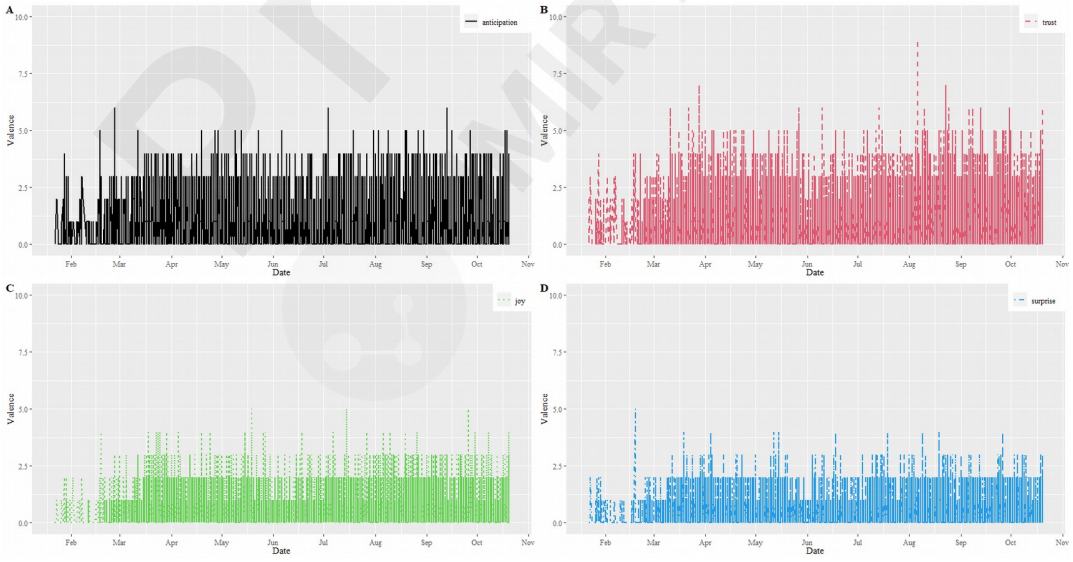
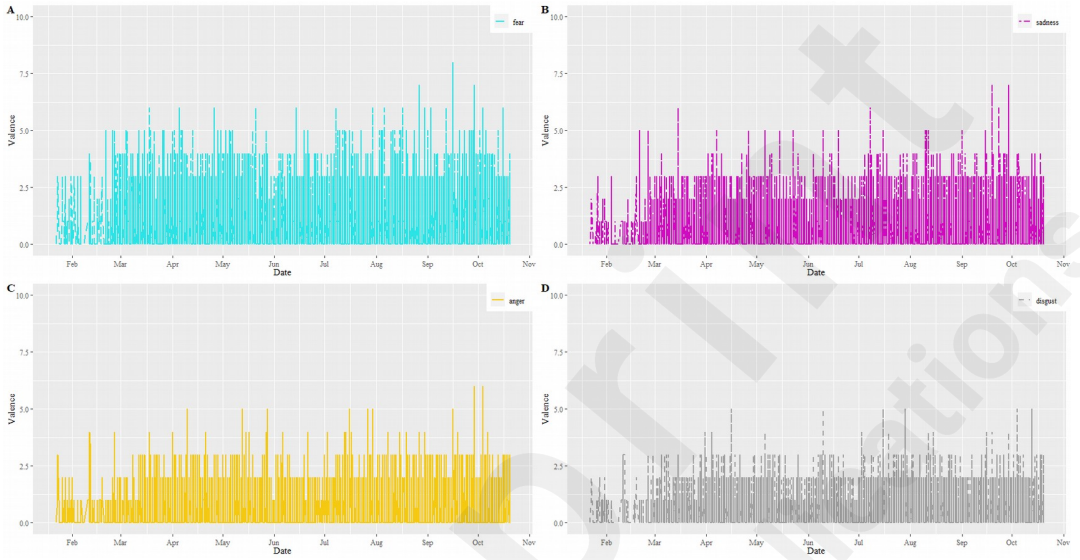


Figure 7 shows the scores of negative emotions such as anger, disgust, fear and sadness for all tweets. The scores grew in the first three months of 2020, and around one-third of the tweets were associated with these negative emotions. Among them,

fear was the most significant one (14%). There were other emotions such as sadness (8%), anger (7%) and disgust (5%) besides fear. On the other hand, nearly 22% of the tweets were emotionally neutral.

Figure 7. Distributions of emotion valences regarding fear (A), sadness (B), anger (C), and disgust (D) between Jan 2020 and Oct 2020.



Discussion

Principal Results

Three latent topics underlie the public opinion about COVID-19 vaccines among Australian Twitter users from January 22 to October 20, 2020. The latent topic 1 was about different attitudes and actions towards COVID-19 and its vaccinations. Pro-vaxxers recognized the consequences of the COVID-19 pandemic and would support vaccine trials. Those who were skeptical about vaccines were affected by misinformation and adverse effects, which are rare in statistics. Some Twitter users gave a low priority to COVID-19 and hence its vaccinations among other unrelated problems. The latent topic 2 showed that some Twitter users advocated infection control measures, had confidence in COVID-19 vaccine trials and rebutted tweets that were derived from conspiracy theories or misinformation. They argued that infections and deaths from COVID-19 had overtaken previous pandemics, and other measures such as wearing masks and social distancing should be followed when mass vaccination is yet to come. The latent topic 3 was centered on the baseless claims, conspiracy theories, complaints, and misconceptions about various measures against COVID-19, including vaccines, drugs, virus testing, lockdown, and herd immunity. The major pitfalls in these tweets are that the misinformation could not be supported with any valid, i.e., scientific evidence, and those complaints were not directly

associated with any solutions. Other significant findings were that nearly two-thirds of the sentiments in the tweets related to COVID-19 vaccines were found positive. Of those tweets analyzed, 17% of the emotions were linked with trust and 14% were associated with anticipation. However, 14% contained emotion of fear, and 8% were expressing sadness. Overall, less than one-third of the tweets' sentiments were classified as negative, and one-third of the tweets were associated with the four negative emotions (i.e., fear, sadness, anger, and disgust).

Evidence Before This Study

In the past decade, machine learning has been applied to explore the topics of contents and sentiments about vaccinations from Twitter users. Some studies examined tweets related to vaccinations in general, while some studies analyzed vaccination-related tweets focusing on a particular virus or disease, such as the influenza virus which causes respiratory illness, or HPV, which is mainly sexually transmitted. In those studies, they identified both positive and negative sentiments towards vaccinations and the one being neutral. Nevertheless, the outcomes of sentiment categories and the topics identified from Twitter users varied across studies focusing on different countries, years, viruses and diseases.

For example, Jamison, Broniatowski [78] generated 100 topics using LDA in which nearly half were annotated as pro-vaccination, and less than 30% were coded as anti-vaccination from English vaccine-relevant tweets between 2014 and 2017. However, Raghupathi, Ren [55] found that both positive and negative sentiments were around 40% among English tweets in the first half of 2019. On the other hand, the composition of sentiments in non-English tweets could be different from that in English tweets. In Italy, Tavoschi, Quattrone [79] used support vector machine (SVM) to classify term frequency–inverse document frequency (TF-IDF) of tweets between 2016 and 2017, and found that 60% were neutral, 23% were against vaccination and only 17% were pro-vaccination. The researchers also found that the number of pro-vaccine tweets had become higher than the number of anti-vaccine tweets when news about the law and government's requirement for vaccination, the soaring of positive cases or deaths were broadcast [80].

Furthermore, the topics identified were not entirely similar across studies. For instance, Jamison, Broniatowski [78] summarized five pro-vaccine themes and five anti-vaccine themes from 100 topics; and Raghupathi, Ren [55] identified three focuses such as the search for better vaccines, the disease outbreak, and the debates between pro-vaxxers and anti-vaxxers regarding measles in particular. Chan,

Jamieson [67], who studied influenza vaccination in the United States, used LDA to create ten topics in which some of them share similar attributes with the themes from Jamison, Broniatowski [78], such as vaccine science, safety concerns and conspiracy theories. Some, but not all, of the similar themes, focuses and topics could also be seen in the analyses of tweets about vaccination regardless of virus types such as those in the studies surrounding HPV vaccinations [66, 81-84].

Added Value of This Study

This study is the first topic modeling and sentiment analysis on Australian tweets about COVID-19 vaccinations. As COVID-19 has turned into a pandemic, it is necessary to explore and summarize public opinion and sentiments in the discussion about the COVID-19 vaccine so as to prepare for the promotion of vaccination, which needs to be strengthened. This study used a traditional NLP technique, the LDA, to identify three latent topics in the tweets associated with COVID-19 vaccinations, they were "Attitudes towards COVID-19 and its vaccination", "Advocating infection control measures against COVID-19", and "Misconceptions and complaints about COVID-19 control". Furthermore, this study discovered that positive sentiment in the discussion about the COVID-19 vaccine was higher than the negative one; and relatively larger proportions of emotions were trust and anticipation but also fear. This study also visualizes the word clouds, counts of word pairs and the correlations between words, which offer supplementary angles in interpreting the results. For example, high-frequency words and word pairs that commonly appeared together were intuitively presented.

Australian population has been the focus of research on tweets related to vaccine. Taking HPV vaccine as an example, nearly one-fifth of Australian Twitter users expressed health concerns about the vaccination [82]; and around one-third of the exposure to information on Twitter were associated with misinformation or adverse effects of the vaccine [83]. Our study provides new insights into the Australian topics of discussion and sentiments towards vaccine against COVID-19, which is now a global pandemic and causes over 900 deaths in Australia [3] and over 1.8 million deaths worldwide [4] as of early Jan 2021. By assessing public opinion and sentiments associated with COVID-19 vaccination, governments and health agencies could plan, implement and adjust a timely promotion of vaccination to achieve the crucial herd immunity as soon as possible.

Implications of All the Available Evidence

In the previous studies results, we do not see a prevalent objection or opposition, in

terms of topics identified or sentiments, towards vaccination regardless of virus types. A number of topics' focuses or themes shared a certain level of similarity across studies concerning different viruses. For instance, topics of safety, scientific evidence and conspiracy theories were commonly found across studies. Topics like scandals of vaccine, misinformation and disease outbreaks were identified in some other studies. These results indicated that the public was concerned about the benefits and risks of vaccination at the individual and social levels, and the type of virus or disease when deciding whether to receive the vaccination.

In our study, besides fabricated information such as microchips in vaccines and flu vaccine causing COVID-19 deaths, some Twitter users thought that COVID-19 was not disastrous enough compared with other existing global crises, and that the pandemic is being politicized or commercialized. These conspiracy theories, along with other anti-vaccine propagandas such as touting natural herd immunity, indicated that the risks of deaths, complications or sequela arisen from COVID-19 to the others, or maybe oneself, were acceptable to some people in public.

Though the Australian opinion showed more positive sentiment related to COVID-19 vaccinations, the positive sentiment was not a leading vast majority compared to the negative one. That means more works need to be done to promote vaccination so as to achieve herd immunity to protect vulnerable and minority groups. We suggest that rigorous science that is easily understandable might swamp the biased, fabricated or outdated information in public. Governments should build and strengthen the public's confidence in COVID-19 vaccination, if it is not mandatory, i.e., required by law, beyond arranging vaccine delivery logistically and vaccine administration clinically.

Limitations

Our results represent the Twitter users in the Australian public, which is a different approach from national survey statistics. Furthermore, not only year and/or country of concern, but also analysis methods might lead to variation in topics and sentiments towards vaccinations. For supervised learning such as SVM, a training set is required but it needs to be manually labeled which might carry some subjectivity in categorizing tweets into pre-defined topics for training. However, the advantage is that the set could be used to validate the model performance and then test a large dataset. Considering unsupervised learning such as LDA, Dirichlet Multinomial Mixtures (DMM), and k-means of TF-IDF, the primary limitation is the subjectivity in defining the topics created [55, 67]. In addition, a sound reason or calculation is needed to support the pre-set number of topics which would affect the results.

Some previous studies generated a rather high number of topics (30-100) in LDA or DMM model, and then manually grouped the topics into themes [66, 78, 83]. However, there was risk of bias since the content of each topic was not reported in detail, and the contents of the themes could be mixed which is difficult to interpret. Furthermore, the manual grouping also contained the risk of subjectivity. In the current study, we adopted the LDA which was similar to the one used by Chan, Jamieson [67]. We identified three latent topics in which importance of words were visualized; the frequency of word pairs and correlations between words provide additional results corresponding to the topic contents.

Regarding sentiment analysis, the number of emotion categories were limited at eight [76, 85], but emotion is an abstract and broad concept which may involve as much as 27 categories [86]. Furthermore, words with spelling mistakes could not be identified and analyzed in the algorithm. With respect to each term for the development of emotion lexicon by Mohammad and Turney [76], only five individuals in the public were recruited to annotate a term against each of the eight emotions. The emotions of a term were annotated without considering possible contexts. Moreover, the inter-rater reliability statistics were not reported though the agreement percentages were apparently high.

Future Directions

Our study adopted an unsupervised machine learning method, the LDA, for topic analysis. Future studies could investigate supervised learning to train classifiers to categorize tweets into different topics and sentiments based on a recognized theoretical framework. Such a framework could be proposed after extensive literature review and qualitative synthesis; and manual annotations should be as transparent, objective and reliable as possible. Results from supervised learning following the same theoretical framework could be compared across the analyses of different datasets, for example, the results from different countries as shown by Shapiro, Surian [82]. Spatiotemporal analysis of tweets about COVID-19 vaccination could also be attempted in the future. Similar studies have been conducted on Twitter data to study emergency department visits for influenza-like illness in New York City [87], COVID-19 related stress symptoms in the United States [88], and communicating the risk of MERS infections in South Korea [89]. Furthermore, individual reactions to the COVID-19 vaccine in the tweets could be monitored over time and tested for correlations between frequencies of identified topics or emotions, important real events, and health indicators such as vaccination coverage, infection rate and death

rate. Besides misinformation and conspiracy theories spreading on social media, future research should explore personal values which might hinder collective healthcare strategies and positive outcomes.

Conclusions

Our new findings indicate that the Australian public possessed different attitudes towards COVID-19 and its vaccination. Moreover, some of them had misconceptions and complaints about COVID-19 and infection control measures, while others advocated pharmaceutical and non-pharmacological measures against COVID-19. Nonetheless, in our sentiment analysis, the level of positive sentiment in the public may not be strong enough to further a high vaccination coverage to achieve vaccination-induced herd immunity, which is collective, to protect oneself and the others. Hence, for those who are not subject to contraindications, receiving the vaccination is not merely a personal choice, but also a social responsibility. Governments should explore public opinion and sentiments towards COVID-19 vaccination, and get the public psychologically prepared for vaccination with evidence-based, authorized and understandable information, besides supporting the biomedical development, storage, delivery and clinical administration of vaccines.

Acknowledgments

The work was supported in part by the Murdoch New Staff Startup Grant (SEIT NSSG).

Conflicts of Interest

None declared.

Abbreviations

API: application programming interface
COVID-19: coronavirus disease 2019
DMM: Dirichlet multinomial mixtures
HPV: human papillomavirus
LDA: latent Dirichlet allocation
MERS: Middle East respiratory syndrome
MMR: measles, mumps, and rubella
NASA: National Aeronautics and Space Administration
NLP: natural language processing
PCV13: 13-valent pneumococcal conjugate vaccines
SAO: Smithsonian Astrophysical Observatory

Jan 2021

SARS-CoV-2: severe acute respiratory syndrome coronavirus 2
SVM: support vector machine
TF-IDF: term frequency–inverse document frequency

Jan 2021

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020 Feb 15;395(10223):497-506. PMID: 31986264. doi: 10.1016/s0140-6736(20)30183-5.

2. Hong Kong Centre for Health Protection. Countries/areas with reported cases of Coronavirus Disease 2019 (COVID 19) (Last updated on January 4, 2021, 11 am). 2020.

3. Australian Department of Health. Coronavirus (COVID-19) current situation and case numbers. 2020; Available from: https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers.

4. World Health Organization. WHO coronavirus disease (COVID-19) dashboard. 2020 [cited 2020 22nd Sep]; Available from: https://covid19.who.int/.

5. Debata B, Patnaik P, Mishra A. COVID -19 pandemic! It's impact on people, economy, and environment. Journal of Public Affairs. 2020. doi: 10.1002/pa.2372.

6. Song L, Zhou Y. The COVID-19 pandemic and its impact on the global economy: What does it take to turn crisis into opportunity? China & World Economy. 2020;28(4):1-25. doi: 10.1111/cwe.12349.

7. Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. Medical Letter on the CDC & FDA. 2020:217.

8. He J, Guo Y, Mao R, Zhang J. Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. Journal of Medical Virology. 2020 Jul 21. PMID: 32691881. doi: 10.1002/jmv.26326.

9. Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. Journal of Infection. 2020 Jun;80(6):656-65. PMID: 32283155. doi: 10.1016/j.jinf.2020.03.041.

10. Grant MC, Geoghegan L, Arbyn M, Mohammed Z, McGuinness L, Clarke EL, et al. The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. PLoS One. 2020;15(6):e0234765-e. PMID: 32574165. doi: 10.1371/journal.pone.0234765.

11. Baksh M, Ravat V, Zaidi A, Patel RS. A systematic review of cases of acute respiratory distress syndrome in the coronavirus disease 2019 pandemic. Cureus. 2020;12(5):e8188-e. PMID: 32566429. doi: 10.7759/cureus.8188.

12. Hasan SS, Capstick T, Ahmed R, Kow CS, Mazhar F, Merchant HA, et al.

Mortality in COVID-19 patients with acute respiratory distress syndrome and corticosteroids use: A systematic review and meta-analysis. Expert Review of Respiratory Medicine. 2020 Jul 31. PMID: 32734777. doi: 10.1080/17476348.2020.1804365.

13. Zhang Z, Zhang L, Wang Y. COVID-19 indirect contact transmission through the oral mucosa must not be ignored. Journal of Oral Pathology & Medicine. 2020 May;49(5):450-1. PMID: 32281674. doi: 10.1111/jop.13019.

14. Alimohamadi Y, Taghdir M, Sepandi M. Estimate of the basic reproduction number for COVID-19: A systematic review and meta-analysis. Journal of Preventive Medicine and Public Health. 2020 May;53(3):151-7. PMID: 32498136. doi: 10.3961/jpmph.20.076.

15. Billah A, Miah M, Khan N. Reproductive number of COVID-19: A systematic review and meta-analysis based on global level evidence. Respiratory Therapeutics Week. 2020:386.

16. Lei H, Xu X, Xiao S, Wu X, Shu Y. Household transmission of COVID-19-a systematic review and meta-analysis. Journal of Infection. 2020:S0163-4453(20)30571-5. PMID: 32858069. doi: 10.1016/j.jinf.2020.08.033.

17. Rehman M, Tauseef I, Aalia B, Shah SH, Junaid M, Haleem KS. Therapeutic and vaccine strategies against SARS-CoV-2: Past, present and future. Future Virol. 2020:10.2217/fvl-020-0137. PMID: PMC7386380. doi: 10.2217/fvl-2020-0137.

18. Badgujar KC, Badgujar VC, Badgujar SB. Vaccine development against coronavirus (2003 to present): An overview, recent advances, current scenario, opportunities and challenges. Diabetology & Metabolic Syndrome. 2020 Sep-Oct;14(5):1361-76. PMID: 32755836. doi: 10.1016/j.dsx.2020.07.022.

19. Spencer JP, Trondsen Pawlowski RH, Thomas S. Vaccine adverse events: Separating myth from reality. American Academy of Family Physicians. 2017;95(12):786-94.

20. Maglione MA, Das L, Raaen L, Smith A, Chari R, Newberry S, et al. Safety of vaccines used for routine immunization of U.S. children: A systematic review. Pediatrics. 2014 Aug;134(2):325-37. PMID: 25086160. doi: 10.1542/peds.2014-1079.

21. Fine P, Eames K, Heymann DL. "Herd immunity": A rough guide. Clinical Infectious Diseases. 2011 Apr 1;52(7):911-6. PMID: 21427399. doi: 10.1093/cid/cir007.

22. Lopez A, Mariette X, Bachelez H, Belot A, Bonnotte B, Hachulla E, et al. Vaccination recommendations for the adult immunosuppressed patient: A systematic review and comprehensive field synopsis. Journal of Autoimmunity. 2017 Jun;80:10-27. PMID: 28381345. doi: 10.1016/j.jaut.2017.03.011.

23. Bansal P, Goyal A. COVID-19 – Challenges ahead of vaccination in

immunocompromised patients. General Internal Medicine and Clinical Innovations. 2020;5:1-3. doi: 10.15761/GIMCI.1000196.

24. Hussain A, Ali S, Ahmed M, Hussain S. The anti-vaccination movement: A regression in modern medicine. Cureus. 2018;10(7):e2919-e. PMID: 30186724. doi: 10.7759/cureus.2919.

25. Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 1998 Feb 28;351(9103):637-41. PMID: 9500320. doi: 10.1016/s0140-6736(97)11096-0.

26. Taylor B, Miller E, Farrington CP, Petropoulos MC, Favot-Mayaud I, Li J, et al. Autism and measles, mumps, and rubella vaccine: No epidemiological evidence for a causal association. Lancet. 1999 Jun 12;353(9169):2026-9. PMID: 10376617. doi: 10.1016/s0140-6736(99)01239-8.

27. Fombonne E, Chakrabarti S. No evidence for a new variant of measles-mumps-rubella-induced autism. Pediatrics. 2001;108(4):e58-e. doi: 10.1542/peds.108.4.e58.

28. Farrington CP, Miller E, Taylor B. MMR and autism: Further evidence against a causal association. Vaccine. 2001 Jun 14;19(27):3632-5. PMID: 11395196. doi: 10.1016/s0264-410x(01)00097-4.

29. DeStefano F, Thompson WW. MMR vaccine and autism: An update of the scientific evidence. Expert Review of Vaccines. 2004 Feb;3(1):19-22. PMID: 14761240. doi: 10.1586/14760584.3.1.19.

30. Peltola H, Patja A, Leinikki P, Valle M, Davidkin I, Paunio M. No evidence for measles, mumps, and rubella vaccine-associated inflammatory bowel disease or autism in a 14-year prospective study. Lancet. 1998;351(9112):1327-8. doi: 10.1016/S0140-6736(98)24018-9.

31. DeStefano F, Chen RT. Negative association between MMR and autism. Lancet. 1999;353(9169):1987-8. doi: 10.1016/S0140-6736(99)00160-9.

32. Editors of The Lancet. Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 2010;375(9713):445-. doi: 10.1016/S0140-6736(10)60175-4.

33. Singh K, Mehta S. The clinical development process for a novel preventive vaccine: An overview. J Postgrad Med. 2016 Jan-Mar;62(1):4-11. PMID: 26732191. doi: 10.4103/0022-3859.173187.

34. Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, et al. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: A preliminary report of a phase 1/2, single-blind, randomised controlled trial. Lancet. 2020 Aug 15;396(10249):467-78. PMID: 32702298. doi: 10.1016/s0140-

6736(20)31604-4.
35. Zhu F-C, Guan X-H, Li Y-H, Huang J-Y, Jiang T, Hou L-H, et al. Immunogenicity and safety of a recombinant adenovirus type-5-vectored COVID-19 vaccine in healthy adults aged 18 years or older: A randomised, double-blind, placebo-controlled, phase 2 trial. Lancet. 2020;396(10249):479-88. doi: 10.1016/S0140-6736(20)31605-6.
36. World Health Organization. The push for a COVID-19 vaccine. 2020 [cited 2020 22nd Sep]; Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines.
37. U.S. Centers for Disease Control and Prevention. Immunization schedules. 2020; Available from: https://www.cdc.gov/vaccines/schedules/hcp/schedule-changes.html.
38. U.S. Food and Drug Administration. FDA insight: Vaccines for COVID-19, Part 1. 2020; Available from: https://www.fda.gov/news-events/fda-insight/fda-insight-vaccines-covid-19-part-1.
39. U.K. Department of Health and Social Care. Distributing vaccines and treatments for COVID-19 and flu. 2020; Available from: https://www.gov.uk/government/consultations/distributing-vaccines-and-treatments-for-covid-19-and-flu.
40. Kortum P, Edwards C, Richards-Kortum R. The impact of inaccurate internet health information in a secondary school learning environment. Journal of Medical Internet Research. 2008;10(2):e17. doi: 10.2196/jmir.986.
41. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm--an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine. 2012 May 28;30(25):3778-89. PMID: 22172504. doi: 10.1016/j.vaccine.2011.11.112.
42. World Health Organization. Smallpox vaccines. 2020; Available from: https://www.who.int/csr/disease/smallpox/vaccines/en/.
43. Khuroo MS, Khuroo M, Khuroo MS, Sofi AA, Khuroo NS. COVID-19 vaccines: A race against time in the middle of death and devastation! J Clin Exp Hepatol. 2020:10.1016/j.jceh.2020.06.003. PMID: 32837093. doi: 10.1016/j.jceh.2020.06.003.
44. Funk CD, Laferrière C, Ardakani A. A snapshot of the global race for vaccines targeting SARS-CoV-2 and the COVID-19 pandemic. Front Pharmacol. 2020;11:937-. PMID: 32636754. doi: 10.3389/fphar.2020.00937.
45. Cyranoski D, Mallapaty S. Scientists relieved as coronavirus vaccine trial restarts — but question lack of transparency. Nature. 2020 14 September 2020.
46. Mahase E. Covid-19: Johnson and Johnson vaccine trial is paused because of unexplained illness in participant. BMJ. 2020;371:m3967. doi: 10.1136/bmj.m3967.

47. Ortiz-Sánchez E, Velando-Soriano A, Pradas-Hernández L, Vargas-Román K, Gómez-Urquiza JL, Cañadas-de la Fuente GA, et al. Analysis of the anti-vaccine movement in social networks: A systematic review. International Journal of Environmental Research and Public Health. 2020;17(15):1-11. doi: 10.3390/ijerph17155394.
48. Megget K. Even covid-19 can't kill the anti-vaccination movement. BMJ. 2020 Jun 4;369:m2184. PMID: 32499217. doi: 10.1136/bmj.m2184.
49. Bruns A, Harrington S, Hurcombe E. 'Corona? 5G? or both?': The dynamics of COVID-19/5G conspiracy theories on Facebook. Media International Australia, Incorporating Culture & Policy. 2020:1329878. doi: 10.1177/1329878X20946113.
50. Ahmed W, Vidal-Alaball J, Downing J, López Seguí F. COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. Journal of Medical Internet Research. 2020 May 6;22(5):e19458. PMID: 32352383. doi: 10.2196/19458.
51. Twitter. Investor fact sheet. 2020.
52. Khanna P. Sentiment analysis: An approach to opinion mining from Twitter data using R. International Journal of Advanced Research In Computer Science. 2017;8(8):252-6. doi: 10.26483/ijarcs.v8i8.4716.
53. Zimbra D, Abbasi A, Zeng D, Chen H. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. ACM Transactions on Management Information Systems (TMIS). 2018;9(2):1-29. doi: 10.1145/3185045.
54. Du J, Xu J, Song H, Liu X, Tao C. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. Journal of Biomedical Semantics. 2017;8(1):9-. doi: 10.1186/s13326-017-0120-6.
55. Raghupathi V, Ren J, Raghupathi W. Studying public perception about vaccination: A sentiment analysis of tweets. International Journal of Environmental Research and Public Health. 2020;17(10):3464. doi: 10.3390/ijerph17103464.
56. Kearney M. rtweet: Collecting and analyzing Twitter data. Journal of Open Source Software. 2019;4(42):1829. doi: 10.21105/joss.01829.
57. Rinker TW. qdapRegex: Regular expression removal, extraction, and replacement tools. 0.7.2. Buffalo, New York: University at Buffalo; 2017; Available from: http://github.com/trinker/qdapRegex.
58. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. Journal of Statistical Software. 2008;25(5):1-54. doi: 10.18637/jss.v025.i05.
59. Feinerer I, Hornik K. tm: Text mining package. R package version 0.7-8. 2020; Available from: https://CRAN.R-project.org/package=tm.
60. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength natural language processing in Python. Zenodo; 2020; Available from: https://doi.org/10.5281/zenodo.1212303.

61. Řehůřek R, Sojka P, editors. Software framework for topic modelling with large corpora. LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010; Valletta, Malta: ELRA.
62. Sedgewick R, Wayne K. 3.5 Searching Applications. Algorithms. 4 ed. New Jersey: Addison-Wesley Professional; 2014.
63. SAO/NASA Astrophysics Data System. SAO/NASA ADS abstract service stopword list. SAO/NASA Astrophysics Data System; 2020 [cited 2020 7th November]; Available from: http://adsabs.harvard.edu/abs_doc/stopwords.html.
64. Robinson D, Misra K, Silge J. widyr: Widen, process, and re-tidy a dataset. 2020; Available from: https://cran.r-project.org/web/packages/widyr/index.html.
65. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003;3(4-5):993-1022.
66. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. Journal of Medical Internet Research. 2016;18(8):e232-e. doi: 10.2196/jmir.6045.
67. Chan M-pS, Jamieson KH, Albarracin D. Prospective associations of regional social media messages with attitudes and actual vaccination: A big data and survey study of the influenza vaccine in the United States. Vaccine. 2020;38(40):6236-47. doi: 10.1016/j.vaccine.2020.07.054.
68. Nikita M, Chaney N. ldatuning: Tuning of the latent Dirichlet allocation models parameters. 2020; Available from: https://cran.r-project.org/web/packages/ldatuning/index.html.
69. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN. On finding the natural number of topics with latent Dirichlet allocation: Some observations. Berlin, Heidelberg: Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 391-402.
70. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing. 2009;72(7):1775-81. doi: 10.1016/j.neucom.2008.06.011.
71. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(Supplement 1):5228-35. doi: 10.1073/pnas.0307752101.
72. Deveaud R, SanJuan E, Bellot P. Accurate and effective Latent Concept Modeling for ad hoc information retrieval. Document Numérique. 2014;17(1):61-84. doi: 10.3166/dn.17.1.61-84.
73. Grün B, Hornik K. Topicmodels: An R package for fitting topic models. Journal of Statistical Software. 2011;40(13):1-30. doi: 10.18637/jss.v040.i13.
74. Jockers ML. Syuzhet: Extract sentiment and plot arcs from text. 2015;

Jan 2021

Available from: https://github.com/mjockers/syuzhet.

75. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D, editors. The Stanford CoreNLP natural language processing toolkit. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2014.

76. Mohammad S, Turney P, editors. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; 2010; LA, California.

77. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Computational Intelligence. 2013;29(3):436-65. doi: 10.1111/j.1467-8640.2012.00460.x.

78. Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, et al. Adapting and extending a typology to identify vaccine misinformation on Twitter. American Journal of Public Health. 2020;110(Suppl 3):S331-S9. doi: 10.2105/AJPH.2020.305940.

79. Tavoschi L, Quattrone F, D'Andrea E, Ducange P, Vabanesi M, Marcelloni F, et al. Twitter as a sentinel tool to monitor public opinion on vaccination: An opinion mining analysis from September 2016 to August 2017 in Italy. Human Vaccines & Immunotherapeutics. 2020;16(5):1062-9. doi: 10.1080/21645515.2020.1714311.

80. D'Andrea E, Ducange P, Bechini A, Renda A, Marcelloni F. Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Systems with Applications. 2019;116:209-26. doi: 10.1016/j.eswa.2018.09.009.

81. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine. 2017;35(23):3033-40. doi: 10.1016/j.vaccine.2017.04.060.

82. Shapiro GK, Surian D, Dunn AG, Perry R, Kelaher M. Comparing human papillomavirus vaccine concerns on Twitter: A cross-sectional study of users in Australia, Canada and the UK. BMJ Open. 2017;7(10):e016869-e. doi: 10.1136/bmjopen-2017-016869.

83. Dyda A, Shah Z, Surian D, Martin P, Coiera E, Dey A, et al. HPV vaccine coverage in Australia and associations with HPV vaccine information exposure among Australian Twitter users. Human Vaccines & Immunotherapeutics. 2019;15(7-8):1488-95. doi: 10.1080/21645515.2019.1596712.

84. Luo X, Zimet G, Shah S. A natural language processing framework to analyse the opinions on HPV vaccination reflected in Twitter over 10 years (2008 - 2017). Human Vaccines & Immunotherapeutics. 2019;15(7-8):1496-504. doi:

10.1080/21645515.2019.1627821.

85. Plutchik R. A general psychoevolutionary theory of emotion. . Emotion: Theory, Research, and Experience. 1980;1(3):3–33.

86. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences of the United States of America. 2017 Sep 19;114(38):E7900-e9. PMID: 28874542. doi: 10.1073/pnas.1702247114.

87. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. Journal of Medical Internet Research. 2014;16(10):e236-74. doi: 10.2196/jmir.3416.

88. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. International Journal of Environmental Research and Public Health. 2020;17(14):4988. doi: 10.3390/ijerph17144988.

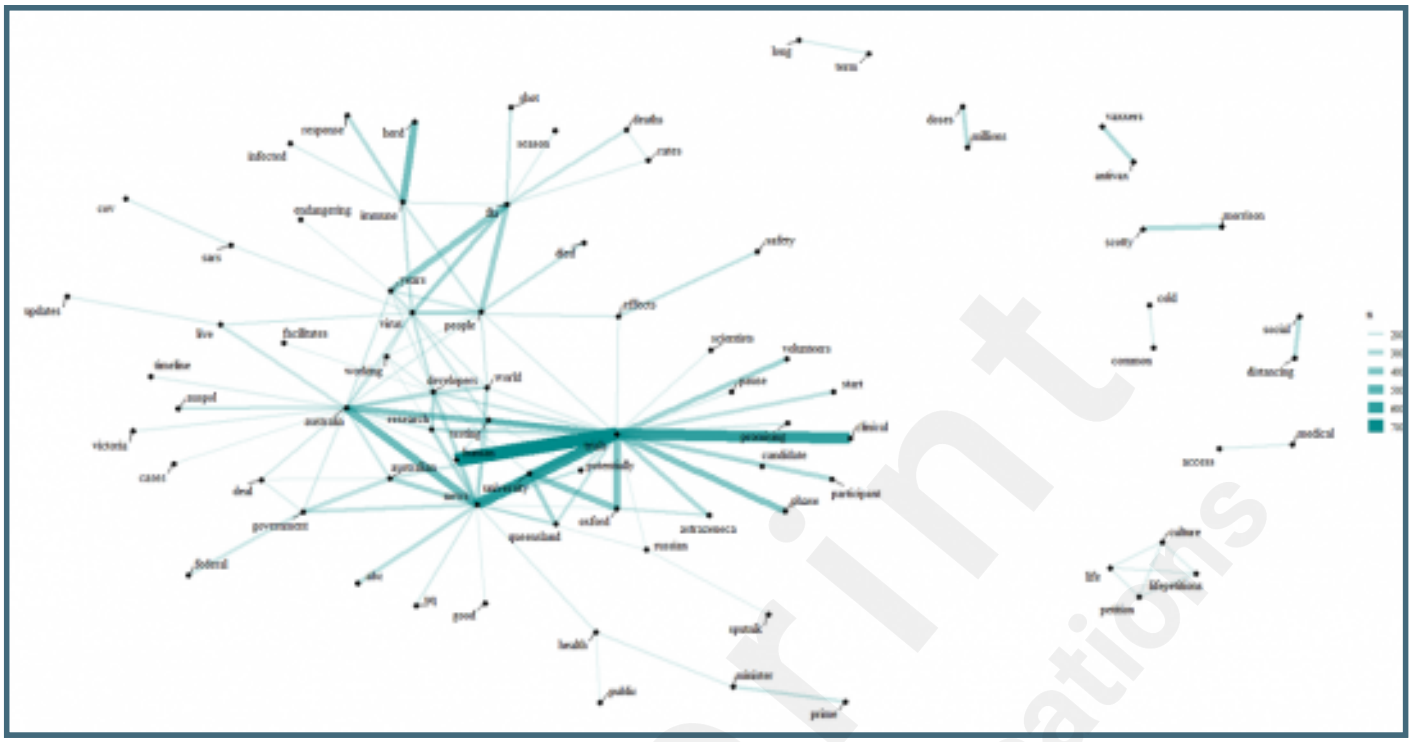
89. Kim I-H, Feng C-C, Wang Y-C, Spitzberg BH, Tsou M-H. Exploratory spatiotemporal analysis in risk communication during the MERS outbreak in South Korea. Professional Geographer. 2017;69(4):629-43. doi: 10.1080/00330124.2017.1288577.

Supplementary Files

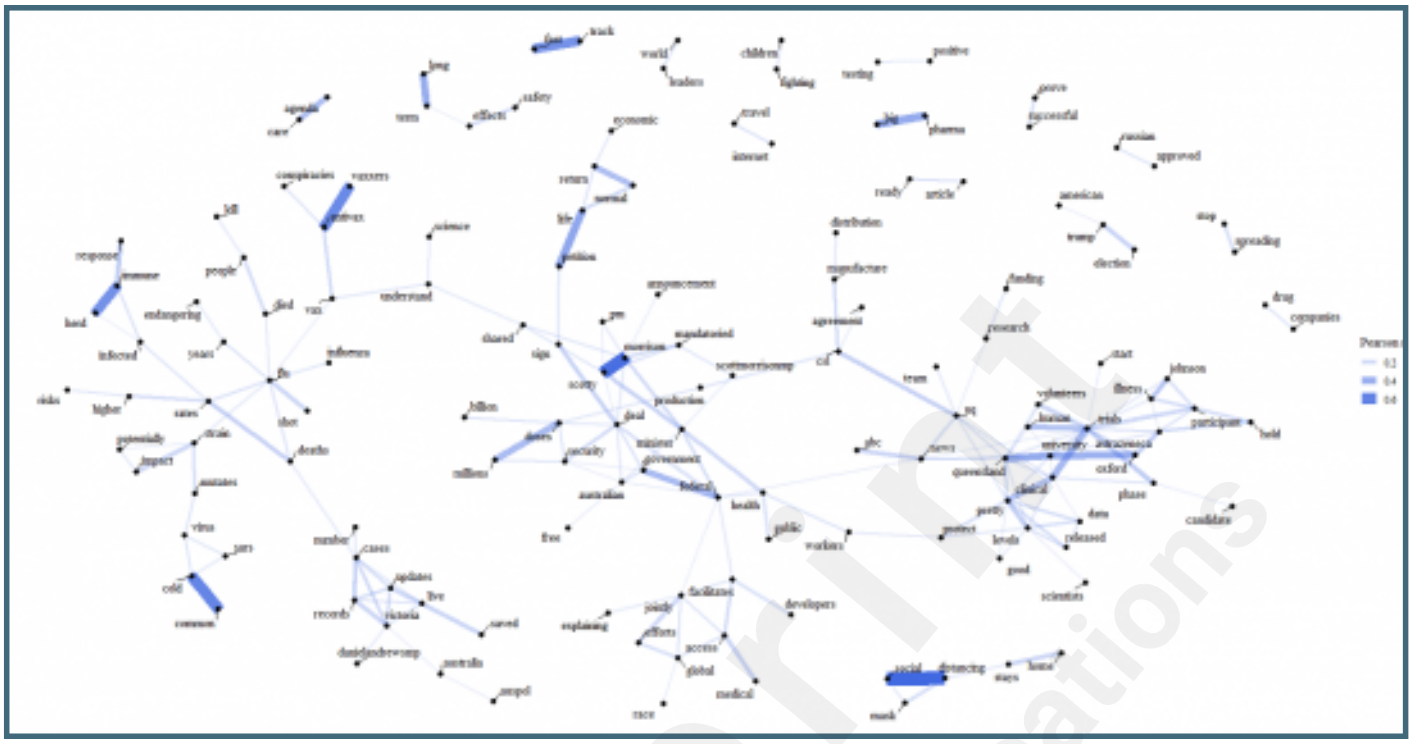
Figures

Word cloud of word tokens with count over 250 in the corpus.

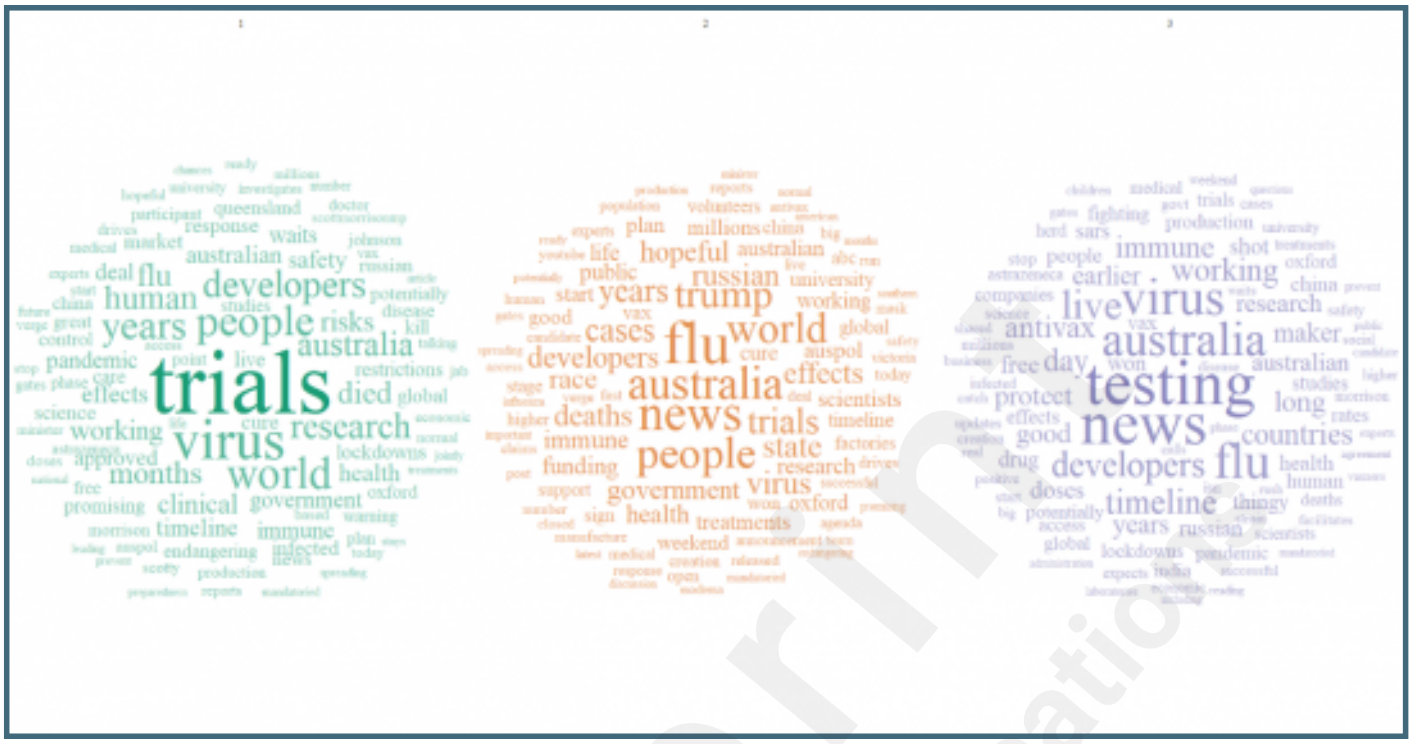
Network of word pairs with counts above 150 in the corpus.



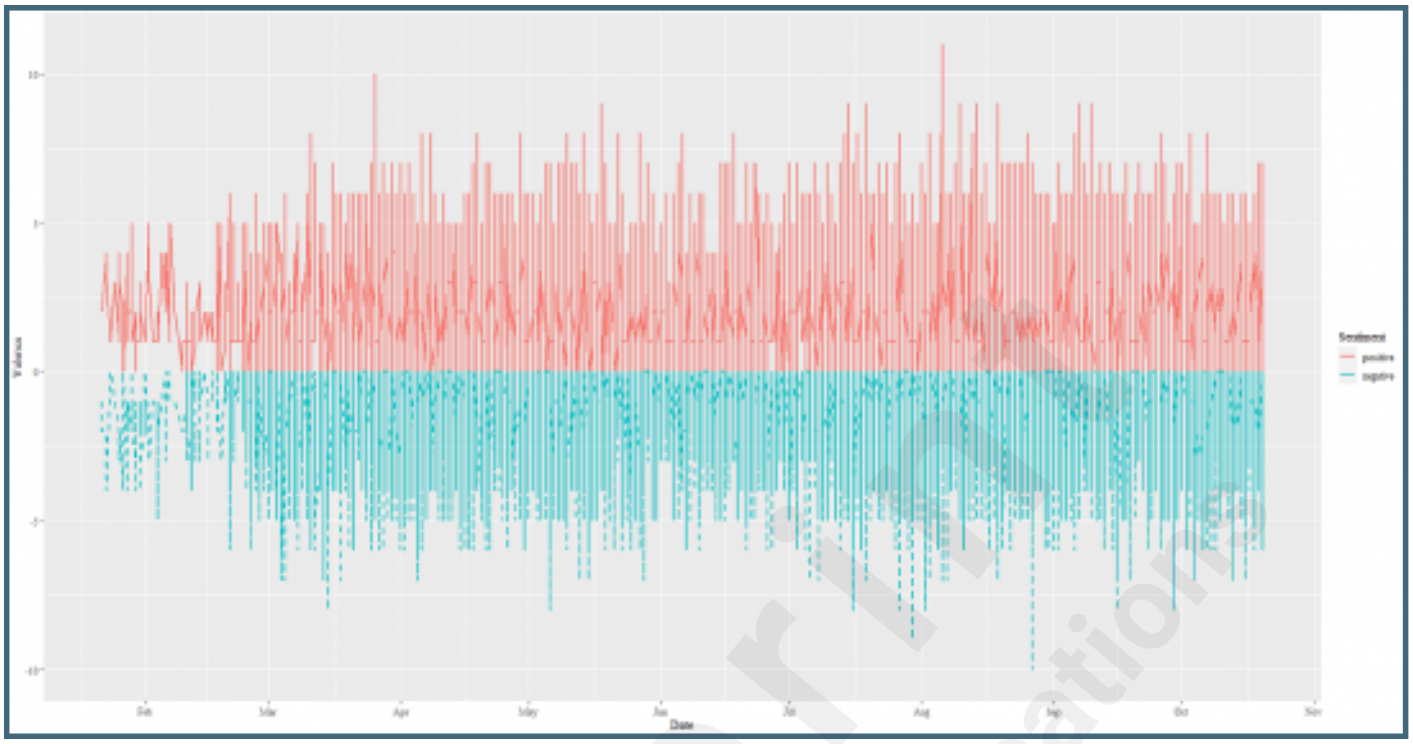
Network of correlations (r > 0.1) between word tokens with count above 250 in the corpus.



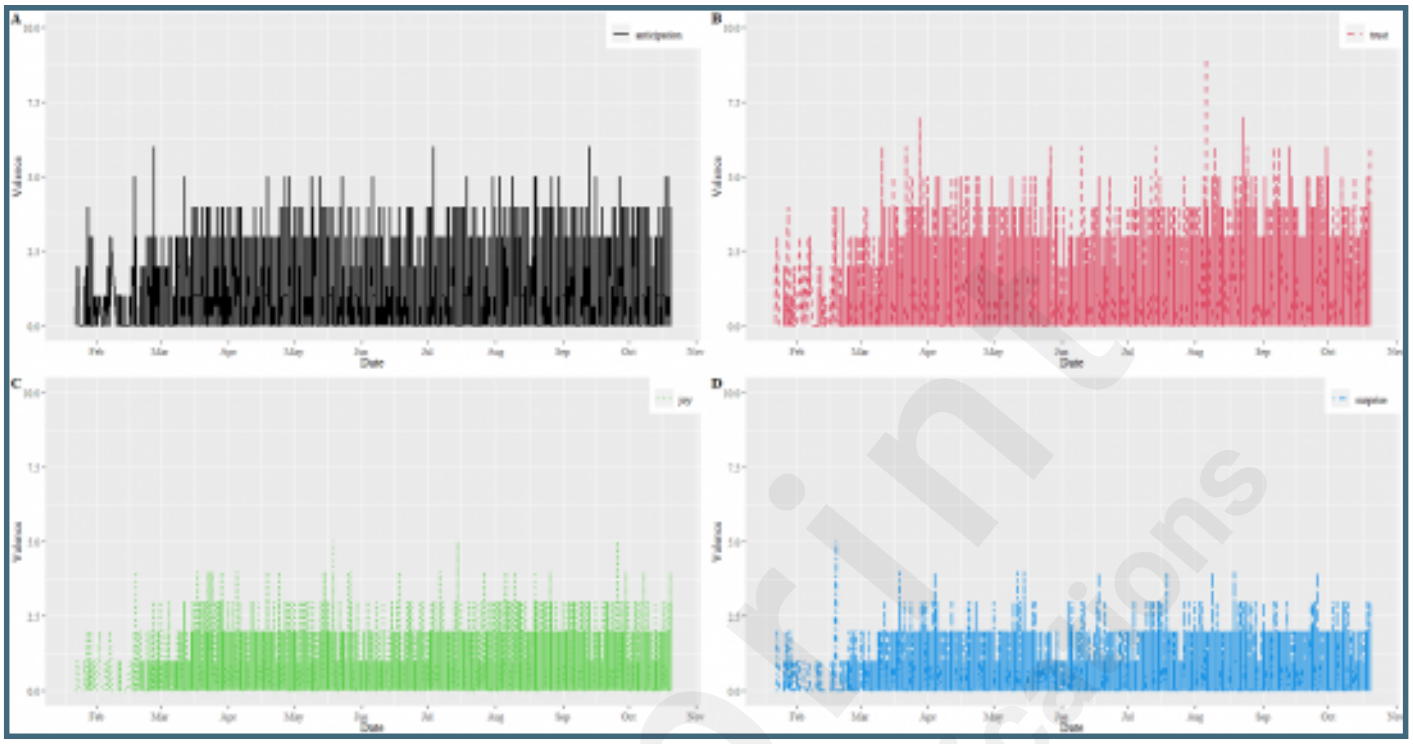
Word distributions over three topics in the LDA model.



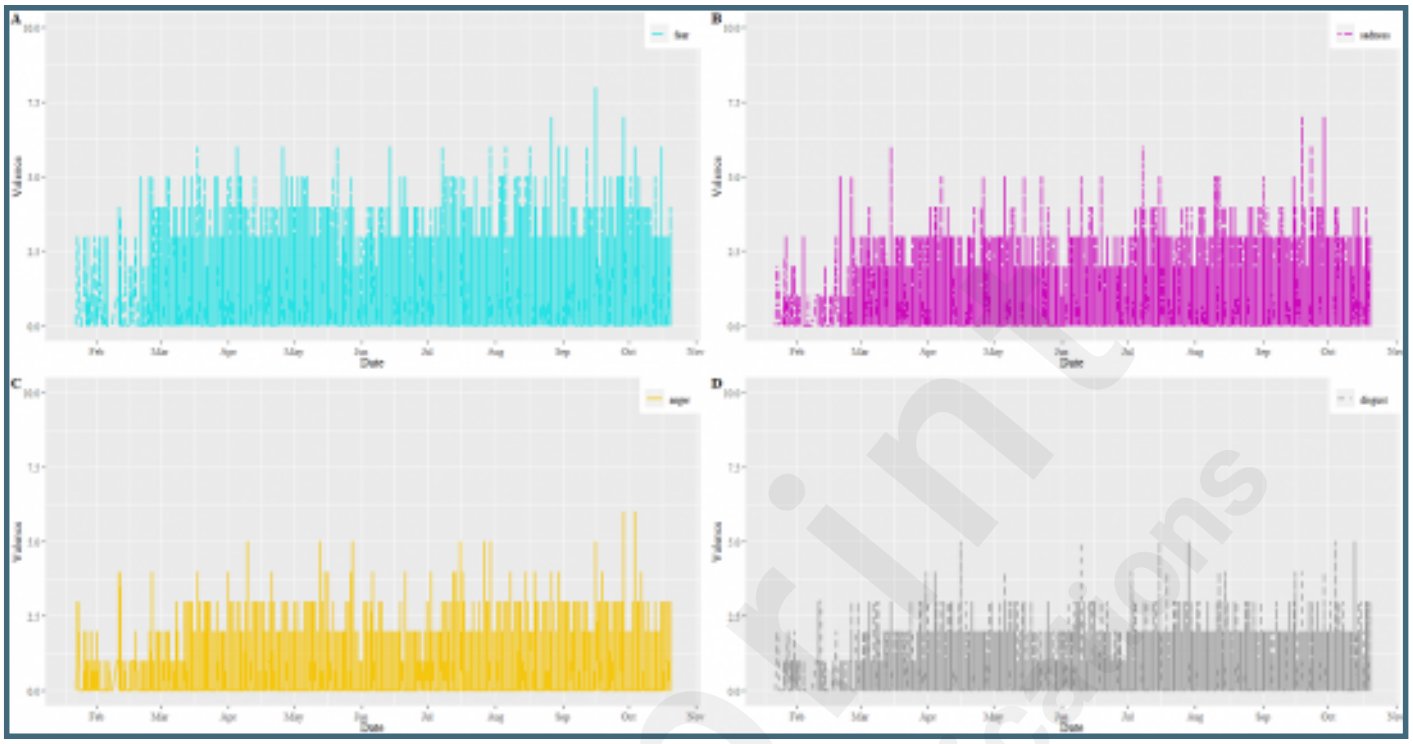
Distributions of sentiment valences between Jan 2020 and Oct 2020.



Distributions of emotion valences regarding anticipation (A), trust (B), joy (C), and surprise (D) between Jan 2020 and Oct 2020.



Distributions of emotion valences regarding fear (A), sadness (B), anger (C), and disgust (D) between Jan 2020 and Oct 2020.



Multimedia Appendixes

Number of tweets collected between January 22 and October 20, 2020.
URL: https://asset.jmir.pub/assets/a984bec433e708f151b3e880578805da.doc

Plot of word tokens against counts sorted.
URL: https://asset.jmir.pub/assets/00d1d6714c93cbab3d89269d4bda8f7f.doc

Plot of word pairs against counts sorted.
URL: https://asset.jmir.pub/assets/b037d6738eaa9a9d18ae03ae7e8ea625.doc

Plot of number of topics against LDA tuning scores.
URL: https://asset.jmir.pub/assets/7ee7dfca19d98a559e251aa2fe894698.doc

Top 100 probability (beta) distributions of words in each topic.
URL: https://asset.jmir.pub/assets/1f92c440a6dc03c292b8bd9cc60a9806.docx

Top 20 probability (theta) distributions of topic 1 in tweet samples.
URL: https://asset.jmir.pub/assets/1fa280318cd0c5315cc9ccd4a5b55bcd.docx

Top 20 probability (theta) distributions of topic 2 in tweet samples.
URL: https://asset.jmir.pub/assets/6cd64570abb4b7cdb40719a1a0a80b10.docx

Top 20 probability (theta) distributions of topic 3 in tweet samples.
URL: https://asset.jmir.pub/assets/be448218793b77a6b292028d01df6a04.docx