DOI: 10.1002/pra2.233

SHORT PAPERS

Uncovering temporal differences in COVID-19 tweets

Han Zheng | Dion H.-L. Goh | Chei S. Lee | Edmund W. J. Lee | Yin L. Theng

Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

Correspondence
Han Zheng, Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore.
Email: han019@ntu.edu.sg

Abstract

In the fight against the COVID-19 pandemic, understanding how the public responds to various initiatives is an important step in assessing current and future policy implementations. In this paper, we analyzed Twitter tweets using topic modeling to uncover the issues surrounding people's discussion of the disease. Our focus was on temporal differences in topics, prior and after the declaration of COVID-19 as a pandemic. Nine topics were identified in our analysis, each of which showed distinct levels of discussion over time. Our results suggest that as the pandemic progresses, the concerns of the public vary as new developments come to light.

KEYWORDS

COVID-19, pandemic, temporal differences, topic modeling, twitter

1 | INTRODUCTION

In December 2019, pneumonia of unknown cause was reported in Wuhan, China. Initially thought to be a localized problem, this disease has now been declared the COVID-19 pandemic, infecting about 2 million people worldwide and claiming more than 120,000 lives as of April 2020 (World Health Organization [WHO], 2020b). To stem the progress of COVID-19, governments around the world have instituted a variety of measures including lockdowns of various degrees, public health campaigns, work from home initiatives and online learning.

A critical prong in the multifaceted fight against this disease is the behavior of the public in conforming to government directives as well as taking various protective measures such as washing of hands and social distancing (World Health Organization [WHO], 2020a). Hence, understanding how the public responds to COVID-19 initiatives is an important step in assessing current policy implementations and guiding future policy development. Here, social media postings have the potential to provide a glimpse into people's responses to the disease as numerous messages urging positive public health behaviors have emerged on various platforms, along with news updates, personal opinions and anecdotes.

The present research aims to uncover the issues surrounding the discussion of COVID-19 on Twitter. We employ topic modeling in our analysis of Twitter tweets. This technique facilitates the automated discovery of patterns that reflect the underlying topics in a corpus of documents (Sharma & Sharma, 2017). Of particular interest are temporal differences in topics, notably prior to the declaration of COVID-19 as a pandemic by the WHO on March 11, 2020, and after this characterization. Although the disease had already reached high levels of spread and severity worldwide before 11 March, the pandemic label would have presumably spurred governments and individuals to pay more attention to COVID-19 and adopt measures to curb it.

There are two reasons for using Twitter tweets in our research. First, this social media platform is currently in active use by governments, organizations and individuals for COVID-19 information sharing. Second, it is arguably an important source of information, and has been used

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

in studies of other disease outbreaks (e.g. Signorini, Segre, & Polgreen, 2011).

2 | RELATED WORK

User-generated postings such as tweets are excellent sources of public health information (Sinnenberg et al., 2017). As compared to traditional public health surveillance methodologies (e.g. surveys), the data from Twitter have the advantages of being "naturally occurring", inexpensive to get, and contain high velocity granular data (Lee & Yee, 2020). The act of tweeting reflects the degree of public attention and collective public sentiments toward certain health issues, and thus would provide potentially useful leading signals for public health researchers to act on (Kuehn, 2015). In the context of infectious diseases, Twitter data has been used to understand and map the spread of malaria (Fung et al., 2017), H1N1 (Chew & Eysenbach, 2010), and Ebola (Liang, 2018), to name a few examples.

By incorporating temporal components when analyzing tweets, one could uncover critical variations in the spread of COVID-19 information down to a granular level, such as the evolution of discussions on specific days as the disease spreads, and monitor the spread of the disease (Chen, Hossain, Butler, Ramakrishnan, & Prakash, 2016). This allows researchers to effectively engage in dissemination science, by enabling public health organizations to be targeted in developing strategic messaging efforts. After all, past research has documented that time matters when examining tweets in public health contexts, and temporal distribution of COVID-19 information could provide a nuanced understanding of how people communicate, which text alone cannot give (Stefanidis et al., 2017).

3 | METHODOLOGY

The dataset used in this study was from an ongoing project that actively collected COVID-19 tweets from January 28, 2020 (Chen, Lerman, & Ferrara, 2020), leveraging Twitter's search API with a list of keywords and accounts related to COVID-19 (e.g., "coronavirus", "corona", "Covid-19", "Covid"). Until April 10, 2020, this project had collected around 94.67 million tweets. Since we focused on the tweets before and after the declaration of COVID-19 as a pandemic on March 11, 2020, we selected two weeks of tweets between March 4, 2020 and March 18, 2020. The project only released the Tweet IDs of the collected tweets. Thus, we used the software Hydrator to extract the tweets for this timeframe (Summers, 2017). There was a total of 18.8 million tweets during this two-week period, and the number of tweets per day ranged from 913,230 to 3,408,778. Due to the large data size and to facilitate processing, we randomly sampled 5% of the tweets on each day, and the final samples constituted 940,837 tweets. This random sampling approach is consistent with prior research (e.g. Cavazos-Rehg et al., 2016; DiGrazia, McKelvey, Bollen, & Rojas, 2013).

Data were analyzed using R statistical software version 3.5.1. First, we eliminated non-English tweets and duplicate tweets in the dataset. Next, we preprocessed the tweets by removing the "RT" (retweet) text and usernames, URL links, punctuations, and numbers. We tokenized the tweets into single words and converted all words to lower case. Further, we removed a list of standard stopwords such as "the," "is," and "are," plus additional stopwords that frequently appeared in the tweets (e.g., "COVID-19," "coronavirus," "virus," etc.). Also, we used the Porter stemmer to stem the words into their root forms. Finally, to reduce the dimensionality of data, we removed sparse terms that did not appear very often. After preprocessing, 258,290 valid English tweets that consisted of 1,450,595 words and 1,509 unique words were used for further analysis.

Latent Dirichlet Allocation (LDA) topic modelling was employed to identify the common COVID-19 topics discussed on Twitter. It is an unsupervised machine learning method to uncover the hidden semantic structures from a given textual corpus and assign individual documents to a fixed set of topics (Blei, Ng, & Jordan, 2003). We used the Gibbs sampling algorithm as it allows iterative steps through configurations to estimate optimal model fit (Geman & Geman, 1984). To select the best number of topics for the corpus, we ran several models ranging from 2 to 20, in intervals of 1. For the quality evaluation of these models, we considered two data-driven metrics (Cao, Xia, Li, Zhang, & Tang, 2009; Deveaud, SanJuan, & Bellot, 2014) and interpretability of the topics in each model. Cao et al.'s (2009) metric suggests that when the average cosine distance of topics reaches the minimum, the LDA model performs best. Deveaud et al.'s (2014) metric posits that the optimal number of topics would be the one with the maximum information divergence. The analyses resulted in a decision to run LDA with nine topics for the corpus.

4 | RESULTS

Table 1 shows the nine topics derived from our LDA topic modelling. To manually assign topic names, the top 10 terms based on beta values in each topic were taken

TABLE 1 Nine topics generated by the LDA topic modeling

Topic label	Top 10 words in the topic	Rate %	Example
Topic 1: Mortality of COVID-19	Peopl, infect, die, flu, mani, kill, rate, million, disease, risk	15.37	"*Coronavirus has so far killed 27 people in the US, 19 of which were at one senior living facility*." (March 10)
Topic 2: Origin of COVID-19	China, world, countri, live, chines, start, Wuhan, show, thank, read	14.41	"*Wuhan Huanan Seafood Market was claimed to be the origin of COVID2019*." (March 4)
Topic 3: Preventive measures	Spread, take, hand, use, stop, home, need, stay, keep, way	12.34	"*The CDC says you should avoid shaking hands due to coronavirus during a press conference...*" (March 14)
Topic 4: Trumps' responses to pandemic	Trump, pandem, respons, presid, call, american, realdonaltrump, media, news, lie	12.05	"*The Trump administration explains that the Europe ban trump announced does not only exempt the UK but...*" (March 12)
Topic 5: Reporting of new cases	Case, new, death, state, report, first, confirm, break, itali, update	10.04	"*9 coronavirus deaths now reported in Washington state and only 27 confirmed cases*." (March 4)
Topic 6: Organizing healthcare resources	Test, health, posit, cdc, public, quarantin, social, patient, care, hospital	9.38	"*Patients seeking information on coronavirus are being asked to use NHS111 online for general information...*" (March 4)
Topic 7: Coping with pandemic	Well, hope, happen, man, worri, talk, feel, love, look, better	10.24	"*I pray for everyone! I ask that we all have good healthy bodies! Please heal those who are sick and suffering!*" (March 9)
Topic 8: Reports of lockdowns	Due, outbreak, close, cancel, week, school, travel, fear, concern, day	8.68	"*Italy will shut all schools from today for 10 days...*" (March 5)
Topic 9: Government help and support	Work, govern, need, help, global, sick, hous, fight, busi, support	7.47	"*The emergency coronavirus package as summarized by today's house vote schedule*." (March 4)

into account. A beta value refers to the probability of a term belonging to a given topic. Thus, a higher beta value indicates the term can better describe the topic.

In addition, we examined tweets in each topic to help in the labeling. To illustrate, for topic 3, the key words were "hand," "home," and "stay." The focus of this topic might be related to preventive measures in response to COVID-19 such as washing hands and staying home. We thus examined the associated tweets for topic 3. For example, one user on 14 March wrote that "The CDC says you should avoid shaking hands due to coronavirus during a press conference..." Similarly, another user posted "Wearing a face mask when you have a cold or flu should become the norm as it is in Japan." on 4 March. As such, we labelled topic 3 as "Preventive measures." In this way, we assigned names to the other eight topics as presented in Table 1.

Next, we sought to uncover temporal differences in the COVID-19 tweets. First, as the number of tweets varied across the days, we divided the number of tweets in each topic per day by the total number of tweets per day to get a topic weightage score for each day. Second, we visualized the trend of how each topic weightage changed during the 2 weeks (see Figure 1).

Overall, compared to the week prior to the pandemic declaration on March 11, there were more discussions on preventive measures (topic 3), organizing healthcare resources (topic 6), and government help and support (topic 9) in the second week. In contrast, less attention was paid on mortality rates of COVID-19 (topic 1) and reporting of new cases (topic 5) after the declaration. Interestingly, discussions on topic 2 (origin of COVID-19) and topic 4 (Trumps' responses to pandemic) fluctuated before it experienced a sharp increase on March 17. Finally, topic 7 (coping with the pandemic) and topic 8 (reports of lockdowns) had a steady increase in the first week and reached its peak after the declaration, followed by a sharp decrease thereafter.

5 | DISCUSSION

We found that the topics generated reflect the diversity of the narratives surrounding COVID-19. With the rapidly

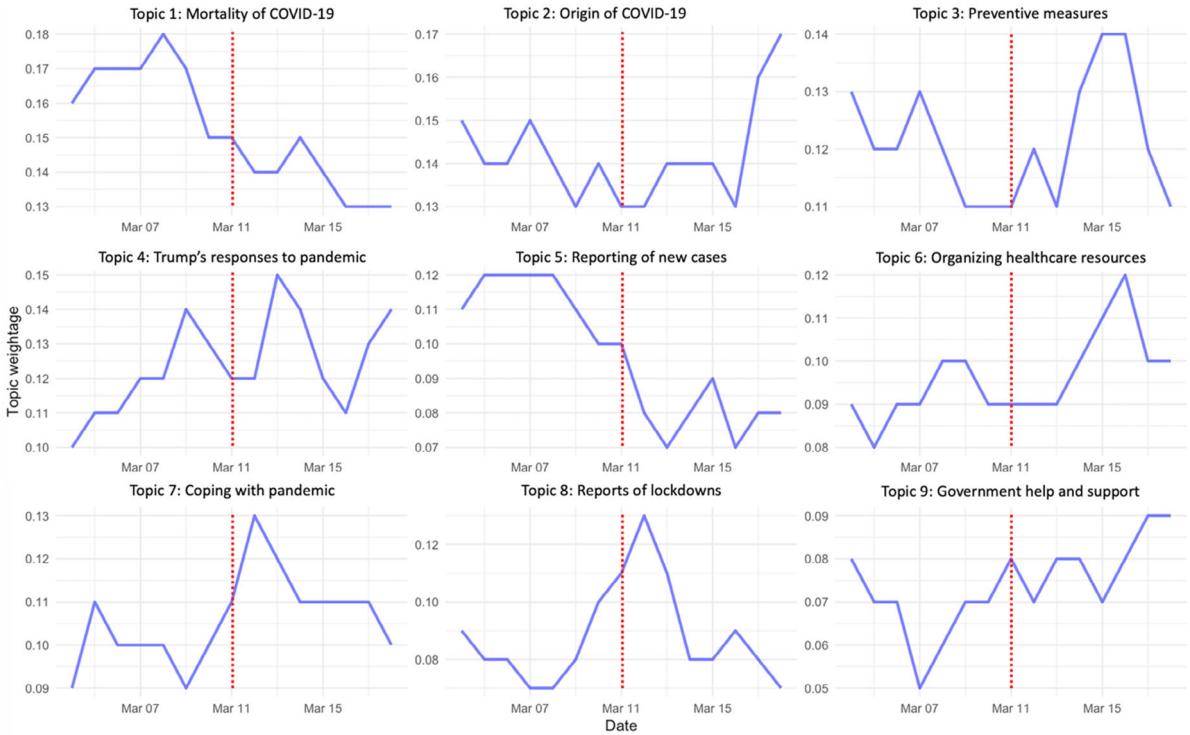

FIGURE 1 Topic change with time

changing situation, these topics were constantly evolving in response to what was happening in real life. Put differently, our results revealed what was on people's minds as well as the social conversations surrounding the pandemic declaration.

One interesting finding is that the discussions of topics reflected the volatility and social effects of COVID-19. For example, one of the topics we found was the origins of the virus (topic 2). A possibility for the interest is that there were multiple narratives on this topic, and the truth of the origin remains elusive. The volume of discussion on this topic was initially high but waned until March 11, 2020. However, the real-life political exchanges and tensions between the US and Chinese officials as well as the US president labelling the pandemic a "Chinese virus" on March 17, 2020 likely triggered more attention on this topic on Twitter. This suggests that discussions on Twitter are influenced by reports from mainstream media. In particular, as the pandemic evolved during our period of analysis, new issues were reported in the mainstream media, triggering discussions on Twitter.

Our results also demonstrate that people depend on social media platforms (Twitter in this study) to meet various needs during times of uncertainty and crisis. Before the pandemic declaration on March 11, 2020, discussions centered mainly around informational exchanges such as COVID-19 mortality (topic 1) and reports of new cases (topic 5). After 11 March, conversations were not only informational, but were also emotional where people supported each other during the lockdown (topic 8) and helped each other cope with the pandemic (topic 7). This finding is consistent with the notion of audience-media dependency (Ball-Rokeach & DeFleur, 1976; Lee, 2012) in which an audience is impacted not only by media content but also by the society in which they consume the content.

To conclude, our findings suggest that social media platforms such as Twitter play important roles to meet people's needs during the pandemic. Next, discussions are influenced by what people read in the mainstream media and possibly other sources (e.g. Topic 4 and 5). Hence it is essential that these platforms put in place fact-checking mechanisms quickly to reduce ambiguity and misinformation. Further, our results show that government and other decision-makers may use Twitter to uncover ongoing discussions that may help craft official responses to ongoing developments or chart new policy directions.

A limitation of our research is that due to the large volume of data, we were not able to analyze all the tweets. Consequently, the topics uncovered may deviate from the themes that people actually discussed online. Further, our nine topics are a two-week snapshot of Twitter discussions that may not capture new conversation topics as the pandemic develops over time. Other social

media platforms may also yield different sets of topics. Hence, it would be worthwhile to analyze new tweets as they become available as well as content from other social media platforms to ascertain the stability of our nine topics. Finally, because there were differences in how countries responded to COVID-19, it would be interesting to examine geographical variations in discussions of the disease.

REFERENCES

Ball-Rokeach, S. J., & DeFleur, M. L. (1976). A dependency model of mass-media effects. *Communication Research*, *3*(1), 3–21.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3* (Jan), 993–1022.

Chen, L., Hossain, K. S. M. T., Butler, P., Ramakrishnan, N., & Prakash, B. A. (2016). Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery*, *30*, 681–710.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72* (7–9), 1775–1781.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One*, *5*(11), 1–13.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. J. (2016). A content analysis of depression-related tweets. *Computers in Human Behavior*, *54*, 351–357.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public Coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2), e19273.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS One*, *8*(11), e79449.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84.

Fung, I. C. H., Jackson, A. M., Ahweyevu, J. O., Grizzle, J. H., Yin, J., Tsz, Z. H. Z., ... Fu, K. W. (2017). #Globalhealth twitter conversations on #malaria, #HIV, #TB, #NCDs, and #NTDS: A cross-sectional analysis. *Annals of Global Health*, *83*(3–4), 682–690.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741.

Kuehn, B. M. (2015). Twitter streams fuel big data approaches to health forecasting. *JAMA*, *314*(19), 2010–2012.

Lee, C. S. (2012). Exploring emotional expressions on YouTube through the lens of media system dependency theory. *New Media Society*, *14*(3), 457–475.

Lee, E. W. J., & Yee, A. Z. H. (2020). Toward data sense-making in digital health communication research: Why theory matters in the age of big data. *Frontiers in Communication*, *5* (11), 1–10.

Liang, H. (2018). Broadcast versus viral spreading: The structure of diffusion cascades and selective sharing on social media. *Journal of Communication*, *68*(3), 525–546.

Sharma, H., & Sharma, A. K. (2017). Study and analysis of topic modelling methods and tools – A survey. *American Journal of Mathematical and Computer Modelling*, *2*, 84–87.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic. *PLoS One*, *6*(5), e19467.

Summers, E. (2017). DocNow hydrator for tweets. GitHub repository. Retrieved from https://github.com/DocNow/hydrator.

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, *107*(1), e1–e8.

Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P. L., Jacobsen, K. H., ... Crooks, A. (2017). Zika in twitter: Temporal variations of locations, actors, and concepts. *JMIR Public Health Surveillance*, *3*(2), e22.

World Health Organization (WHO). (2020a). Coronavirus disease (COVID-19) advice for the public. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public.

World Health Organization (WHO). (2020b). Coronavirus disease (COVID-19) pandemic. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

How to cite this article: Zheng H, Goh DH-L, Lee CS, Lee EWJ, Theng YL. Uncovering temporal differences in COVID-19 tweets. *Proc Assoc Inf Sci Technol*. 2020;57:e233. https://doi.org/10.1002/pra2.233