Tracking COVID-19 Discourse on Twitter in North America: Topic Modeling and Aspect-based Sentiment Analysis

Hyeju Jang, Emily Rempel, David Roth, Giuseppe Carenini, Naveed Z. Janjua

Submitted to: Journal of Medical Internet Research
on: November 02, 2020

Table of Contents

Tracking COVID-19 Discourse on Twitter in North America: Topic Modeling and Aspect-based Sentiment Analysis

Hyeju Jang[1, 2] MSc, PhD; Emily Rempel[2] MSc, PhD; David Roth[2] MSc, PhD; Giuseppe Carenini[1] MSc, PhD; Naveed Z. Janjua[2, 3, 4] MBBS, MSc, DrPH

[1]Department of Computer Science University of British Columbia Vancouver CA
[2]British Columbia Centre for Disease Control Vancouver CA
[3]School of Population and Public Health University of British Columbia Vancouver CA
[4]Centre for Health Evaluation and Outcome Sciences Vancouver CA

Corresponding Author:
Hyeju Jang MSc, PhD
Department of Computer Science
University of British Columbia
ICICS/CS Building 201-2366 Main Mall
Vancouver
CA

Abstract

Background: Social media is a rich source where we can learn about people's reactions to social issues. As COVID-19 has significantly impacted on people's lives, it is essential to capture how people react to public health interventions and understand their concerns.

Objective: We aim to investigate people's reactions and concerns about COVID-19 in North America, especially focusing on Canada.

Methods: We analyze COVID-19 related tweets using topic modeling and aspect-based sentiment analysis (ABSA), and interpret the results with public health experts. To generate insights on the effectiveness of specific public health interventions for COVID-19, we compare timelines of topics discussed with timing of implementation of interventions, synergistically including information on people's sentiment about COVID-19 related aspects in our analysis. In addition, to further investigate anti-Asian racism, we compare timelines of sentiments for Asians and Canadians.

Results: Topic modeling identified 20 topics and public health experts provided interpretations of the topics based on top-ranked words and representative tweets for each topic. The interpretation and timeline analysis showed that the discovered topics and their trend are highly related to public health promotions and interventions, such as physical distancing, border restrictions, hand washing, staying-home, and face coverings. After training the data using ABSA with human-in-the-loop, we obtained 545 aspect terms (e.g., "vaccines", "economy", and "masks") and 60 opinion terms (e.g., "infectious"- negative, and "professional"- positive), which were used for inference of sentiments of 20 selected aspects. The results showed negative sentiments related to overall outbreak, misinformation, and Asians and positive sentiments related to physical distancing.

Conclusions: Analyses using Natural Language Processing (NLP) techniques with domain expert involvement can produce useful information for public health. This study is the first to analyze COVID-19 related tweets in Canada in comparison with tweets in the United States by using topic modeling and human-in-the-loop domain-specific aspect-based sentiment analysis. This kind of information could help public health agencies to understand public concerns as well as what public health messages are resonating in our populations who use Twitter, which can be helpful for public health agencies when designing a policy for new interventions.

(JMIR Preprints 02/11/2020:25431)
DOI: https://doi.org/10.2196/preprints.25431

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

Tracking COVID-19 Discourse on Twitter in North America: Topic Modeling and Aspect-based Sentiment Analysis

Hyeju Jang1, 2, Emily Rempel[2], David Roth[2], Giuseppe Carenini[1], Naveed Z. Janjua[2, 3, 4]

[1]Department of Computer Science, University of British Columbia

[2]British Columbia Centre for Disease Control

[3]School of Population and Public Health, University of British Columbia, Vancouver, Canada

[4]Centre for Health Evaluation and Outcome Sciences (CHÉOS), Vancouver, Canada

{hyejuj, carenini}@cs.ubc.ca
{emily.rempel, David.roth, naveed.janjua}@bccdc.ca

Abstract

Background: Social media is a rich source where we can learn about people's reactions to social issues. As COVID-19 has significantly impacted on people's lives, it is essential to capture how people react to public health interventions and understand their concerns.

Objective: We aim to investigate people's reactions and concerns about COVID-19 in North America, especially focusing on Canada.

Methods: We analyze COVID-19 related tweets using topic modeling and aspect-based sentiment analysis (ABSA), and interpret the results with public health experts. To generate insights on the effectiveness of specific public health interventions for COVID-19, we compare timelines of topics discussed with timing of implementation of interventions, synergistically including information on people's sentiment about COVID-19 related aspects in our analysis. In addition, to further investigate anti-Asian racism, we compare timelines of sentiments for Asians and Canadians.

Results: Topic modeling identified 20 topics and public health experts provided interpretations of the topics based on top-ranked words and representative tweets for each topic. The interpretation and timeline analysis showed that the discovered topics and their trend are highly related to public health promotions and interventions, such as physical distancing, border restrictions, hand washing, staying-home, and face coverings. After training the data using ABSA with human-in-the-loop, we obtained 545 aspect terms (e.g., "vaccines", "economy", and "masks") and 60 opinion terms (e.g., "infectious"- negative, and "professional"- positive), which were used for inference of sentiments of 20 selected aspects. The results showed negative sentiments related to overall outbreak, misinformation, and Asians and positive sentiments related to physical distancing.

Conclusions: Analyses using Natural Language Processing (NLP) techniques with domain expert involvement can produce useful information for public health. This study is the first to analyze COVID-19 related tweets in Canada in comparison with tweets in the United States by using topic modeling and human-in-the-loop domain-specific aspect-based sentiment analysis. This kind of information could help public health agencies to understand public concerns as well as what public health messages are resonating in our populations who use Twitter, which can be helpful for public health agencies when designing a policy for new interventions.

Keywords: COVID-19; twitter; topic modeling; aspect-based sentiment analysis; racism; Anti-Asians; Canada; North America

Introduction

Globally, more than 31 million people have been diagnosed with COVID-19 infection and more than 1 million people have died as of October 12, 2020. Since there is no vaccine or effective treatment yet, governments across the world have implemented wide-ranging non-pharmaceutical interventions such as hand hygiene, face masks, contact tracing, isolation and quarantine, and physical (social) distancing through banning mass gatherings and lockdowns to reduce the transmission of SARS-CoV-2. Impact of COVID-19 and measures to prevent transmission has generated a lot of discussion among general population, medical and public health professionals and government officials. Some of this discourse is happening in social media such as Twitter.

During this pandemic, people have been using social media such as Twitter to share news, information, opinions and emotions about COVID-19 [1,2], similar to previous infectious disease outbreaks such as Ebola. In the Eloba outbreak, public health organizations helped contain Ebola by monitoring conversations on social media and spreading accurate information about the disease [3-6]. As we can see from these past successes, social media is an important source to learn about people's reactions and concerns. This information can assist public health authorities in monitoring and surveillance of health information, concerns and behaviors, and designing interventions to reduce impact of pandemic. Understanding both people's information needs, misinformation, hate speech/discrimination, compliance with preventative measures and other reactions to COVID-19 and where their concerns lie helps to tailor public health strategy and ultimately create better informed interventions.

Topic modeling and sentiment analysis have been widely used to identify issues and people's opinions in public health and is being used to understand COVID-19 related issues as well (Table 1). Analyses were conducted to identify patterns of health communications in diverse kinds of data sources and communities/locations. While some works investigated news articles [7] or research papers [8], most research focused on social media such as Reddit posts [9] and tweets [10-16]. Conversations in particular communities were examined such as tweets posted by U.S. governors and presidential cabinet members [10] and African American twitter communities [13]. Specific languages and locations were discussed as well, e.g., Chinese news articles [7], Persian/Farsi tweets in Iran [11] and English tweets in California and New York, US [14]. Although all these works investigated people's reactions towards COVID-19, there have been few studies about general public responses in Canada. Furthermore, although sentiment analysis has been broadly used [12-15], the techniques employed in prior work determine sentiment of an overall text rather than capturing opinions towards COVID specific aspects chosen by domain experts, and exploit lexicon built in general domains overlooking that a word's sentiment depends on the domain or context where it is used [17].

Our study aims to investigate Twitter users' reactions to COVID-19 in North America, especially focusing on Canada. We analyze COVID-19 related tweets with topic modeling and Aspect-Based Sentiment Analysis (ABSA) using human-in-the- loop, and interpret the results with public health experts. We examine the sentiment of tweets about COVID-19 related aspects such as social distancing and masks, by using ABSA based on domain specific aspect and opinion terms. The key advantage of our work is that public health experts are actively involved in the computational process with the specific goal of informing public health interventions. Our results are interpreted by these public health experts, and we use a human-in-the-loop ABSA approach to obtain domain specific aspect and opinion terms. To the best of our knowledge, we are the first to directly identify sentiment of COVID-specific aspects.

Table 1 Related work on topic modeling of COVID-19 related data.

Authors	Source	Posters	Time	Location	Language	sentiment
Liu et al. [7]	News articles	News reporters	Jan. 1 – Feb. 20, 2020	Not specified	Chinese	No
Dong et al. [8]	Research papers	Researchers	Unknown – Mar. 20, 2020	Not specified	English	No
Stokes et al. [9]	Reddit posts	public	Mar. 3 – Mar. 31, 2020	Not specified	English	No
Sha et al. [10]	Tweets	State governors, presidential cabinet members, and the president	Jan. 1 – Apr. 7, 2020	U.S.	English	No
Hosseini et al. [11]	Tweets	public	Mar. 13 – Apr. 19, 2020	Iran	Persian/Farsi	No
Sharma et al. [12]	Tweets	public	Mar. 1 – Marcy. 30, 2020	Not specified	English	Yes
Odlum et al. [13]	Tweets	public (African Americans)	Jan. 21 – May 3, 2020	Not specified	English	Yes
Wang et al. [14]	Tweets	public	Mar. 5 – Apr. 2, 2020	California and New York, U.S.	English	Yes
Abd-Alrazaq et al. [15]	Tweets	public	Feb. 2 – Mar. 15, 2020	Not specified	English	Yes
Ordun et al. [16]	Tweets	public	Mar. 24 – Apr. 9, 2020	Not specified	English, Spanish, Italian, French,	No

					Portuguese	
Ours	Tweets	Public	Jan. 21 – May. 31, 2020	Canada and U.S.	English	Yes

Methods

Data and data processing

We use a public twitter dataset about the COVID-19 pandemic, collected by [18] using numerous COVID-19 related keywords such as “coronavirus”, “COVID-19” and “pandemic”. The data collection started on January 28, 2020 (tweets from January 21, 2020), and is still ongoing, which has published over 123 million tweets by May 11, 2020.

For our work, we select tweets from January to May, whose location is Canada and United States (US).[a] Among the 372,711 tweets in total (Canada: 30,235, US: 342,476), we only include tweets written in English using tweet metadata and the spacy-langdetect toolkit[b]. This process resulted in 319,524 tweets in total, 25,595 for Canada, and 293,929 for US. To remove tweet specific keywords and URLs, we use the tweet-preprocessor toolkit[c]. We do not remove hashtags and mentions because they can be informative for our work. We lowercase, tokenize using the Spacy toolkit [19]
. Since the methods we use in this paper are all unsupervised, we do not split the data for training and test.

Topic modeling

We first discover topics in COVID-19 related tweets using a widely used topic modeling approach, Latent Dirichlet Allocation (LDA) [20] To assess changes in topics of discussion over time, we compare timelines of topic distributions and timing of implementation of public health interventions for COVID-19.

To discover topics and track the topic change over time, we construct topic models on our Twitter data using LDA[4] implementation in the scikit-learn package [21]. We choose a model with 20 topics among 5, 10, 20, and 50 because 20 topics showed diverse and less redundant topics when manually examined.

To analyze the dynamics of public health relevant topics, we investigate the change in the prevalence of the

[a] Note that tweets with location tags are limited.
[b] https://pypi.org/project/spacy-langdetect/
[c] https://pypi.org/project/tweet-preprocessor/

topics over time. More specifically, we performed a basic analysis based on examination of the estimates of θ, a document-to-topic distribution, produced by the model. We first divide tweets into 10-day time buckets, e.g., Jan. 21 – Jan. 31, Feb. 1 – Feb. 10, and Feb. 11 – Feb. 20. Note that we use time in UTC-12 because tweet timestamps are in the time zone. Then, we compute a mean θ vector for tweets in each bucket as in [22].

Aspect based sentiment analysis

To capture sentiment revealed in tweets towards important aspects of COVID-19, we use ABSA. In our work, aspects can include public health interventions or issues associated with COVID-19 such as “social-distancing”, “reopening”, and “masks”. We investigate people’s opinion (positive/negative) towards these aspects.

We use ABSApp, a weakly-supervised ABSA system [23]. We choose ABSApp because it does not require labeled data for training, and allows manually editing domain-specific aspect and opinion lexicons produced by the method. This feature is particularly beneficial for us because in collaboration with domain experts we can select/add aspects public health agencies are interested in.

Results

Figures 1 and 2 provides some context for our results, by showing mobility and case counts for Canada and US. Since social distancing measures were enacted in the middle of March, 2020 as the daily COVID-19 cases increase, we can see that the activity of going to parks drastically increased whereas other activities such as recreational or work related mobility decreased.

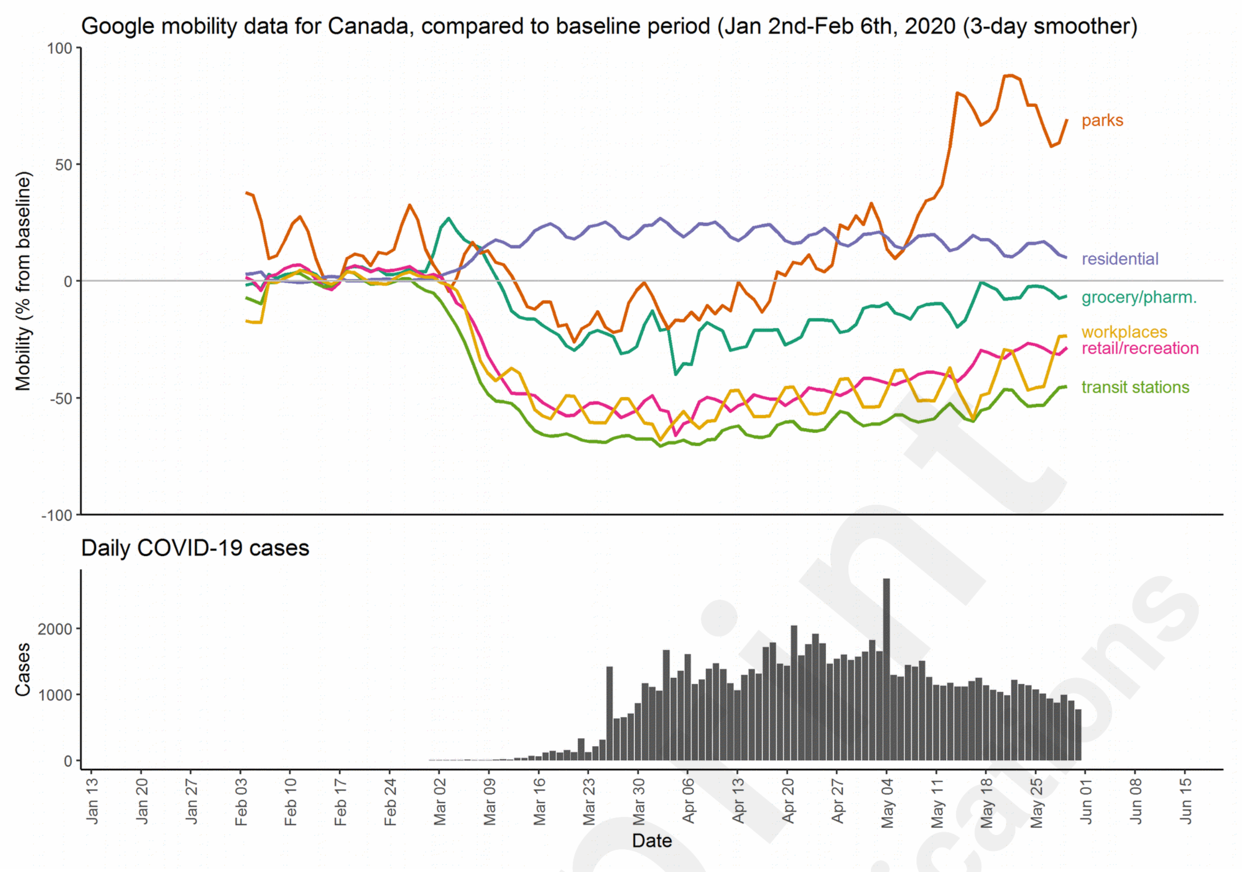


Figure 1 Mobility and case counts for Canada.

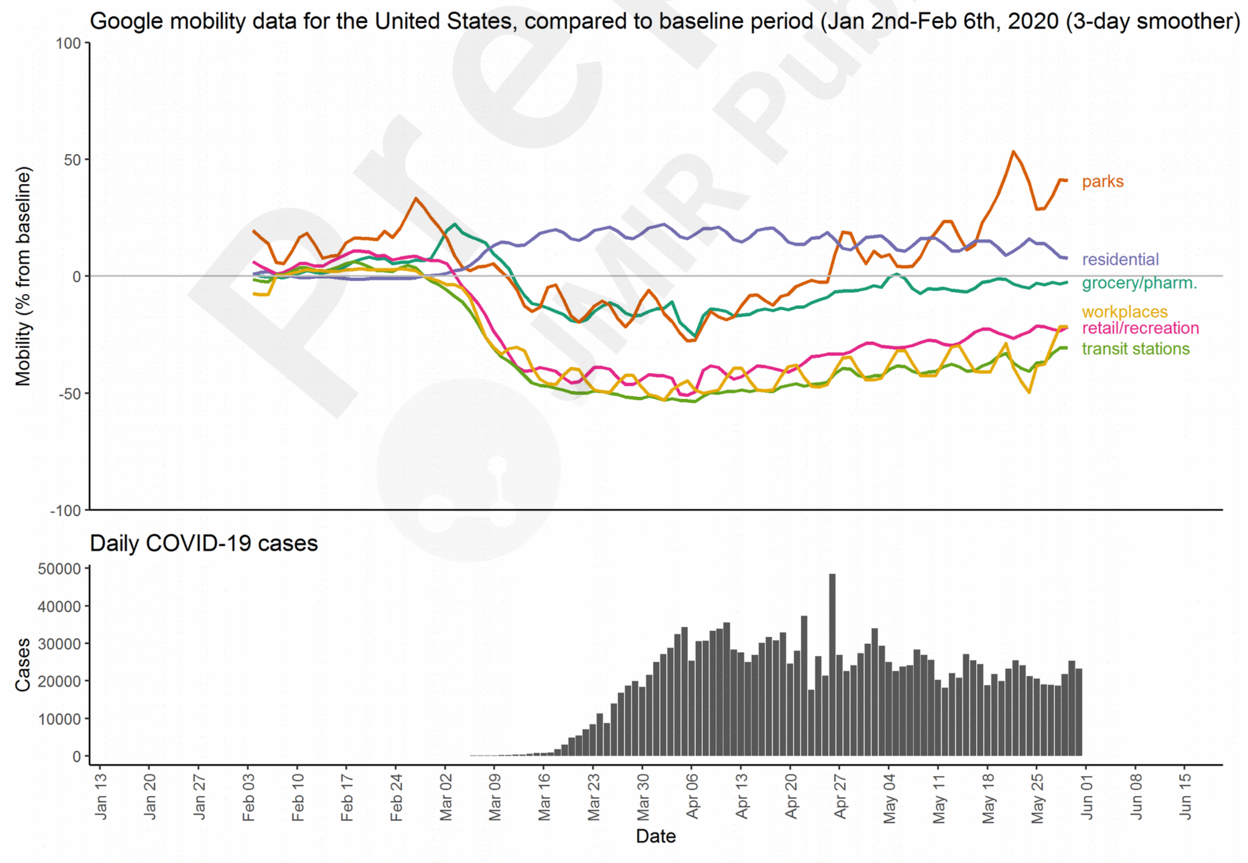


Figure 2 Mobility and case counts for the United States.

Topic modeling

Two public health experts interpreted and labelled topics, the output of LDA, based on top-ranked words and representative tweets for each topic. The discovered topics are highly related to public health promotions and interventions, such as physical distancing, border restrictions, hand washing, staying-home, and face coverings, as shown in Table 2. Other topics include US President Donald Trump, initial outbreaks in Wuhan, economic concerns and negative reactions. The entire set of topics is listed in the Appendix.

The most prevalent topics in Canada and US show some differences, as can be seen in Table 2. In both countries, age and COVID-19 transmission was the most prevalent topic. The discussion around the initial outbreak in Wuhan and US President Trump's statement was also active in both countries. However, the topic about air travel and regional border restrictions was highly ranked only in Canada whereas the topic was not even listed in the top-10 in the US. Similarly, the topics about COVID-19 being like the flu and staying home were highly ranked in the US tweets but ranked lower than other topics in the Canadian tweets.

Table 2 Top-5 prevalent topics in Canada and United States.

#	Canada	United States
1	Age and COVID-19 transmission, as well as time.	Age and COVID-19 transmission, as well as time.
2	Initial outbreak in Wuhan.	US President Trump's statement.
3	US President Trump's statement.	Early debate on whether coronavirus is like the flu.
4	Thank you notes related to the pandemic mixed with discussion of cruise ship outbreaks.	Initial outbreak in Wuhan.
5	Air travel and regional border restrictions/outbreaks.	The need to stay home and the impact of COVID-19 on essential workers and family.

(a) Canada

(b) United States

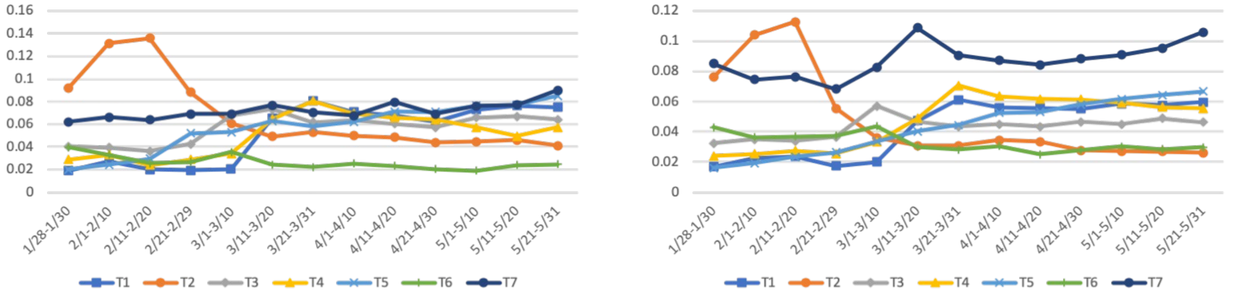


Figure 3 Topic changes over time. T1: social and physical distancing, T2: air travel and regional border restrictions/outbreaks, T3: hand washing and preventive measures, T4: the need to stay home and impact of COVID-19 on essential workers and family, T5: number of tests and cases, T6: masks and face coverings, T7: age and COVI-19 transmission as well as time.

Based on the mean θ vector for each bucket, we draw graphs of public health relevant topics over time as shown in Figure 3. First, we observe that the patterns in the US tweets and Canadian tweets are very similar. Although there are slight differences, the overall increase and decrease patterns are almost identical. For example, the topic about air travel and regional border restrictions (T2) shows a peak in February and drastically decreases.

Second, we can see that the topic trend is highly related to public health interventions. For example, the topic about social distancing (T1) starts to increase in early March after social distancing measures were enacted. Hand washing (T3) also started to be emphasized then. The topic about the need to stay home (T4) starts to increase around the end of March. In Canada, the Federal Quarantine Order was issued on March 24, and in the US, many states issued stay-at-home orders around that time as well. Discussion about the number of tests and cases (T5) gradually increases. Interestingly, the topic about masks and face coverings (T6) slightly decreases from March this is possibly because public health institutes in both countries announced their position about masks around that time.

Aspect-based sentiment analysis

After training the tweet data using ABSApp, we obtained 806 aspect terms and 211 opinion terms. Two public health experts edited the terms so that aspect terms are related to important aspects they are interested in, and that opinion terms are words that describe sentiment of the domain-specific aspect terms. Editing the lexicons resulted in 545 aspect terms (e.g., “vaccines”, “economy”, and “masks”) and 60 domain specific opinion terms (e.g., “infectious”- negative, and “professional”- positive). Then, these manually edited terms were used for inference of sentiments of 20 selected aspects. The results are shown in Figures 4 and 5. Overall, the sentiments between Canada and US show similar patterns. We observe that the sentiments about the coronavirus itself is dominantly negative. With this, the twitter users’ reactions to misinformation appear to be more negative than being positive, suggesting the frustration about the situation and misinformation. The mixed sentiments about masks might reflect the conflicting massaging around using masks. The negative sentiments towards Asians may imply the Anti-Asian sentiments escalated due to COVID-19.

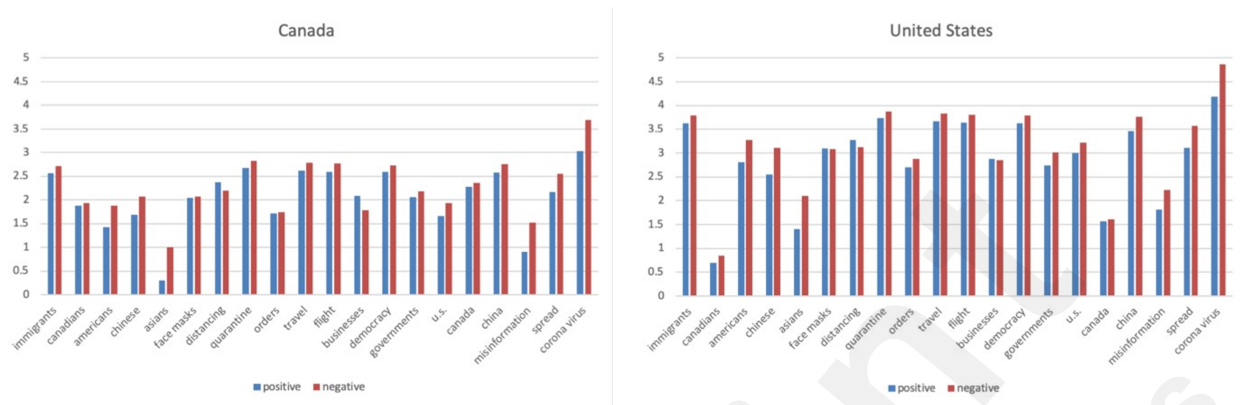


Figure 4 Aspect-based sentiment analysis results. x-axis: selected aspects, y-axis: # of positive occurrences and # of negative occurrences in log scale.

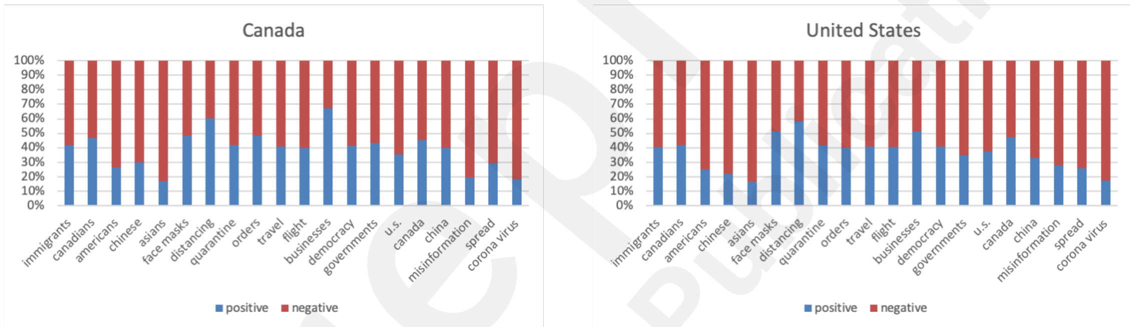


Figure 5 Aspect-based sentiment analysis results for selected aspects. y-axis: the ratio between # of positive occurrences and # of negative occurrences.

To further investigate the possible stigma for Asians, we observed words that frequently co-occur with the aspect words, Chinese and Asians. The top-ranked words in negative tweets include “virus”, “racist”, “racism”, “fucking”, “attacks”, “ass”, “assaults”, “blame”, and “hate”, and the top-ranked words in positive tweets include “fucking”, “racism”, “respectful”, “kind”, “street”, “disgusting”, and “crying”. We list sample tweets that show positive and negative sentiments in Table 3, respectively.

Table 3 Sample tweets showing positive or negative sentiments towards Asians.

Positive	Negative

- “You should not be afraid of Asians but you should be absolutely terrified of the PEOPLE THAT DONT COVER THEIR MOUTHS/NOSES DURING A COUGH AND/OR SNEEZE.” - “French Asians hit back at racism with 'I'm not a virus” - “Y’all realize that the coronavirus ain’t exclusive to Chinese people right?? mfs look for any excuse to be racist bruh”	- “Oriental Asians always starting some fuckin outbreak...” - “Yea I’m holding my breath round all Asians till this coronavirus shit clear up call it wat u think it is.” - “No Asians allowed in my shop after the outbreak.”

Figure 6 displays sentiment changes over time towards Asians and Canadians. While sentiments about Canadians are overall more positive except in February and after middle May, sentiments about Asians are mostly negative. Especially in the beginning of March when COVID-19 started to be serious in North America, we can see a spike in the number of negative tweets about Asians, and then it drastically reduces after that, which might suggest that there were some campaigns or awareness about anti-Asian racism.

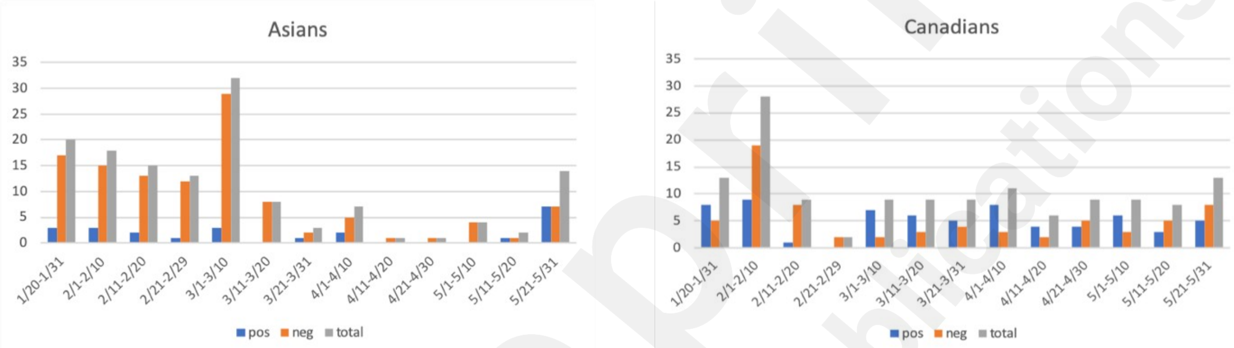


Figure 6 Sentiment changes over time for Asians and Canadians. y-scale: # of positive occurrences, # of negative occurrences, and # of total occurrences.

Discussion

In this study, using topic modeling and ABSA on twitter data from North America, we identified various topics related to physical distancing, travel and boarder restrictions, hand washing and preventive measures, face masks, stay at home orders, and number of cases and testing. Travel and border restrictions were major discussion points in February which were taken over by other topics such as physical distancing later in time. ABSA analysis identified various negative themes related to overall outbreak, anti-Asian racism and misinformation and positive occurrences related to physical distancing. These data demonstrate Twitter users’ focused on discussing and reacting to public health interventions during first phase of the pandemic. This kind of information could help public health agencies to understand public concerns as well as what public health messages are resonating in our populations who use Twitter. For example, public health agencies in North America have focused their messaging around encouraging hand hygiene, limiting physical contact when sick and staying home to prevent infection. We can see this messaging echoing in the topics around hand washing, staying home, mask-wearing and social/physical distancing.

The change over time in the age topic shows an area for improving our public messaging. Early public health data showed a strong correlation between age and severe COVID-19 outcomes [24,25]. However, we now know that SARS-CoV-2 transmission does not show the same correlation [26]. A renewed focus on messaging to highlight this difference would likely benefit our public health response.

Our findings that tweets reflect public health interventions are aligned with other studies. Abd-Alrazaq et al. [15] performed topic modeling on tweets before mid-March, and their results focused more on the virus itself, e.g., its origin, impact on people, and the economy, but did not show conversations about public health interventions. However, in the studies using tweets including March and April, topics related to social distancing policies such as school closure, stay home orders, and work from home commonly emerged in tweets posted by U.S. governors and presidential cabinet executives [10], Reddit posts [9], tweets in English, Spanish, Italian, French, and Portuguese [16], tweets in California and New York [14] and tweets in Iran [11].

Depending on tweets used for analysis, other studies report some interesting topics different from topics drawn from tweets in Canada. For example, topics related to government and political issues were observed in the studies on tweets by U.S. governors [10] and on tweets in Iran [11] whereas our analysis only showed Trump’s statement as a topic rather than overall political issues regarding COVID-19.

Our ABSA provides sentiments towards specific aspects by considering sentence structures, while most prior works performing sentiment analysis use algorithms to decide a sentiment of an entire text. For this reason, these studies are generally not suitable for identifying a sentiment of a given aspect. For instance, Wang et al. [14] computed the average sentiment scores of tweets by each day and each hour rather than obtaining sentiments for aspects. Yin et al. [27] related sentiment for each tweet to the topic the tweet belongs to, and then investigated overall sentiment of each topic. Therefore, it is not straightforward to compare our ABSA results with other sentiment analysis results.

However, our ABSA results, especially related to racism and discrimination against Asians are also observed in other research using different study methods. Zhu et al. [28] qualitatively analyzed a small number of 1,366 tweets to examine translanguaging swears around “Chinese virus”. Topic modeling on English tweets in March and April [29] showed a topic related to racism with top-ranked words such as “Chinese” and “pig”. A survey in the United States also showed prejudicial attitudes among Americans toward Chinese Americans [30]. These findings show that ABSA has the potential to track stigma and other negative consequences related to COVID-19. Our communities of Asian ethnicity have experienced unprecedented stigma and discrimination due to COVID-19. Chinese Canadians and other East Asians are experiencing hatred expressed as assaults, verbal threats, and feeling unsafe in the society. As our analysis suggests, if we monitor the change in discrimination over time using social media as a stream in real time, we could develop counter-acting messages and measures in specific geographic areas whenever there is a spike in such incidents.

The study has following limitations: We used only a small set of Twitter data because tweets with the location information were limited compared to the whole data set. This has affected other studies using social media data in similar fashion. In addition, although ABSA allows capturing more nuanced sentiments towards specific aspects, it also has the limitation the current state-of-the-art sentiment analysis techniques have: it cannot properly handle figurative languages such as sarcasm. However, this study is still useful for public health interventions and messaging because this gives insights at how public opinion has been affected by public intervention measures. Understanding people's reactions to the COVID-19 pandemic is important to public health agencies because it informs how public health agencies should frame their health messaging.

Conclusion

In this paper we presented the exploratory results of topic modeling and ABSA on COVID-19 related tweets in North America, especially focusing on Canada. We compared topic modeling and ABSA results of Canada and the US, and also showed public health intervention related topic changes over time. Our analyses demonstrated that Twitter conversations about COVID-19 are highly aligned with public health interventions. In our work, public health experts were actively involved in the computational process as well as interpretation of the results. The human-in-the-loop ABSA allowed manually editing aspect and opinion lexicons, and as a result, our analysis showed sentients towards the aspects public health experts were interested in by leveraging the domain-specific lexicons. Our results suggest that monitoring twitter user's reactions about COVID-19 related aspects can be beneficial for public health policy makers.

Conflicts of Interest

None declared.

Appendices

Table 1. LDA generated topics and their interpretations.

#	Representative words	Interpretation
T1	social, distancing, outside, park, walk, quarantinelife, stayhome	Social and physical distancing, including spending time outside during quarantine.
T2	corona, beer, stupid, die, cure, flu, drink, cold	Early debate on whether coronavirus is like the flu and around corona beer sales.
T3	china, travel, canada, russia, flights, trade, border	Air travel and regional border restrictions/outbreaks.
T4	hands, wash, health, public, use, need, safety	Hand washing and what people can do to prevent COVID-19.
T5	home, stay, safe, work, sick, family, essential	The need to stay home and the impact of COVID-19 on essential workers and family.
T6	positive, testing, tested, cases,	This topic focuses on data, particularly

	patients, hospital, data	number of tests and cases.
T7	masks, wear, face, hand, sanitizer, gloves, n95	Things we can do to prevent COVID-19, e.g., masks and face coverings.
T8	trump, china, americans, hoax, cdc, democrats, pandemic	US President Trump's statement of whether COVID-19 is a hoax and his discussion of China.
T9	students, pandemic, nyc, petition, climate, college, university	A mix of discussion around school closures, the climate and the outbreak in New York City.
T10	trump, house, white, president, press, vote, conference	USA politics including the white house press conferences and the US election.
T11	test, cdc, facts, vaccine, fake, lab, control	Lab testing for COVID-19 and vaccination discussions, as well as discussing 'fake' tests available.
T12	china, cases, death, wuhan, outbreak, spread, rate	Initial outbreak in Wuhan and its associated case and death statistics.
T13	time, old, years, day, feel, life, long	A mix of discussion around age and COVID-19 transmission, as well as time.
T14	cases, deaths, york, million, state, total, cuomo	The statistics around deaths, particularly cases and deaths in New York City.
T15	thanks, help, support, cruise, community, team, proud	Thank you notes related to the pandemic mixed with discussion of cruise ship outbreaks.
T16	health, need, care, public, emergency, fighting, stigma, curve, job	The need for health care related to addressing the COVID-19 pandemic in the USA.
T17	money, pay, paper, toilet, buying, water, price	General economic concerns including pay and bulk buying.
T18	stayhome, lockdown, order, quarantine, florida, beach, california	Quarantine and lockdown orders, particularly in Florida beaches and California.
T19	shit, fucking, ass, wow, damn, dumb, hell	Negative reactions to COVID-19 and emotional usage of swearing.
T20	break, trip, spring, school, classes, summer, quarnantinelife	COVID-19 school closures and spring break.

References

1. Yu M, Li Z, Yu Z, He J, Zhou J. Communication related health crisis on social media: a case of COVID-19 outbreak. Current Issues in Tourism Taylor & Francis; 2020 Apr 13;0(0):1–7.

2. Zhao Y, Cheng S, Yu X, Xu H. Chinese Public's Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study. Journal of medical Internet research 2020;22(5):e18825–13.

3. Carter M. How Twitter may have helped Nigeria contain Ebola. BMJ British Medical Journal Online 2014 Nov 19;:1–2.

4. Dalrymple KE, Young R, Tully M. “Facts, Not Fear”: Negotiating Uncertainty on Social Media During the 2014 Ebola Crisis. Science Communication 2016 Jul 10;38(4):442–467.

5. Guidry JPD, Jin Y, Orr CA, Messner M, Meganck S. Ebola on Instagram and Twitter_ How health organizations address the health crisis in their social media engagement. Public Relations Review Elsevier; 2017 Sep 1;43(3):477–486.

6. MPH MOE, PhD SYR. What can we learn about the Ebola outbreak from tweets? American Journal of Infection Control Elsevier Inc; 2015 Jun 1;43(6):563–571.

7. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. Journal of medical Internet research 2020;22(4):e19118–12.

8. Dong M, Cao X, Liang M, Li L, Liu G, Liang H. Understand Research Hotspots Surrounding COVID-19 and Other Coronavirus Infections Using Topic Modeling. medRxiv preprint 2020 Mar 26;:1–13.

9. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling. J Gen Intern Med. Journal of General Internal Medicine; 2020 Jun 30;35(7):2244–2247. PMID: 32399912

10. Sha H, Hasan MA, Mohler G, Brantingham PJ. Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. arXiv preprint 2020;arXiv:2004.11692.

11. Hosseini P, Hosseini P, Broniatowski DA. Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP. arXiv preprint 2020 May 18;arXiv:2005.08400.

12. Sharma K, Seo S, Meng C, Rambhatla S, Liu Y. Covid-19 on Social Media: Analyzing Misinformation in Twitter Conversations. arXiv preprint 2020 Oct;arXiv:2003.12309.

13. Odlum M, Cho H, Broadwell P, Davis N, Patrao M, Schauer D, et al. Application of Topic Modeling to Tweets as the Foundation for Health Disparity Research for COVID-19. The importance of Health Informatics in Public Health during a Pandemic 2020 Jun 10;:24–27.

14. Wang X, Zou C, Xie Z, Li D. Public Opinions towards COVID-19 in California and New York on Twitter. medRxiv Cold Spring Harbor Laboratory Press; 2020 Jul 12;:1–17.

15. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. Journal of medical Internet research 2020;22(4):e19016–9.

16. Ordun C, Purushotham S, Raff E. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. arXiv preprint 2020 May 7;arXiv:2005.03082.

17. Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. 2016. p. 1–11.

18. Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health Surveill 2020;6(2):e19273–9.

19. Honnibal M, Montani I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017.

20. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:993–1022.

21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011 Oct 11;12:2826–2830.

22. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences 2004 Mar 24;101(suppl. 1):5228–5235.

23. Pereg O, Korat D, Wasserblat M, Mamou J, Dagan I. ABSApp: A Portable Weakly-Supervised Aspect-Based Sentiment Extraction System. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 1–6.

24. Liu Y, Mao B, Liang S, Yang J-W, Lu H-W, Chai Y-H, et al. Association between ages and clinical characteristics and outcomes of coronavirus disease 2019. Eur Respir J European Respiratory Society; 2020 Apr 20;20(5):2001112–7. PMID: 32312864

25. Nikolich-Zugich J, Knox KS, Rios CT, Natt B, Bhattacharya D, Fain MJ. SARS-CoV-2 and COVID-19 in older adults: what we may expect regarding pathogenesis, immune responses, and outcomes. GeroScience GeroScience; 2020 Apr 8;42:505–514.

26. Public Health Ontario. COVID19-What We Know So Far about Infection in Children [Internet]. N Engl J Med. 2020 [cited 2020 Oct 25]. Available from: https://www.publichealthontario.ca/-/media/documents/ncov/what-we-know-children-feb-21-2020.pdf?la=en

27. Yin H, Yang S, Li J. Detecting Topic and Sentiment Dynamics due to COVID-19 Pandemic Using Social Media. arXiv preprint 2020 Jul 6;arXiv:2007.02304.

28. Zhu H. Countering COVID-19-related anti-Chinese racism with translanguaged swearing on social media. Multilingua 2020 Sep 1;39(5):607–616.

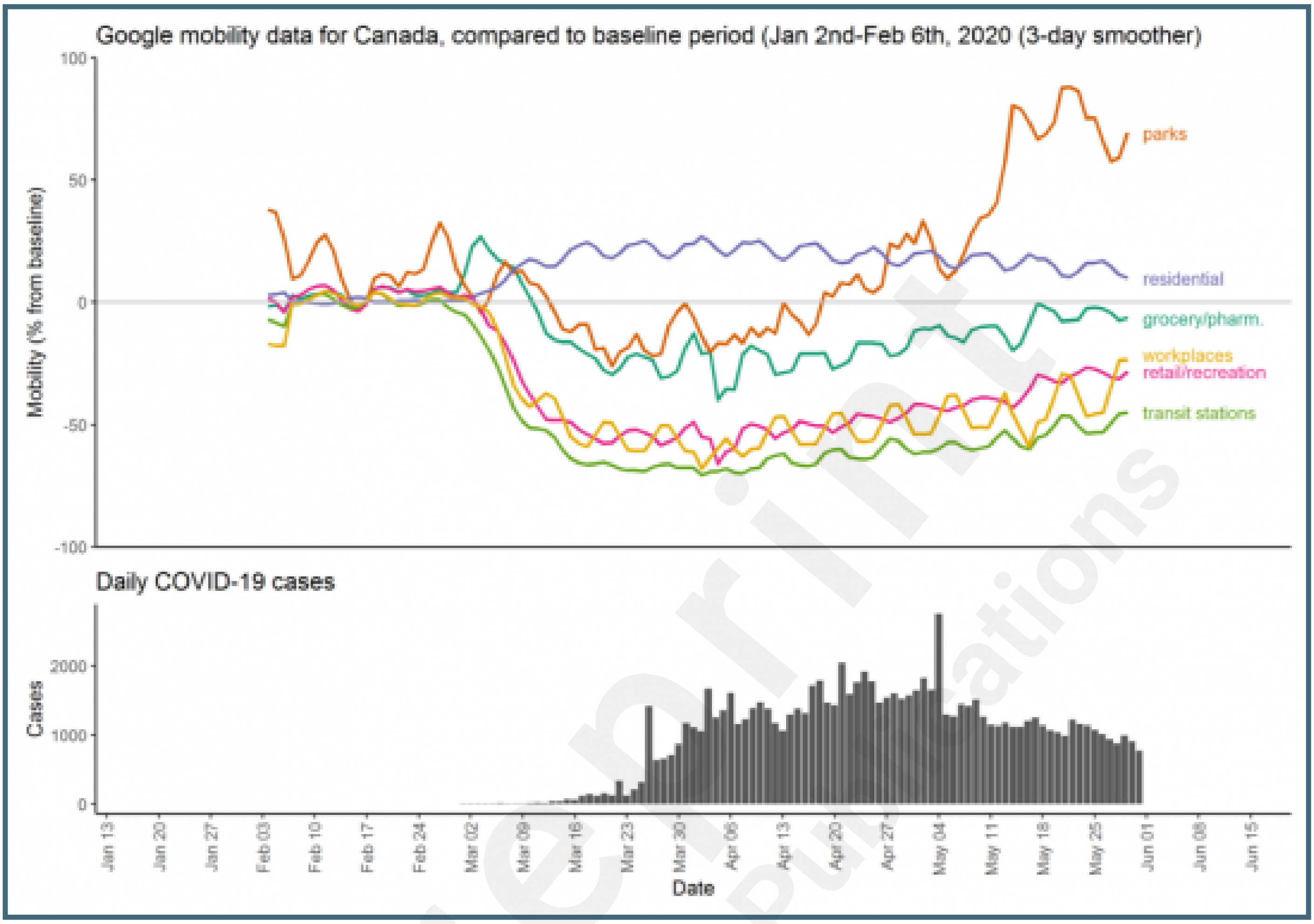
29. Chen L, Lyu H, Yang T, Wang Y, Luo J. In the Eyes of the Beholder: Analyzing Social Media Use of Neutral and Controversial Terms for COVID-19. arXiv preprint 2020 Sep 13;arXiv:2004.10225.

30. Croucher SM, Nguyen T, Rahmani D. Prejudice Toward Asian Americans in the Covid-19 Pandemic: The Effects of Social Media Use in the United States. Front. Commun. Frontiers; 2020 Jun 12;5:1–12.

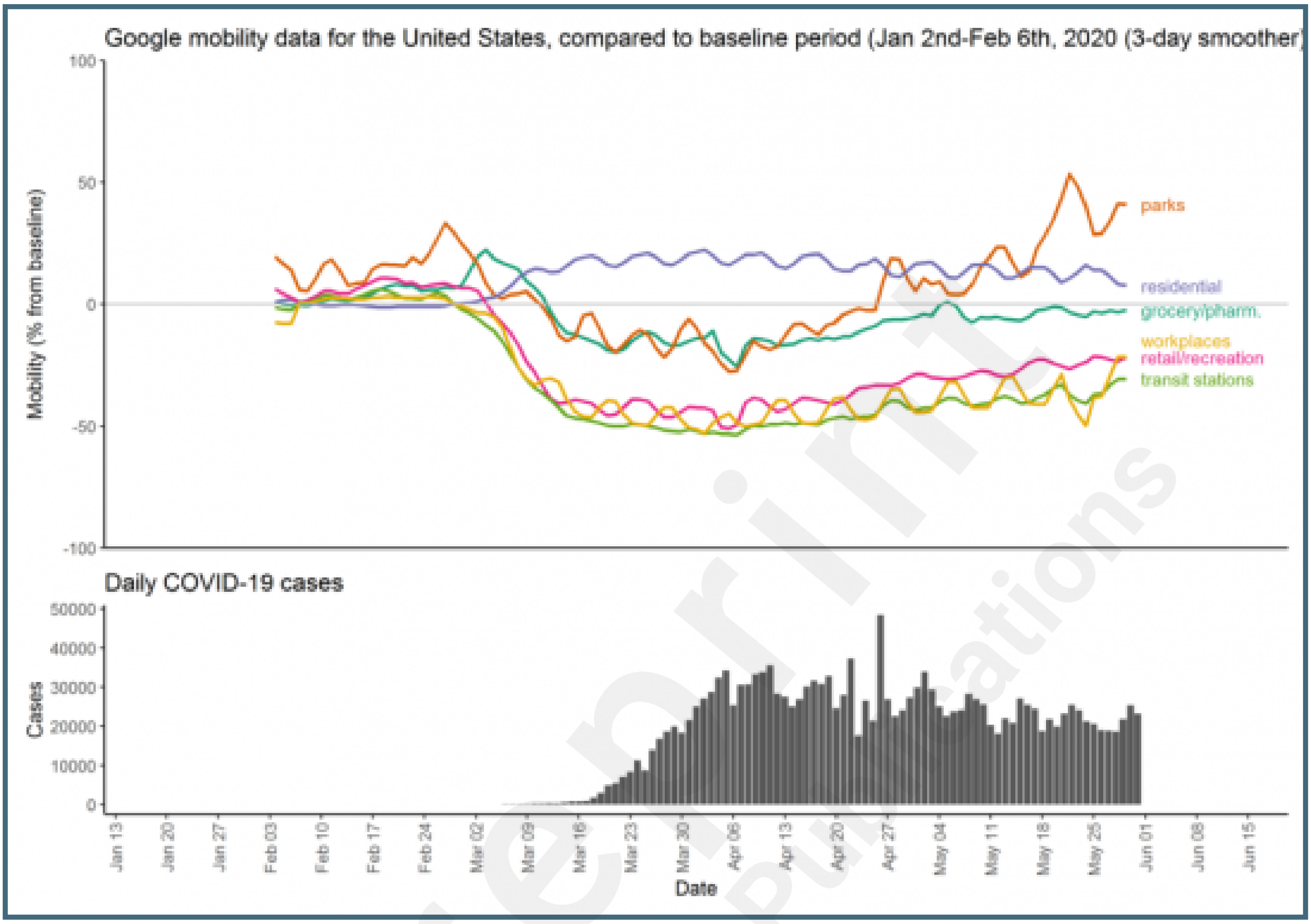
Supplementary Files

Figures

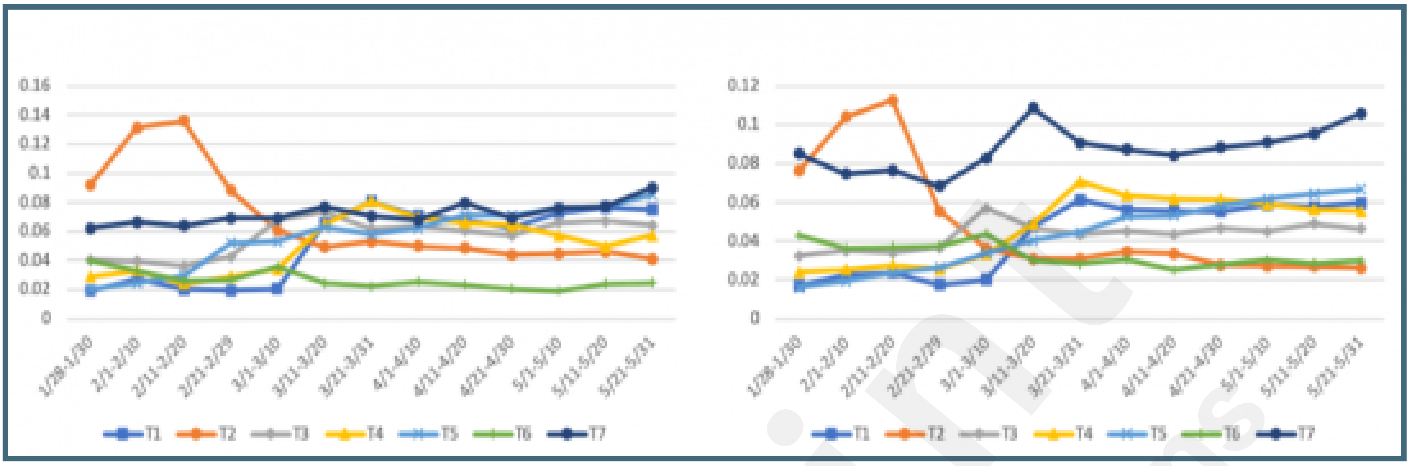
Mobility and case counts for Canada.



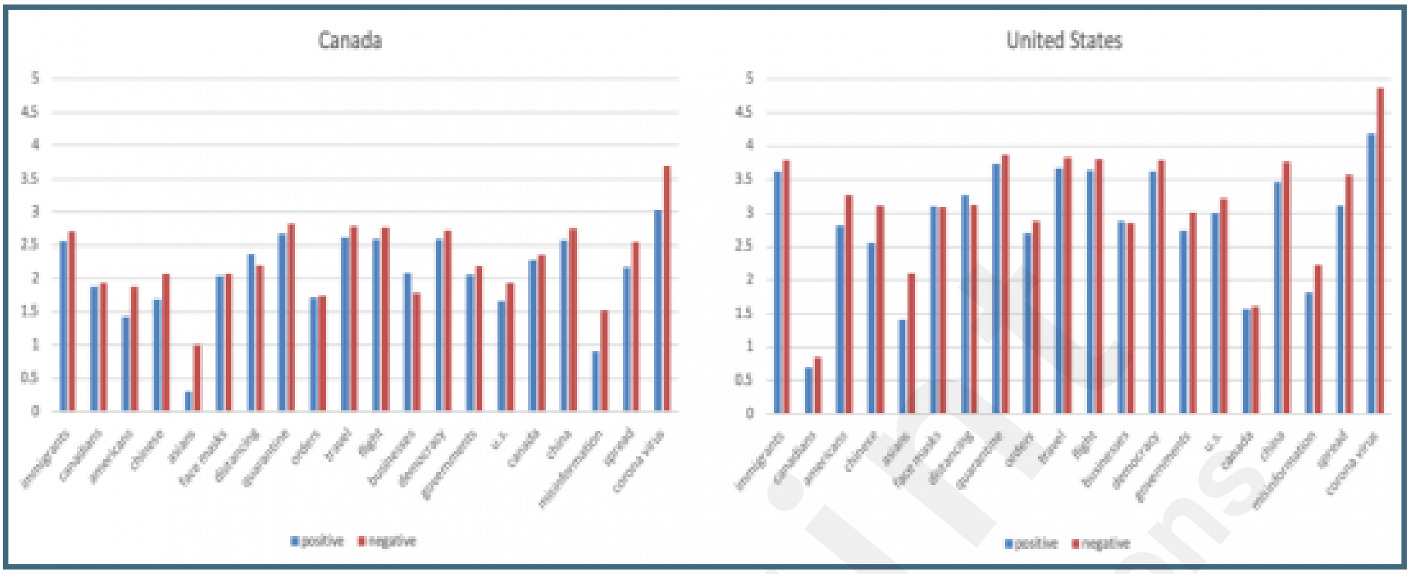
Mobility and case counts for the United States.



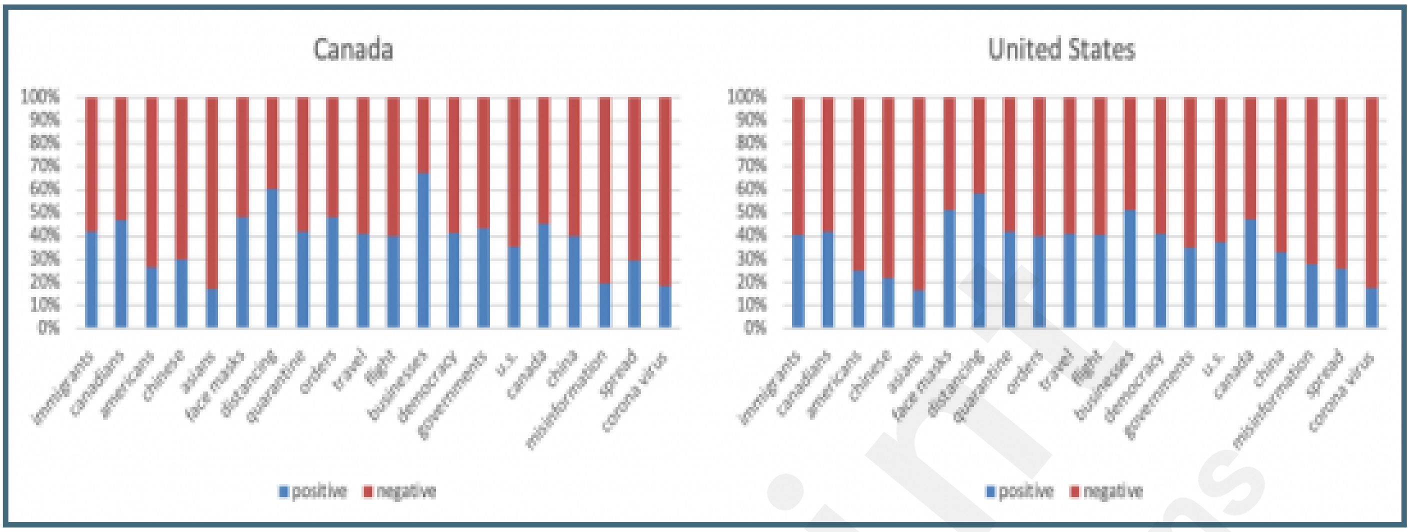
Topic changes over time. T1: social and physical distancing, T2: air travel and regional border restrictions/outbreaks, T3: hand washing and preventive measures, T4: the need to stay home and impact of COVID-19 on essential workers and family, T5: number of tests and cases, T6: masks and face coverings, T7: age and COVI-19 transmission as well as time.



Aspect-based sentiment analysis results. x-axis: selected aspects, y-axis: # of positive occurrences and # of negative occurrences in log scale.



Aspect-based sentiment analysis results for selected aspects. y-axis: the ratio between # of positive occurrences and # of negative occurrences.



Sentiment changes over time for Asians and Canadians. y-scale: # of positive occurrences, # of negative occurrences, and # of total occurrences.