Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling

J Gen Intern Med
DOI: 10.1007/s11606-020-05889-w

INTRODUCTION

Given the rapidly changing nature of the coronavirus disease 2019 (COVID-19) pandemic, real-time monitoring of COVID-19 cases and deaths has been widely embraced.[1] The pandemic has also been accompanied by an "infodemic," an overabundance of information and misinformation.[2] Public response to the pandemic and infodemic is important, but undermeasured.[3] Real-time analysis of public response could lead to earlier recognition of changing public priorities, fluctuations in wellness, and uptake of public health measures, all of which carry implications for individual- and population-level health.[3] To test this hypothesis, we measured daily changes in the frequency of topics of discussion across 94,467 COVID-19-related comments on an online public forum in March, 2020.

METHODS

Reddit is the 19th most popular website in the world with 420 million monthly active users.[4] Between March 3 and March 31, 2020, we obtained all comments from the "Daily Discussion Post" on "r/Coronavirus," the most popular COVID-19 subreddit with 1.9 million members. We defined 50 discussion topics, groups of commonly co-occurring words, using a machine learning based approach to natural language processing, latent Dirichlet allocation (LDA).[5]

For each of the 50 topics, we reviewed the ten words and comments most associated with each topic.[6] We identified topics that fell into three categories of interest: response to public health measures, impact on daily life, and sense of pandemic severity. We tracked daily variations in the average prevalence of topics across all comments. In order to improve visualization of patterns of topic change, we used locally estimated scatterplot smoothing (LOESS) lines. To quantify the degree of change in prevalence, we compared 4-day periods using the two-proportion z-test. We used R version 3.6.1 for all analyses. All data was publicly available, and the study was considered exempt under University of Pennsylvania Institutional Review Board guidelines.

RESULTS

In the 29 days between March 3 and March 31, we collected 94,467 posts from r/Coronavirus daily discussion threads, with peak activity between March 15 and 17 (16% of comments). Of the 50 LDA topics (available by request), ten pertained to the three categories of interest. Other topics included those related to news sharing, political discussions, and discussions about the science of COVID-19. Table 1 shows key topic words and representative comments, and Figure 1 displays the change in topic frequency over time by category. In the "public health measures" category, for instance, "hand washing" became less prevalent throughout March (2.7% from March 3 to March 6 vs 1.9% from March 28 to March 31, $p < .001$; two-proportion z-test). "Impact on daily life" topics showed "travel" peaking early and dropping throughout the month (3.2% March 3–March 6 vs 1.0% March 28–March 31, $p < .001$) and concern regarding "personal finances" increasing (1.5% March 3–March 6 vs 2.1% March 28–March 31, $p = .003$). "Sense of pandemic severity" evolved over the month, with fewer comments comparing COVID-19 with the flu (2.3% March 3–March 6 vs 1.8% March 28–March 31, $p = .04$) and mid- to late-month growth in comments reporting numbers of cases and deaths (2.1% March 12–March 15 vs 2.7% March 28–March 31, $p = .001$).

DISCUSSION

This analysis indicates that longitudinal topic modeling of Reddit content is effective in identifying patterns of public dialogue and could be used to guide targeted

Prior Presentations *These data have not been described or published elsewhere.*

Received April 20, 2020
Accepted April 28, 2020

Published online: 12 May 2020

Table 1 Latent Dirichlet Allocation Topics from a Coronavirus Subreddit Throughout March, 2020, with a Collection of Top Words Used to Define the Topic and a Redacted Representative Reddit Comment

Topic Category	Topic	Top words	Redacted representative Reddit comment (to preserve user anonymity)
Public health measures	Hand washing	hands, wash, touch, use, water, soap	"At least get them to wash hands as soon as they get back and wash clothes"
	Outdoor safety	stay, people, away, home, outside, safe	"It's okay to go for a walk, just try to stay at least 6 feet from others."
	Masks	masks, wear, face, n95, use, make	"What type of filter to insert in a cotton mask? Ordering some cotton masks with an insert to add a filter. Would an air conditioner filter work?"
Daily life impact	Food and supplies	food, grocery, people, store, toilet, buy	"Just went to my local grocery store this morning. The place was packed with folks… saw a ton of people buying paper towels, toilet paper etc.that aisle was almost empty."
	Travel changes	travel, back, trip, US, flight, cancel	"Going to a wedding in Canada next month. What are the odds travel is banned between the last weeks of April?"
	School closing	school, closed, still, public, kids, university	"Gov has closed all K-12 schools in [state] starting Monday until early April."
	Personal finances	work, get, pay, money, need, help	"My work just closed until further notice. I work in food service industry. What are my options for govenrment financial assistance? I do not have paid sick leave or paid time off."
Sense of pandemic severity	Number of cases and deaths	cases, number, deaths, new, confirmed	"So if these numbers are correct, US is now third in total cases behind China and Italy, and FIRST in new cases, surpassing Italy. And we are supposed to be ~10 days behind Italy."
	Comparison to flu	flu, like, coronavirus, much, bad, worse	"There is no way this virus is as bad as people are saying it is. Do not about 61,000 people die every year from flu?"
	Danger to elderly	rate, death, mortality, age, higher, risk	"The case fatality rate in Italy was 1.0%, but with a much more elderly population, in which coronavirus death rate is much higher"

interventions. For instance, comparisons to the flu were embraced by the public. Early recognition of this reality could have led to more specific information dissemination campaigns and earlier public acknowledgement of disease severity. Questions about safely spending time outdoors peaked in mid-March, representing a missed opportunity for public guidance. Tracking and responding proactively to common questions, such as what material is best used for a homemade mask, may minimize the spread of misinformation. Notably missing from these Reddit topics were discussions of contact tracing, a growing area of public concern. Limitations of this study include that Reddit users are not representative of all segments of the population, and that Reddit data is not associated with a geographic location. Real-time monitoring of online COVID-19 dialogue holds promise for more dynamically understanding and responding to needs in public health emergencies.

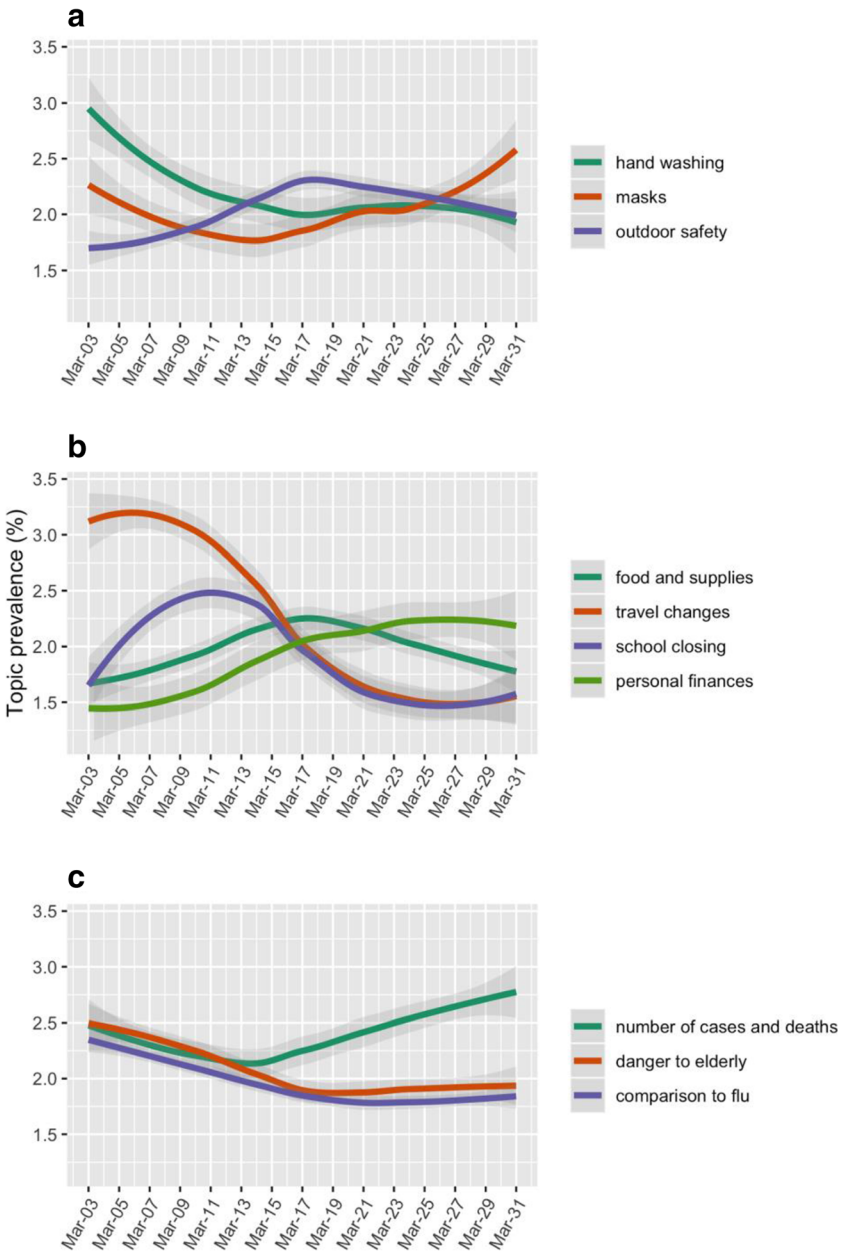

Fig. 1 The change in the prevalence over the month of March, 2020, in Reddit comment content related to a public health measures, b daily life impact, and c sense of pandemic severity. Lines show locally estimated scatterplot smoothing (LOESS) for the daily average prevalence of the topic across all comments; shaded grey area represents the standard error of the LOESS estimation.

Daniel C. Stokes, MS
Anietie Andy, PhD
Sharath Chandra Guntuku, PhD
Lyle H. Ungar, PhD
Raina M. Merchant, MD
Penn Medicine Center for Digital Health, University of Pennsylvania,
Philadelphia, PA, USA

Daniel C. Stokes, MS
Raina M. Merchant, MD
Center for Emergency Care Policy and Research, Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA

Sharath Chandra Guntuku, PhD
Lyle H. Ungar, PhD
Department of Computer and Information Science, University of Pennsylvania,
Philadelphia, PA, USA

Corresponding Author: *Daniel C. Stokes, MS; Penn Medicine Center for Digital Health, University of PennsylvaniaPhiladelphia, PA, USA (e-mail: daniel.stokes@pennmedicine.upenn.edu).*

Compliance with Ethical Standards:

The study was considered exempt under University of Pennsylvania Institutional Review Board guidelines.

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Dong E**, **Du H**, **Gardner L**. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;0(0). doi:https://doi.org/10.1016/S1473-3099(20)30120-1
2. **Zarocostas J.** How to fight an infodemic. *Lancet* 2020;395(10225):676. doi:https://doi.org/10.1016/S0140-6736(20)30461-X
3. **Merchant RM**, **Lurie N**. Social Media and Emergency Preparedness in Response to Novel Coronavirus. *JAMA*. March 2020. doi:https://doi.org/10.1001/jama.2020.4469
4. The top 500 sites on the web. Alexa: an amazon.com company. https://www.alexa.com/topsites. Accessed April 11, 2020.
5. **David M. Blei**, **Andrew Y. Ng**, **Michael I. Jordan**. Latent Dirichlet Allocation. *J Mach Learn Res* 2003;3(January):993-1022.
6. **Guntuku SC**, **Schneider R**, **Pelullo A**, et al. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open* 2019;9(11):e030355. doi:https://doi.org/10.1136/bmjopen-2019-030355