Online hate in the pandemic

Authored by: Dr. Chris Tenove[1] and Dr. Heidi Tworek[2]

With the following case studies:

Hate and harassment targeting health communicators
Authored by: Dr. Tworek and Dr. Tenove with Wilson Dargbeh,[3] Hanna Hett[4] and Oliver Zhang[5]

Hate and the COVID-19 pandemic – An Analysis of B.C. Twitter discourse
Authored by: Matt Canute,[6] Hannah Holtzclaw,[7] Alberto Lusoli[8] and Wendy Hui Kyong Chun[9]

June 2022

[1] Postdoctoral research fellow, School of Public Policy and Global Affairs, University of British Columbia
[2] Canada Research Chair (Tier II) and Associate Professor, School of Public Policy and Global Affairs and Department of History, University of British Columbia
[3] Student, Master of Public Policy and Global Affairs, University of British Columbia
[4] Student, Master of Journalism, University of British Columbia
[5] Student, Master of Public Policy and Global Affairs, University of British Columbia
[6] Data Scientist, Digital Democracies Institute, Simon Fraser University
[7] PhD researcher, Digital Democracies Institute, Simon Fraser University
[8] Postdoctoral researcher, Digital Democracies Institute, Simon Fraser University
[9] Canada 150 Research Chair, Director Digital Democracies Institute, Simon Fraser University

This paper was funded by a grant from British Columbia's Office of the Human Rights Commissioner (BCOHRC), which holds the copyright. The conclusions in this paper do not necessarily reflect the views of B.C.'s Human Rights Commissioner.

Support for impacted communities: The data in this report includes disturbing language and points to trends of online abuse and hate during the pandemic in British Columbia. We recognize this information will be deeply disturbing for many people in our province to hear. This issue, while critical to examine, is extremely challenging, especially for people who have experienced or witnessed instances of online hate and toxicity. British Columbians who experience distress at reading this report or who need immediate help can access a list of crisis lines and emergency mental health supports we have compiled on our website at: bchumanrights.ca/support

Contents

1 Introduction

Since the COVID-19 pandemic reached British Columbia in January 2020, there have been reports of online hate speech as well as offline hate incidents. An inquiry into hate during the pandemic and into potential responses to hate needs to consider the different roles that online communication plays in promoting hate and exacerbating its harms. For instance, research suggests people are frequently exposed to hate speech online.[10,11] In addition, hate groups make extensive use of digital media to recruit members and organize activities.[12,13] Such activities have increased markedly during the pandemic, particularly in right-wing extremist and incel (involuntarily celibate) online forums.[14,15] To address these and other harms, the Canadian government and other governments have proposed new policies to address hate speech and related forms of harmful online communication. Technology companies have developed new approaches to curbing online hate, and civil society groups have made extensive efforts to counteract online hate and support those targeted by it.

This report aims to support the information-gathering and development of policy recommendations as part of BC's Office of the Human Rights Commissioner's inquiry into hate in the pandemic through four contributions. First, we will identify functions and forms of online hate that should be understood and addressed. Second, we will summarize key research findings on online hate in Canada, and we will suggest pandemic-related factors that may have exacerbated online hate. Third, we will summarize ongoing research projects on online abuse of health communicators (conducted by our team at the University of British Columbia) and online hate and counter-speech (conducted by our colleagues at Simon Fraser University), which are described more fully in the appended case studies. These preliminary findings from our research teams are presented to partly address the significant gaps in research on online hate in B.C. We highlight those gaps and suggest steps to address them. Fourth, we identify key actions that may be taken to address online hate, drawing on existing or proposed policies for governments, technology companies and civil society.

Alongside drawing on original research projects at UBC and SFU, this report brings together scholarship from communications and media studies, political science, criminology and history;

---

[10] Matthew Barnidge et al., Perceived Exposure to and Avoidance of Hate Speech in Various Communication Settings, *Telematics and Informatics* 44 (November 1, 2019): 101263.
[11] Canadian Race Relations Foundation, Online Hate and Racism: Canadian Experiences and Opinions on What to Do about It, (Toronto, ON: Canadian Race Relations Foundation and Abacus Data, January 25, 2021).
[12] Anti-Defamation League, 2021 Online Antisemitism Report Card, Anti-Defamation League, 2021.
[13] Cynthia Miller-Idriss, Weaponizing Online Spaces, in *Hate in the Homeland: The New Global Far Right* (Princeton, NJ: Princeton University Press, 2020), 138–60.
[14] Garth Davies, Edith Wu, and Richard Frank, A Witch's Brew of Grievances: The Potential Effects of COVID-19 on Radicalization to Violent Extremism, *Studies in Conflict & Terrorism.* Advance online publication (May 10, 2021): 1–24.
[15] Mackenzie Hart et al., An Online Environmental Scan of Right-Wing Extremism in Canada (London, UK: Institute for Strategic Dialogue, 2021).

policy reports by federal standing committees and international organizations; as well as research by civil society organizations and journalists.

We hope that this report will help individuals and organizations in B.C., including the Human Rights Commissioner, to better understand and address the complex online dimensions that form part of broader problems of hate.

2 What is online hate?

In the 1980s, the early internet provided users with new opportunities to find, engage and mobilize networks of individuals with shared experiences and interests. Many of these early online communities brought together individuals who faced discrimination and persecution for their identity. Others brought together individuals committed to white supremacy and other forms of bigotry. Since that time, there have been continuing efforts to address hostility and conflict in online spaces, including those that target women, LGBTQ2SAI+[16] folks and marginalized ethnic and religious groups. However, members of marginalized groups have sometimes found that these measures undermine their voices and activism, rather than protect them.

To clarify the complex relationships between online communication and hate, we will define key terms and propose a framework that captures different ways that digital media are used to communicate, facilitate and exacerbate hate.

2.1 Defining online hate

There is no widely agreed-upon definition of online hate.[17,18] We describe key forms that online hate can take and how these forms can contribute to hate and its harms. Identifying these different forms is necessary both to understand attempts to measure online hate, such as answering whether it increased in B.C. during the pandemic, and to clarify actions that different institutions can take to address it. Importantly, we argue that online hate includes but is not limited to online "hate speech."

Hate speech is generally understood to be public communication that aims to disparage, threaten or deeply insult people according to their identity or social group affiliations. As the Commissioner observes, Canadian criminal law and B.C.'s *Human Rights Code* agree that hate speech is communication expressed in a public way that "targets a person or group of people with a protected characteristic such as race, religion or sexual orientation" and "uses extreme language to express hatred towards that person or group of people because of their protected characteristic."[19] Similarly, Facebook defines hate speech as: "a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease."[20]

Governments and institutions seek to address hate speech because of the significant harms it may cause. These harms can occur at the individual, collective and societal levels. Individuals may be

[16] Lesbian, Gay, Bisexual, Trans, Queer, 2S (Two Spirit), Asexual/Aromantic and Intersex
[17] Lex Gill, The Legal Aspects of Hate Speech in Canada (Ottawa: Public Policy Forum, 2020).
[18] Alexandra Siegel, Online Hate Speech, in *Social Media and Democracy: The State of the Field and Prospects for Reform*, ed. Nathaniel Persily and Joshua A. Tucker, SSRC Anxieties of Democracy (Cambridge: Cambridge University Press, 2020), 56–88.
[19] BC's Office of the Human Rights Commissioner, Hate Speech Q&A, 2021.
[20] Meta, Hate Speech: Facebook Community Standards, Meta | Transparency Center, 2021.

harmed by hate speech directly from the speech itself, including psychological harm,[21] or if that speech is used to motivate violence or discriminatory actions against them. Hate speech harms groups by using terms (such as racial slurs) or claims (such as that gender minorities are immoral) that promote or justify the subordination, disparagement or violence toward the group and its members.[22] In doing so, hate speech thus draws on and reinforces systemic discrimination against the group.[23] Hate speech harms society by undermining the full political, economic and cultural participation of targeted groups and by exacerbating inter-group conflict.

Hate speech may include private or targeted communication. Part of the legal definition of hate speech is its public nature. However, we argue that this assumption is too limiting for understanding the extent or impacts of online hate. What counts as "public" online communication can be unclear. For instance, most posts on Twitter are accessible to anyone but may only be seen by a small number of people, while private messaging services like WhatsApp can be used to spread messages to hundreds, and closed groups on Facebook may be accessed by thousands. Furthermore, individuals may be directly targeted by explicit hate speech in private messages between individuals, such as through emails or messaging apps. These private expressions of hate may be sent by individuals engaged in coordinated campaigns. Private speech may also cause some of the impacts of public hate speech, which may include psychological harm and chilling or silencing effects.[24] Furthermore, social media platforms can and do regulate hate speech in some forms of direct or targeted messaging, such as WhatsApp messages, Twitter direct messages and Facebook groups. It is important to examine the impacts of such actions, such as whether they reduce the harms of targeted hate speech or potentially silence other forms of speech.

Hate incidents. As defined by the Commissioner for the inquiry, hate incidents may include either public or private hate speech.[25] The inquiry defines "hate incidents" as "actions and speech rooted in prejudice that, in the view of the person who experiences or witnesses it, are: aimed at a person or a group of people because of their actual or perceived individual, collective or intersecting characteristics including age, disability, gender expression or identity, ethnicity, Indigenous identity, place of origin, race, immigration status, religion, sex, sexual orientation and social condition; and intended to, or does, significantly dehumanize, humiliate, degrade, injure, silence and/or victimize the targeted individual or group."[26]

Hate promotion. Hate speech is also usually understood to be *explicit.* This, too, is overly limiting for the study of online hate. First, groups committed to hateful ideologies use online communication to identify, persuade, recruit or coordinate individuals. This communication can

---

[21] Katharine Gelber and Luke McNamara, Evidencing the Harms of Hate Speech, *Social Identities* 22, no. 3 (2016): 324–41.
[22] Canadian Human Rights Tribunal, Warman v. Kouba, No. T1071/5205 (November 22, 2006).
[23] Katharine Gelber, Differentiating Hate Speech: A Systemic Discrimination Approach, *Critical Review of International Social and Political Philosophy* 24, no. 4 (2021): 393-414.
[24] Danielle Keats Citron, *Hate Crimes in Cyberspace* (Cambridge, MA: Harvard University Press, 2014).
[25] BC's Office of the Human Rights Commissioner, Inquiry Details. Inquiry into Hate in the Pandemic, 2021.
[26] *Ibid*.

promote hate or exacerbate its harms, even if explicitly hateful language is not used. For instance, hate groups arrange logistical details online, and white supremacist and jihadist groups use "positive" messages about themselves and their activities to promote their superiority over others.[27] Second, members of hate groups often develop new terms or other forms of communication to promote hate in ways that only members of their in-group will understand and to evade detection by social media platforms. We use the term "hate promotion" to refer to these uses of ambiguous, coded or implicit forms of communication, particularly when used to promote and further the aims of hate-based groups.

Unequal impacts of harmful communication. There are many different forms of harmful online communication beyond hate speech, both public and private. These include forms of speech deemed illegal in Canada, such as uttering threats, harassment, defamatory libel and non-consensual distribution of intimate images, as well as countless forms of "lawful but awful" speech, including insults and health disinformation.[28] These forms of harmful communication may not explicitly refer to protected characteristics such as gender, race or religion. However, they may disproportionately target members of marginalized groups or deepen the harms experienced by those who are targeted for hate speech. For instance, individuals who have been targeted by hate speech by some actors may also face insults, false claims or the exposure of their private information. In other words, it is important to identify when harmful communication, either through intentional deployment or in its unequal patterns of targeting, may be used to achieve the functional aims of hate speech, including harming individuals and maintaining the subordination of marginalized groups.

Online hate goes beyond social media platforms. Facebook, Twitter, YouTube and other major social media platforms have been the focus of much recent public debate regarding online hate. These platforms are important given their large user bases in Canada, their key role in our media systems and evidence of their role in mobilizing hate groups and facilitating hate crimes. However, online hate and hate promotion also occur on alternative social media platforms (e.g., Gab, Parler, Rumble), gaming platforms (e.g., Roblox), encrypted communication channels (e.g., Telegram), crowdfunding and money transfer sites (e.g., Paypal, GoFundMe, Wesearchr) and dedicated websites.[29,30]

Online hate is not separate from offline hate. While this report focuses on online communication, we do not suggest that what happens online can be separated from what happens offline. Online communication is often used to organize or publicize offline hate activities. For instance, investigation of the 2019 attack on a mosque in Christchurch, New Zealand, revealed that the killer made hateful remarks online before the attack, received encouragements online to

---

[27] Soraya Binetti et al., Stop the Virus of Disinformation: The Malicious Use of Social Media by Terrorist, Violent Extremist and Criminal Groups during the COVID-19 Pandemic (Torino: United Nations Interregional Crime and Justice Research Institute, November 2020).
[28] Canadian Commission on Democratic Expression, Harms Reduction: A Six-Step Program to Protect Democratic Expression Online (Ottawa: Public Policy Forum, January 2021), 32.
[29] Anti-Defamation League, "2021 Online Antisemitism Report Card."
[30] Miller-Idriss, "Weaponizing Online Spaces."

commit hate-motivated violence, publicized the violence online and had his actions used by broader online communities to promote further hate and violence.[31] It is thus more accurate to refer to online hate as the use of digital media technologies to promote hate and deepen its harms through, for example, hate speech, encouragement to violent actions and amplification of hateful messages. To capture these nuances, some experts refer to "technology-facilitated" violence and discrimination.[32]

Furthermore, the impact of hate speech and other forms of online abuse on women and racialized individuals is exacerbated by their offline experiences of threat, hostility and discrimination, as we detailed in research on candidates in the 2019 federal election.[33]

2.2 Implications for this inquiry understanding and responding to online hate in B.C.

1. Hate incidents are not limited to *public* communication, as in Canadian criminal and human rights law on hate speech. Hate incidents can also include private or targeted communication. Attention should also be given to hate promotion, which may not itself be explicit hate speech but may facilitate the commission of hate incidents. Finally, the policies and activities of social media platforms should be examined with respect to the actions they take regarding public hate speech, private hate speech and hate promotion. We therefore include recommendations for social media companies in Section 5 of this paper.

2. Online hate speech has characteristics that may distinguish it from other forms of harmful online communication, but that broader category should be taken into account when investigating the harms of online hate and developing responses. For instance, many forms of harmful online communication may be used to endanger or silence members of marginalized groups. Social media companies' policies also influence the extent and disproportionate impact of these diverse forms of harmful speech and could be improved.

3. Online hate occurs across a wide range of online spaces and services, and efforts to address online hate need to go beyond a narrow focus on major social media platforms.

4. Online hate must be understood and addressed by taking into account "offline" hate, since online communication can encourage, facilitate and publicize offline hate and exacerbate its harms. More generally, digital media not only facilitate hate but archive and mirror societies' hateful and discriminatory beliefs and actions.

---

[31] Royal Commission of Inquiry, Ko Tō Tātou Kāinga Tēnei Report: Royal Commission of Inquiry into the Terrorist Attack on Christchurch Masjidain on 15 March 2019 (Wellington, New Zealand, 2020).
[32] Suzie Dunn, Technology-Facilitated Gender-Based Violence: An Overview, Supporting a Safer Internet (Waterloo: Centre for International Governance Innovation, 2020).
[33] Chris Tenove and Heidi Tworek, Trolled on the Campaign Trail: Online Incivility and Abuse in Canadian Politics (Vancouver: Centre for the Study of Democratic Institutions, University of British Columbia, October 2020).

5. Online hate disproportionately targets and affects marginalized groups, even though communication that denigrates people on the basis of protected characteristics can also target identity groups that are not marginalized.[34] Hate speech is often employed to maintain the social position of dominant groups[35] and should be addressed as part of broader efforts to address systemic exclusions and injustices.

[34] For instance, white male journalists and politicians on occasion face disparagement and even wishes of violence that invoke their identity (unpublished research by Chris Tenove).
[35] Gelber, "Differentiating Hate Speech."

3 Online hate in the pandemic context

Online hate was identified as a serious problem long before the pandemic. In Canada, civil society groups have raised alarm bells regarding the frequency and impact of online hate,[36,37,38] and Parliament released a detailed report in 2019 calling for action on the issue.[39] Evidence suggests that experiences of online hate have increased significantly in Canada during the pandemic. It is less clear whether hate promotion, or activities by groups driven by hateful ideologies, have also increased. Major gaps in evidence exist.

We first outline the evidence regarding online hate in Canada during the pandemic, focusing on its frequency, forms, targeting and impact. We identify key gaps in knowledge, including the fact that very little research is focused on British Columbia. We then identify key pandemic-related factors that may exacerbate online hate and its harms, primarily drawing on research conducted outside of Canada.

3.1 Online hate in Canada before and during the pandemic

Police-reported hate crimes have steadily increased over the last five years and rose dramatically in 2020. Across Canada in 2020, there were 2,669 hate crimes reported to police, the highest number reported in a year, and a 37% increase compared to 2019.[40] This rise was primarily driven by increasing rates of reported hate crimes targeting people on the basis of race and ethnicity, especially Black, East or Southeast Asian, South Asian and Indigenous populations. There were 196 more of these reported incidents targeting race or ethnicity in B.C. in 2020 compared to 2019. Statistics Canada has not yet revealed how many of these reported crimes had an online component, but 6.2% of reported hate crimes in 2019 were also recorded by police as cybercrimes.[41]

Police-reported hate crimes represent a tiny fraction of all hate incidents, including online hate. Statistics Canada has estimated that two-thirds of victims of hate-motivated crimes would not report them to police.[42] Furthermore, many forms of online hate may not rise to the legal category of hate crimes.

---

[36] Canadian Race Relations Foundation, "Online Hate and Racism: Canadian Experiences and Opinions on What to Do about It."
[37] Canadian Coalition to Combat Online Hate, Open Letter from the Canadian Coalition to Combat Online Hate (Center for Israel and Jewish Affairs, November 16, 2021).
[38] National Council of Canadian Muslims, NCCM Recommendations: National Summit on Islamophobia (Ottawa: National Council of Canadian Muslims, July 2021).
[39] Standing Committee on Justice and Human Rights, Taking Action to End Online Hate (Ottawa: House of Commons, Canada, June 2019).
[40] Greg Moreau, Police-Reported Crime Statistics in Canada, 2020 (Ottawa: Statistics Canada, Government of Canada, July 27, 2021).
[41] Greg Moreau, Police-Reported Hate Crime in Canada, 2019 (Ottawa: Statistics Canada, Government of Canada, March 29, 2021).
[42] Standing Committee on Justice and Human Rights, "Taking Action to End Online Hate," 19.

Surveys provide a different picture of experiences of online hate in Canada. The Canadian Race Relations Foundation commissioned a survey of over 2,000 adult Canadian residents in January 2021[43] that asked whether they had *experienced* or *seen* online content that may amount to hate speech. They found:

- 7% of Canadians experienced racist comments or content, and rates were three times higher for racialized respondents (14%) compared to white respondents (5%); a further 40% had seen racist content;
- 9% experienced sexist content, and a further 34% had seen it;
- 6% experienced content inciting violence, and a further 36% had seen it;
- 6% experienced homophobic content, and a further 32% had seen it.

The survey further found that 78% of respondents were concerned about the spread of hate speech online, and the vast majority wanted the government to take more aggressive action to address it, including greater regulation of social media companies and greater efforts to hold perpetrators to account for what they say, share and do online.

These findings broadly align with a non-random survey conducted by the Mosaic Institute in Ontario, which primarily drew on information from racialized adults under 40. It found that 38% of Black, Indigenous, Jewish and Muslim respondents felt unsafe due to something they had experienced online, and 17% of respondents had experienced online hate regarding COVID-19.[44]

With respect to hate promotion, research by the Institute for Strategic Dialogue (ISD) found that Canadian right-wing extremist activity seemed to increase in many online forums during the first year of the pandemic.[45] They found that extremist voices have capitalized on people's increased time online, the anxiety and loss of control many have felt, and resentment of government response to the pandemic. Conspiracy theories were frequently advanced, including regarding China's potential role in the pandemic and influence in Canada, as well as antisemitic and anti-Muslim narratives. The ISD researchers also tracked white supremacist channels on Twitter promoting violence, as well discussions of murdering and harming women by members of incel forums.

Some of the best-known individuals in COVID conspiracy movements have a long history of antisemitism, including questioning the number of Jewish victims in the Holocaust. The Canadian Anti-Hate Network observed that some people protesting pandemic-related health measures were previously involved in far-right movements, such as yellow vest demonstrators or anti-Muslim groups.[46] Despite these linkages, it is important to remember that most people who

---

[43] Canadian Race Relations Foundation, "Online Hate and Racism: Canadian Experiences and Opinions on What to Do about It."
[44] Mosaic Institute, Through Our Eyes: Understanding the Impact of Online Hate on Ontario Communities." (Toronto: Mosaic Institute, 2021).
[45] Hart et al., "An Online Environmental Scan of Right-Wing Extremism in Canada."
[46] Rachel Bergen, Antisemitic Rhetoric Continues to Be Used by Some Opponents of COVID-19 Measures, CBC News, October 10, 2021.

question public policies responding to COVID-19, including vaccines and lockdowns, do not support hate groups.

Some research on the first wave of COVID-19 suggests that the pandemic exacerbated existing trends. Searches for violent, far-right extremist material rose an average of 18.5 percent in Canadian cities during spring 2020.[47] Canadians seem to have been a key driver of global anti-Chinese and anti-Asian rhetoric: one study ranked Canadians as the fourth largest cohort in an examination of this content on Twitter.[48]

Finally, to understand Canada, it is crucial to understand trends in the United States. This is because Canadians engage more with non-Canadian, and especially U.S.-based, accounts on social media than with Canadian accounts. A 2021 study by McGill's Media Ecosystem Observatory looked at nearly 200,000 Canadian Twitter accounts and found that only about 18% of the accounts that Canadians follow are Canadian. By contrast, "an astonishing 57% are based in the United States, with the rest of the world accounting for only a quarter of follows."[49] Canadians also retweet U.S.-based accounts ten times as often as Canadian accounts.

The evidence, summarized here, suggests that there seems to have been a significant increase in online hate during the pandemic and that it is particularly likely to have targeted racial or ethnic groups. There is also reason to believe that this abuse is intersectional, meaning that it targets people for their gender as well as racial identities. However, the evidence base to date is very limited. There is good reason to believe that most experiences of online hate have not been reported either to social media platforms or police. There is little research on whether and how the pandemic itself might be related to shifts in online hate. We anticipate more findings on these topics in the months and years to come. In the meantime, we turn to research primarily beyond Canada to shed light on those questions.

3.2 How might the pandemic influence online hate?

Pandemics have historically exacerbated societal tensions and often led to scapegoating. While this has played out differently during different pandemics, many of the narratives and attacks echo prejudices of the past. A literature review by Jonathan Corpus Ong concluded that alongside COVID-19, the United States was experiencing a "secondary contagion of racism."[50] The same may apply to Canada, though not in quite the same ways as the United States. There are many overlapping and intertwined reasons why pandemics can accelerate and amplify online hate. Below, we briefly identify and describe the main reasons explored by researchers up to now.

---

[47] Moonshot, The Impact of COVID-19 on Canadian Search Traffic (London: Moonshot, June 8, 2020).
[48] Moonshot, COVID-19: Conspiracy Theories, Hate Speech and Incitements to Violence on Twitter (London: Moonshot, April 29, 2020), 3.
[49] Taylor Owen et al., Understanding Vaccine Hesitancy in Canada: Attitudes, Beliefs, and the Information Ecosystem (Media Ecosystem Observatory, McGill University and University of Toronto, December 2020), 17.
[50] Jonathan Corpus Ong, The Contagion of Stigmatization: Racism and Discrimination in the 'Infodemic' Moment, V1.0, MediaWell, Social Science Research Council, February 4, 2021.

Scapegoating in pandemics has commonly happened for centuries. This often focuses on othering particular groups who can then be blamed for the pandemic. This generally affects minority groups and migrants more than other groups.[51]

- The anti-Asian racism exhibited during the COVID-19 pandemic has longer historical roots. European imperial powers in the 19th century often attributed disease origins to Asia, even though the attribution was often erroneous.[52] Canada too has a long history of racial exclusion during disease outbreaks. For example, when a Chinese laundry worker seemed to be the first person to contract smallpox during an outbreak in Calgary in 1892, municipal authorities quarantined him, separated out the Chinese community, and the laundry where the man worked was burned down.[53] Focusing on the Chinese origins of COVID-19 drew on long-standing problematic stereotypes about Asia as the "origin" of diseases. This intertwined with and promoted anti-Asian racism online and offline in British Columbia and elsewhere.

- More broadly, pandemics and epidemics have often exacerbated "othering." Political scientists have noted a "long history of associating immigrants and disease in America and the problematic impact that association has on attitudes toward immigrants."[54] During the Ebola outbreak in West Africa in 2014, for example, mainstream media used images and concepts that portrayed the entire continent of Africa as "a dirty, diseased place to be feared."[55] This sentiment was amplified and spread online, including by future U.S. president Donald J. Trump, who advocated for travel bans and falsely claimed that Ebola was spreading all over Africa.[56] Researchers have also found that pandemics often bolster anti-foreigner sentiment, such as during the avian flu outbreak of 2009-2010.[57] Anti-immigrant and anti-foreigner sentiments can, and often do, intertwine with hatred against people from those ethnic backgrounds. This draws on older stereotypes about certain groups of people as "perpetual foreigners."

- Other groups, such as Jews, are also often scapegoated during pandemics, and older stereotypes have emerged in new forms online. A research report by the UK's

---

[51] Amanuel Elias et al., Racism and Nationalism during and beyond the COVID-19 Pandemic, *Ethnic and Racial Studies* 44, no. 5 (2021): 783–93.
[52] Mark Harrison, Contagion: How Commerce Has Spread Disease (New Haven, CN: Yale University Press, 2012), chap. 6.
[53] Kristin Burnett, Race, Disease, and Public Violence: Smallpox and the (Un)Making of Calgary's Chinatown, 1892, *Social History of Medicine* 25, no. 2 (2012): 362–79.
[54] Kim Yi Dionne and Laura Seay, 8. American Perceptions of Africa during an Ebola Outbreak, in *Ebola's Message* (Cambridge, MA: MIT Press, 2020).
[55] *Ibid*.
[56] Aaron Rupar, Trump Is Facing a Coronavirus Threat. Let's Look Back at How He Talked about Ebola, *Vox*, February 26, 2020.
[57] Franciska Krings et al., Preventing Contagion With Avian Influenza: Disease Salience, Attitudes Toward Foreigners, and Avoidance Beliefs, *Journal of Applied Social Psychology* 42, no. 6 (2012): 1451–66.

Community Security Trust[58] identified five main online narratives of antisemitism during the pandemic, most of which draw on centuries-old prejudices:

1. The virus as real but a Jewish conspiracy;
2. The virus as fake but a Jewish conspiracy;
3. Portraying Jews as primary spreaders of the virus (calling it "Jew flu," much as others may have called it "Wuhan flu");
4. Cheering for Jewish deaths;
5. Trying to kill Jews with COVID-19 (the "Holocough").

A sixth category emerged increasingly from spring 2021 onwards: the comparison of public health orders and vaccine passports to the Holocaust.[59,60] Such comparisons frequently appeared at demonstrations against vaccine passports and mandates, including some protestors wearing yellow stars to evoke a deeply historically inaccurate parallel with the Jews forced to wear yellow stars in Nazi-occupied Europe.

Many of these stereotypes and sentiments are conveyed not just in text but often in memes (images or videos with text intended to be humorous). Such memes can also employ discriminatory stereotypes such as portraying Jewish people with long, hooked noses.

- Much of this racism is also spurred by broader problematic practices, such as place-based naming for variants.[61] While the World Health Organization (WHO) replaced the naming scheme with Greek letters for variants in spring 2021, there continue to be issues with place-based blaming, e.g., Omicron was *identified* first by scientists in South Africa. Such issues also manifested in institutional practices like selective and discriminatory travel bans for southern African countries. Such sentiments and actions could potentially amplify online hate.

3.3 Relationship between hate and misinformation/conspiracy theories.

- One prominent theme for theories during the pandemic has focused on "sinister origins" or shadowy forces behind COVID-19 and vaccines.[62] These go beyond concerns about profit-seeking pharmaceutical companies to allege geo-political

[58] Community Security Trust, Coronavirus and the Plague of Antisemitism (London, UK: Community Security Trust, 2020).
[59] Alistair Steele, Disgust Growing over Vaccine Protesters' Holocaust Comparisons, CBC News, September 15, 2021.
[60] Bill Kaufmann, Former CPS Hate Crimes Officer Who Compared Vaccine Mandates to Holocaust Is Suspended, *Calgary Herald*, December 3, 2021.
[61] Heidi Tworek, Why Disease Names Matter, *The Globe and Mail*, March 24, 2021.
[62] Brian Hughes et al., Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation, *International Journal of Environmental Research and Public Health* 18, no. 14 (2021): 1–18.

machinations by China or "globalist" actors such as Bill Gates. Some of these conspiracy theories are explicitly or implicitly racist.[63]

- Many leaders, including the Director-General of the World Health Organization, have decried the huge spread of conspiracy theories and misinformation during the COVID-19 pandemic as an "infodemic," though some scholars suggest the metaphor is misleading.[64] While researchers are still trying to understand the exact contours of conspiracy theories online, it is clear that those conspiracy theories serve to amplify online hate against particular groups.

More online. Public health responses to the pandemic forced many people to take more of their lives online, including to social media spaces.

- According to a Canadian Internet Use Survey, 75% of those older than 15 conducted "various Internet-related activities" more often since the start of the pandemic.[65] Such a dramatic increase in internet usage offers more opportunities for people to engage and find their communities online, including communities of hate.

Politicization of the pandemic and public health responses may prove fertile ground for radicalization or cross-movement networking.

- Research by the Institute for Strategic Dialogue found that right-wing extremist and incel forums were frequently dominated by conversations around the pandemic and often resistance to public health policies.[66] A Canadian Anti-Hate Network investigation found that anti-mask and anti-lockdown protests brought together diverse extreme and fringe movements, potentially leading to intermingling and more effective mobilization.[67] Similar overlapping groups were seen in some of the vitriolic protests against Justin Trudeau during the 2021 federal election campaign[68] and at the "Freedom Convoy" purportedly combatting a cross-border mandate for truckers in early 2022.[69]

- U.S. politics in 2020 was particularly oriented toward racial conflict. There was also massive mobilization against racism, particularly with Black Lives Matter protests in

---

[63] Kate Starbird, Emma S Spiro, and Kolina Koltai, Misinformation, Crisis, and Public Health—Reviewing the Literature, VV1.0, MediaWell, Social Science Research Council, June 25, 2020.
[64] Felix M Simon and Chico Q Camargo, Autopsy of a Metaphor: The Origins, Use and Blind Spots of the 'Infodemic,' *New Media & Society*, Advance online publication (July 20, 2021).
[65] Howard Bilodeau, Abby Kehler, and Nicole Minnema. 2021. Internet Use and COVID-19: How the Pandemic Increased the Amount of Time Canadians Spend Online (Ottawa: Statistics Canada, Government of Canada, June 24, 2021).
[66] Hart et al., "An Online Environmental Scan of Right-Wing Extremism in Canada."
[67] Canadian Anti-Hate Coalition, Despite Social Media Bans, QAnon Is Reaching Across Canada's Extreme and Fringe Movements, Canadian Anti-Hate Network, October 26, 2020.
[68] Alex Boutilier and Grant LaFleche, Who's behind the Justin Trudeau Protests? This Newmarket Mom Is among the Influencers Urging People to Show Up, *The Toronto Star*, August 30, 2021.
[69] Rachel Gilmore, Some Trucker Convoy Organizers Have History of White Nationalism, Racism, Global News, January 29, 2022.

summer 2020 after the murder of George Floyd. Such demonstrations occurred worldwide. In Canada, they also involved raising awareness of and protesting against police violence against Indigenous people. The pandemic thus occurred in the context of increasing anti-racism mobilizations (and backlash to them).

- In some cases, racist, misogynist or antisemitic narratives were deployed as part of protests against measures to prevent and slow the spread of COVID-19. In Ohio, for example, one state representative used antisemitic slurs in a Facebook post to push back against orders from Dr. Amy Acton, the Jewish director of Ohio's health department. Acton's communication skills were widely praised, with one *New York Times* video in May 2020 calling her "the leader we wish we all had."[70] Yet by summer 2020, Acton resigned after experiencing significant online and offline antisemitic abuse, including protests outside her private home.[71]

- Research into mobilizations against lockdowns found four main narratives in Canada: "misinformation reported by alternative media outlets; the adoption of anti-lockdown positions by politicians; extremist groups latching on to COVID-19 conspiracy theories to attract new members; and opposition to public health measures by the religious far-right."[72] The researchers conclude that Canada's far-right movements are not as popular as equivalents in the U.S. and U.K., and "in Canada tend to be consumed with interpersonal and intergroup dynamics that prevent a larger, more cohesive movement from emerging. Nevertheless, there are reasons to be concerned."

[70] Sanya Dosani and Adam Westbrook, Opinion | The Leader We Wish We All Had, *The New York Times*, May 5, 2020.
[71] Meredith Deliso, 'Unsafe': Women in Public Health Facing Pushback and Threats for Coronavirus Response, ABC News, July 3, 2020.
[72] Stephanie Carvin, Kurt Phillips, and Amarnath Amarasingam, The Real Virus Is Fear: Anti-Lockdown Mobilization in Canada, *Policy Options*, January 21, 2022.

4 Case studies: Health communication and B.C. Twitter

To date, there is little specific research into online hate in B.C. and how the pandemic may have exacerbated online hate specifically in the context of British Columbia. We thus produced two case studies to provide BCOHRC with preliminary research on the interaction between the pandemic and online hate.

The two case studies take two different methodological approaches to understand online hate during the pandemic and provide two different windows into the phenomenon. The first case study (see Case Study A) uses more qualitative methods to explore how online abuse affects people and how it might intersect with their gender, race and/or religion. It draws from the report authors' research project into the online abuse of health communicators in Canada during the pandemic. We conducted interviews with a range of people who communicate around the pandemic, including healthcare workers, university-based experts, public health officials and journalists. We made particular effort to interview people with a range of identities to understand how online abuse might affect them. For this case study, we also completed a news scan about online abuse of health communicators and conducted a thematic analysis of some abusive tweets. We argue that the experiences of health communicators can illuminate broader dynamics and impacts of online hate during the pandemic.

The second case study (see Case Study B) is a quantitative analysis of anti-Asian tweets on Twitter in British Columbia during the pandemic. The study was conducted by the Digital Democracies Institute at Simon Fraser University as part of a wider project on online hate. The case study indicates that anti-Asian hate speech has increased during the pandemic, while explaining the algorithmic tools used for big data investigations and their limitations. The case study also explores the concept of "counterspeech," speech that pushes back against hate, and shows how counterspeech can often get misidentified as hate speech by algorithms.

The two different approaches in the case studies provide some initial evidence into online hate during the pandemic in B.C. They also highlight some of the challenges of investigating these areas. We anticipate that future studies will provide more concrete evidence on the dimensions of online hate during the pandemic in B.C and Canada. Such studies might look at content in languages other than English and on platforms beyond Twitter. Although we do not yet have the full picture of how the pandemic has exacerbated online hate, the case studies and broader research indicate that online hate has increased and worsened during the pandemic. This makes it important to consider possible solutions and responses, even as scholars and civil society groups continue to work on understanding the problem.

5 Responses to online hate

The problem of online hate is multi-dimensional. While it occurs online, it cannot be addressed solely by social media platforms or other technology companies, although they have a significant role to play. Even if there is no silver bullet, we suggest some key policy responses. We also incorporate some lessons from efforts in other jurisdictions.

5.1 Improve platform action on online hate, including through new regulatory frameworks

- Part of the difficulty in assessing online hate is the comparative lack of transparency from online platforms. As the Canadian Commission on Democratic Expression observed: "One of the central challenges faced by researchers, journalists, policy communities, social media users and, soon, regulators is that the platform ecosystem is a black box."[73] The European Union's (EU) voluntary Code of Conduct on Countering Illegal Hate Speech Online requires signatory platform companies to submit to yearly monitoring. However, the code and its monitoring methodology have significant limitations, as monitoring is periodic and there is no independent auditing to ensure its accuracy.[74] Further voluntary efforts at transparency by platforms have come under fire for arbitrarily withdrawing access from researchers and for not providing full sets of data. In response, a draft bill in the United States and the draft *Digital Services Act* in the EU both have provisions around transparency for researchers. Transparency requirements, if well designed,[75] could enable researchers to contribute to evidence-based policy-making and provide important information to regulators and civil society groups seeking to address issues such as online hate.[76]

- Canada, like many other jurisdictions, seeks to move beyond voluntary agreements with platforms and pursue new regulatory frameworks requiring them to address hate speech promptly. Germany has been an early mover in policymaking in this area with its 2018 Network Enforcement Law (NetzDG). The law requires platforms with more than two million unique users in Germany to create a simple complaint mechanism for users to flag posts that seem to violate one of 22 statutes of German speech law. Platforms have to address the complaint within 24 hours or face fines of up to 50 million Euros per post.[77] While the companies have removed hundreds of thousands of

[73] Canadian Commission on Democratic Expression, "Harms Reduction," 34.
[74] Barbora Bukovská, The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An Analysis of Freedom of Expression Implications (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, May 7, 2019).
[75] Heidi Tworek and Alicia Wanless, Time for Transparency From Digital Platforms, But What Does That Really Mean?, *Lawfare,* January 20, 2022.
[76] For further suggestions on transparency, see Mark MacCarthy, Transparency Recommendations for Regulatory Regimes of Digital Platforms (Waterloo: Centre for International Governance Innovation, March, 2022).
[77] Heidi Tworek and Paddy Leerssen, An Analysis of Germany's NetzDG Law (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, April 15, 2019).

pieces of content, Germany continues to have a "white supremacist problem" because officials still underplay the issue and have not undertaken broader structural reforms to address racism.[78] An updated version of NetzDG would have required platforms to provide the Federal Criminal Police Office with the illegal content and the data of users who had posted it. However, the major social media platforms (Meta, YouTube, Twitter and TikTok) have sued the German government over this requirement and the requirement is currently on hold.[79] During the 2019-2021 parliament, Canada drew inspiration from NetzDG, and the Heritage Ministry presented a technical paper for potential legislation around content removal.[80] The legislation proposed the creation of a new, independent regulator and further measures to safeguard freedom of expression and involve civil society. However, the proposal also came under criticism for the extent of proactive monitoring and filtering required of private companies, and for mandatory reporting of certain forms of harmful communication to the RCMP and the Canadian Security and Intelligence Services.[81,82] These raised concerns about undue limitations of freedom of expression and over-involvement of law enforcement, particularly when these mechanisms are combined with the tendencies for algorithmic bias in speech detection with respect to some racialized communities. In light of such criticisms, the Heritage Ministry stated in February 2022 that it would convene an expert panel to consult on revising the legislation.[83] We hope that the government will continue to push forward on meaningful legislation while addressing these and other concerns.

- We note, however, that any attempts to regulate content moderation should ensure that freedom of expression and human rights remain core principles. As the quantitative case study points out, counterspeech can often be swept up and deleted inadvertently if automated content moderation is used. We have also seen authoritarian and authoritarian-leaning governments instrumentalizing the pandemic to censor and suppress freedom of expression.[84] It is important to find and create policies that do not give licence to broad censorship. In turn, this does not have to mean a laissez-faire approach that leaves all moderation to platforms. We also suggest that policies should be less focused on addressing certain problematic phrases or tactics and more focused on creating processes and principles to enable reflexivity, avoid undue influence over

[78] Anna Meier, Germany's White Supremacist Problem—and What It Means for the United States, *Lawfare*, January 30, 2022.
[79] Clothilde Goujard, Big Tech Takes on Germany, *POLITICO*, February 2, 2022.
[80] Canadian Heritage, Technical Paper (Ottawa: Government of Canada, July 29, 2021).
[81] Cynthia Khoo, Lex Gill, and Christopher Parsons, Comments on the Federal Government's Proposed Approach to Address Harmful Content Online (Toronto: Citizen Lab, University of Toronto, September 28, 2021).
[82] Yuan Stevens and Vivek Krishnamurthy, Overhauling the Online Harms Proposal: A Human Rights Approach (Ottawa: Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic, September 2021).
[83] Rachel Aiello, Feds' Planned Crackdown on Harmful Online Content Getting a Revamp, CTV News, February 3, 2022.
[84] Ong, "The Contagion of Stigmatization."

regulation from platforms, and clarify principled public aims, i.e., oriented toward democratic governance.[85]

- One way to help democratize social media policy-making and limit industry or government capture is to create institutional mechanisms to incorporate views from civil society. In previous work with Fenwick McKelvey, we suggested the creation of a social media council to ensure the inclusion of civil society voices, particularly from marginalized communities.[86]

- Platforms themselves should continue to work on how to improve content moderation and platform design, including their algorithmic recommendation systems. Platforms moved comparatively swiftly to create policies on COVID-19 conspiracy theories in 2020.[87] Platforms are also updating their design in various ways. For example, Twitter is trialing a new feature that prompts users to reconsider posts with "offensive" content; this reduced offensive tweets by 6%.[88] Yet, online abuse has remained a huge problem and conspiracy theorists often manage to evade detection. Such problems are often far worse in languages other than English, partly because content moderators in other languages have poorer working conditions, less support and less clear instructions.[89]

5.2 Encourage institutions and civil society to provide support

- While abuse and hate speech occurs online and is often facilitated by the design or algorithms of social media platforms, this online communication has consequences for many dimensions of people's lives. Institutions can provide considerable support for those experiencing abuse and dealing with its consequences.

- Furthermore, people's exposure to these online harms often comes as a result of their professional responsibilities. For professions or experts communicating online, harassment has become an "occupational hazard."[90] Our case study shows that this has become the case for health communicators during the pandemic. Too often within other

[85] Chris Tenove and Heidi Tworek, Processes, People, and Public Accountability: How to Understand and Address Harmful Communication Online (Ottawa: Canadian Commission on Democratic Expression, September 2020).
[86] Chris Tenove, Heidi Tworek, and Fenwick McKelvey, Poisoning Democracy: How Canada Can Address Harmful Speech Online (Ottawa: Public Policy Forum, November 8, 2018).
[87] Heidi Tworek, Platforms Adapted Quickly during the Pandemic — Can They Keep It Up?, Centre for International Governance Innovation, May 14, 2020.
[88] Matthew Katsaros, Kathy Yang, and Lauren Fratamico, Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content, Preprint accepted at the International AAAI Conference on Web and Social Media, December 1, 2021.
[89] Sarah Emerson, Facebook's Spanish-Language Moderators are Calling Their Work a 'Nightmare', *BuzzFeed News*, January 13, 2022.
[90] Chris Russill, Harassment as Occupational Hazard, *Canadian Journal of Communication* 46, no. 4 (2021): 745–49.

professions like journalism[91] or academia,[92] harassment is left to individuals to solve and too few institutional and organizational forms of support are offered. A committee of the Royal Society of Canada proposed that "post-secondary institutions have a readily accessible policy and action plan in place to support scholars who are significantly harassed, threatened, or intimidated."[93] Similarly, we recommend that all institutions employing public-facing people develop organizational structures and plans to support anyone who experiences harassment, threats or intimidation.

- Civil society actors can help provide support, including for individuals who lack a supportive institutional employer. They can also develop tools to help those who have suffered online hate and abuse targeting particular groups. For example, Indigenous lawyer Naomi Sawyers has set up an online tool to enable "victims of online harassment to help create their own reports and to encourage safer online practices."[94] This lowers the barriers for victims who often do not report such online harassment because the reporting mechanisms are too complex, burdensome or unresponsive.

5.3 Empower individuals to address online hate more effectively

- Policy approaches to online abuse and hate often focus on the regulatory relationship between governments and platforms. This leaves out the users who post hate speech and hate promotion and those who are impacted by it. Individuals vary widely in their strategies for navigating online abuse, as we have found in our research on health communicators, politicians and journalists. Some go to great lengths to personally engage antagonistic accounts, seeking to forge a connection and overcome prejudices. More often, individuals develop practices to limit their engagement with hostile voices or toxic content. Some modes of engagement may be more likely to create controversy and abuse than others. For example, users may choose not to engage with those who spread conspiracy theories so as to avoid attracting attention from such groups. This is not a question of policing tone, but of users, platforms and governments learning from research into what might create constructive conversations online or best achieve the aims of communicators.[95]

- Further education on platform tools can help individual users to find some protections for themselves from online abuse. In some cases, people are required or encouraged to engage in social media as part of their work. It is important for employers to give employees explicit permission to take self-protective actions, rather than pressuring

[91] Avery E. Holton et al., 'Not Their Fault, but Their Problem': Organizational Responses to the Online Harassment of Journalists, *Journalism Practice,* Advance online publication (July 5, 2021): 1–16.
[92] Shandell Houlden et al., Support for Scholars Coping with Online Harassment: An Ecological Framework, *Feminist Media Studies,* Advance online publication (February 10, 2021): 1–19.
[93] Julia M. Wright et al., Protecting Expert Advice for the Public: Promoting Safety and Improved Communications (Ottawa: Royal Society of Canada, February 2022), 7.
[94] Naomi Sayers, Tech-Facilitated Violence: Online Harassment, MeetNaomi, December 16, 2021.
[95] Jef Akst, Twitter's Science Stars Fight Misinformation, *The Scientist Magazine*, January 17, 2022.

them to pursue unhealthy over-engagement. We note, further, that individual actions will not be sufficient to address systemic problems.

5.4 Create streamlined processes for dealing with online threats

- Often it is too cumbersome or complicated to report online threats to the police or other government agencies. Sometimes it is not clear to users if something rises to the level of a criminal offence. At other times, responses to complaints are uneven, slow or do not grasp the severity of the threat.
- In some instances, those who experience abuse may be reluctant to report that abuse to the RCMP. This raises questions about whether other bodies might be better placed to be a first port of call for some of these problems, such as reporting abuse. An alternative reporting process, in addition to the police or social media platforms themselves, could generate an important evidence base and act as a point of contact for providing individuals with resources.

While this report has focused on online hate, many of the drivers are historical, social and offline. Addressing online hate requires addressing offline behaviours and broader societal issues. While some of the broader social drivers of hate – online and offline – have been exacerbated by the pandemic, there have also been positive developments. Online communication can serve as both a window to understand and a tool to promote both those negative and positive developments.

CASE STUDY A: Hate and harassment targeting health communicators

Authors: Dr. Heidi Tworek[1] and Dr. Chris Tenove[2]
Researchers: Wilson Dargbeh,[3] Hanna Hett[4] and Oliver Zhang[5]

This case study emerged from a larger project investigating online abuse of health communicators in Canada, funded by a Social Sciences and Humanities Research Council Partnership Engage Grant, no. 892-2021-1100.

A.1 Main findings

- During the pandemic, public health officials, medical practitioners and health experts engaged in unprecedented levels of public communication, including online.
- As the pandemic continued, they faced escalating levels of online abuse, often linked to waves of infections, vaccine mandates and other public health measures, and broader political conflicts.
- Key themes of abuse include accusations of corruption and incompetence, responsibility for widespread injury, and loss of liberties. Health communicators face abuse from individuals who consider public health measures to be too extensive, but also from those who consider the measures to be insufficient.
- Explicit racism, xenophobia and misogyny figure in a small but disturbing proportion of messages. More common are messages that seek to undermine the authority of women or racialized health communicators.
- Online abuse and hate affect the safety and well-being of health communicators, as well as their ability to effectively promote health-related information.
- Abuse toward health communicators, but also toward the vaccine hesitant and other groups, is intertwined with broader patterns of polarization and toxicity online.
- Health communicators require support from employers and other institutions to help them manage online abuse and hate, in addition to more consistent action from social media platforms and law enforcement.

---

[1] Canada Research Chair (Tier II) and Associate Professor, School of Public Policy and Global Affairs and Department of History, University of British Columbia
[2] Postdoctoral research fellow, School of Public Policy and Global Affairs, University of British Columbia
[3] Student, Master of Public Policy and Global Affairs, University of British Columbia
[4] Student, Master of Journalism, University of British Columbia
[5] Student, Master of Public Policy and Global Affairs, University of British Columbia

A.2 Introduction

Accurate and effective health communication is critical to address the COVID-19 pandemic and other public health challenges. Health communicators, from public health officials to university-based experts, have used social media innovatively to engage broad publics and specific communities.[6] They have also tried to address widespread misinformation, created inadvertently, and disinformation, intentionally spread for political and economic aims.

In making these efforts, health communicators have too often faced abuse or threats.[7,8,9] A global survey by *Nature* of scientists who discussed the pandemic on news media or social media found over two-thirds reported negative experiences, 22% received threats of physical or sexual violence, and 15% received death threats.[10] Many Canadian news articles document online abuse of health communicators.[11] Health communicators are not the only professionals to experience online abuse, but our work focuses on them to provide insights into the dynamics of abuse online.

This case study provides a snapshot of findings from an ongoing research project into online abuse of Canadian health communicators led by Heidi Tworek and Chris Tenove at the University of British Columbia, which includes research collaborators at Royal Roads University and the University of Ottawa, as well as the civil society organization, the Institute for Strategic Dialogue. We have used the umbrella term of "health communicators" to encompass several different groups of people communicating about COVID-19: public health officials, health care workers, university-based experts and journalists writing about the pandemic.

The experiences of health communicators are a valuable window into the broader problem of online hate in British Columbia during the pandemic for three reasons. First, many health communicators have faced online abuse during the pandemic, often related to their personal identities. Their experiences help reveal the forms that harassment and hate can take and the impacts they can have. Second, our study reveals relationships between the pandemic context, particularly social and political conflicts over vaccines and other public health measures, and the online abuse and hate that individuals may face. Third, we identify actions that health

---

[6] Heidi Tworek, Ian Beacock, and Eseohe Ojo, Democratic Health Communications during Covid-19: A RAPID Response (Vancouver: Centre for the Study of Democratic Institutions, University of British Columbia, 2020).
[7] Cheryl Clark, Insults, Threats of Violence Still Imperil Public Health Leaders, *MedPage Today*, February 25, 2021.
[8] Sara H. Cody, Dealing With Harassment in Public Health, *Journal of Public Health Management and Practice* 27, no. 4 (2021): 432–33.
[9] Victoria Smith and Alicia Wanless, Unmasking the Truth: Public Health Experts, the Coronavirus, and the Raucous Marketplace of Ideas (Washington, DC: Carnegie Endowment for International Peace, 2020).
[10] Bianca Nogrady, 'I Hope You Die': How the COVID Pandemic Unleashed Attacks on Scientists, *Nature* 598, no. 7880 (October 13, 2021): 250–53.
[11] Among others, see: Penny Daflos, 'We're Just Human Beings': B.C. Doctors Face Abuse, Threats, Doxing amid Pandemic Fatigue, CTV News, May 28, 2021; Sharon Kirkey, 'He Was Looking to Meet Me': Public Health Leaders Face Threats, Harassment over COVID-19, *National Post*, September 29, 2020; Evelyn Kwong, Hacked and Impersonated: Four of Ontario's Top Health-Care Voices on Being Targeted and Harassed on Social Media, *The Toronto Star*, May 13, 2021; Heidi Tworek, As Omicron Surged, So Did Abuse of Health Communicators Online, Centre for International Governance Innovation, January 12, 2022.

communicators and supporting institutions have taken or should take to respond to online abuse and hate. These responses may be relevant for other public emergencies and crises, for other groups that are targets of hostility and hate online.

A.3 Research methods

Our overall study focuses on Canada and pays attention to how factors like race and gender might affect the extent, impact and responses to hate and hostility that Canadian health communicators have encountered during the pandemic. This contrasts with previous studies that have focused on how prominent white male health communicators like Dr. Anthony Fauci in the U.S. or Dr. Christian Drosten in Germany receive considerable online abuse, including death threats, and that social media platforms like Facebook have been slow to remove many of these threats, even when they contravene platforms' own terms of service.[12]

The work here draws on three main methods:

1. A literature review of academic scholarship and a news scan of articles published by Canadian journalism organizations that detail harassment or abuse faced by public health officials, health experts, health practitioners and health journalists. From March 2020 to December 2021, 22 different Canadian news outlets provided original reporting on over 45 different incidents of online harassment and threats against health communicators.

2. Interviews with 22 health communicators, including public health officers, staff of public health agencies, university-based health experts, health practitioners and health journalists. Fifteen interviewees identify as racialized, 11 as women, and 11 are based in B.C.

3. Exploratory analysis of Twitter activity and public engagement of approximately 100 Canadian health communicators on Twitter. This online participant observation is part of a more systematic, forthcoming study.

A.4 Findings

Increased online presence. During the pandemic, many health communicators became more deeply engaged in public communication, including online.

- Public health officials and agencies faced sharp increases in public interest, including in B.C. As one interviewee put it: "The public turned to us in a way we'd never seen before." B.C.'s provincial health officer, Dr. Bonnie Henry, along with other public health officials across the country, became more central to public debates.[13] Dr. Henry is particularly prominent in media sources: she was the most quoted woman in

[12] Avaaz, Scientists under Attack (January 21, 2022).
[13] Giuseppe Valiante, A New Breed of Celebrity in the Age of COVID-19: The Chief Medical Officer, *National Post*, March 23, 2020.

Canadian media for 13 of the 22 months between March 2020 and December 2021.[14] Although Henry herself is not an active user of social media, her frequent press conferences and media coverage dramatically raised her online profile.

- Many health experts and medical practitioners also increased their online engagement. According to interviewees, many did so to address the intense public demand for more information regarding COVID-19 during the early months of the pandemic. This included promoting behaviours such as wearing masks, social distancing and, later, vaccination. Some identified gaps in knowledge in specific communities that they could address, such as the South Asian community, the Black community and the homeless or under-housed.[15] In addition to providing information, some of these health experts and medical practitioners took on explicit advocacy roles, calling for changes in policy.

Key themes of abuse. Health communicators often face criticism or counter-arguments online. Individuals differ in their assessments of when criticism tips over into abuse. Our interviewees and our participant observation on Twitter suggest that more intensely negative or abusive messages often address a limited set of themes. The first four themes align with the findings of Hughes et al,[16] and we draw on their terms:

- "Sinister Origins" includes claims that there has been a cover-up regarding the origins of the COVID-19 virus or vaccines to counter it. Particularly in the initial months of the pandemic, these accusations focused on the role of the Chinese government. Other sinister origins include a "globalist" or New World Order actors, sometimes including Microsoft founder Bill Gates or Jewish philanthropist George Soros. Some of these narratives existed before the pandemic and employ antisemitic tropes.

- "Corrupt Elites" includes accusations that health communicators are primarily motivated by economic or political gain, rather than trying to advance the public good. This includes accusations that health communicators are paid off by Big Pharma or that journalists are taking instructions from political leaders.

- "Causing Injury or Death" includes accusations that health communicators are personally responsible for harm caused by vaccines, mask use or other health measures. This includes fabricated stories regarding major injury or death from vaccines.[17] However, our interviewees also report being attacked for not aggressively promoting

[14] Calculated from https://gendergaptracker.research.sfu.ca/apps/topsources.
[15] For an example of a civil society effort to advocate for a community, and address gaps in information they experienced, see the work of the South Asian Covid Task Force.
[16] Brian Hughes et al., Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation, *International Journal of Environmental Research and Public Health* 18, no. 14 (2021): 1–18.
[17] Stephen Maher, Misinformation from the U.S. Is the next Virus—and It's Spreading Fast," *Macleans*, January 3, 2022.

some health measures or for questioning the efficacy of some measures (e.g., general lockdowns or vaccines for children).

- "Freedom under Siege" includes accusations that health communicators are responsible for public health measures that violate people's rights, including by jeopardizing the viability of businesses (through lockdowns) or employment (through requirements for vaccination). This theme was most closely associated with right-wing political figures. It has also become a core message of the trucker protests and "Freedom Convoy."

- "Incompetence" includes claims that health communicators do not understand science, or fail to enact their duties as medical practitioners, public officials or journalists. They are often told to "do their job." In some cases, individuals were told that complaints would be lodged via professional bodies such as colleges of physicians. Interviewees told us that accusations of incompetence or error from other health professionals or peers caused them particular stress, since they felt their professional identity was under attack.

It is important to acknowledge that criticisms of public health responses, including those that echo these themes, are not necessarily invalid. There are legitimate concerns about the profit-seeking of pharmaceutical companies, the motivations of policymakers, the harms that may result from requiring or not requiring health measures (e.g., vaccine or mask mandates), and the trade-offs in individual freedoms that have resulted from health measures. Interviewees told us that such issues ought to be discussed, but in communication that is neither demeaning nor abusive.

Racism, xenophobia and misogyny. While online abuse has affected almost all health communicators, some have received more explicit attacks on their identity than others, particularly based on their gender, race or religion. Most interviewees reported that explicit racist, xenophobic or misogynistic content is a very small proportion of the negative messaging they get, but they experience it as particularly virulent and disturbing.

- Health communicators' ethnic, racial or religious identity has sometimes been targeted as part of their abuse. The most prominent recipient of such attacks is Canada's chief public health officer. Dr. Theresa Tam has been subjected to abuse targeting her Chinese ethnicity and has been questioned about whether her loyalty is to Canada or China.[18,19] Our own social media analysis corroborated these claims, identifying tweets labelling her as "maoist scum" or a "commie witch," as well as manipulated images, including one showing her saluting China's President Xi Jinping. Tam has also faced abuse that intentionally misrepresents her gender identity. Scholars who document these incidents have noted "a pattern of users mobilizing gendered and racialized

[18] Alex Boutilier, 'Does She Work for Canada or for China?' Conservative MP's Attack on Dr. Theresa Tam Draws Fire, *The Toronto Star*, April 23, 2020.
[19] Steven Zhou, Coronavirus Conspiracies Give Boost to Canada's Far-Right, *Foreign Policy*, May 18, 2020.

discourses to undermine the message, sow public distrust, and challenge the authority of Dr. Tam."[20]

- More generally, health communicators have faced attacks on their race, ethnicity, religion or gender that seek to undermine their credibility or exclude them from participation in public discussion. For instance, a Sikh doctor in B.C. received racist and xenophobic comments after contributing to a television newscast.[21]

- Health communicators also felt targeted by online narratives blaming ethnic communities for the global circulation of the virus. As one B.C.-based South Asian medical practitioner said to us: "You saw from very early on it was called the 'China virus,' and then by the time you got to Delta, it was the 'Indian virus.' But no one called the Alpha variant the 'London virus.' There's a lot of inherent and implicit racism."

- Researchers have identified antisemitic themes in conspiracy theories regarding the origins of the virus and vaccines, tying the pandemic to longstanding conspiracies about the globalist agenda of individuals such as Jewish American investor and philanthropist George Soros, who survived the Nazi occupation of Hungary as a child. In addition, Jewish groups have denounced those protestors who equate restrictions imposed on the unvaccinated in Canada with the Jewish experience during the Holocaust.[22]

- Women observe that attacks often have a misogynistic element, whether in tone or explicit content.[23] Many women we interviewed believed that they encountered challenges to their authority or expertise that would not happen if they were a man. In the case of racialized women, attacks may be intersectional, meaning that they focus on intertwined identities of race and gender.

Violence and threats of violence. Health communicators have faced threats to their safety online, and some of threats relate to targeting of their identity.

- Reports have described death threats against Bonnie Henry and other public health officials.[24,25] Several of our interviewees had received explicit threats of violent action against them and had filed reports with police.

---

[20] Anna Calasanti and Bailey Gerrits, 'You're Not My Nanny!' Responses to Racialized Women Leaders during COVID-19," *Politics, Groups, and Identities*, Advance online publication (June 30, 2021): 1–18.
[21] Bridgette Watson, Sikh Doctor Subjected to Racist Comments Following B.C. Newscast Says It's a 'Reality Check,' *CBC News*, March 4, 2021.
[22] Rachel Bergen, Antisemitic Rhetoric Continues to Be Used by Some Opponents of COVID-19 Measures, *CBC News*, October 10, 2021.
[23] April Lawrence, Women Leaders Speaking out after Learning Dr. Bonnie Henry Has Received Death Threats, *CHEK News*, September 23, 2020.
[24] Lawrence, "Women Leaders Speaking out after Learning Dr. Bonnie Henry Has Received Death Threats."
[25] Mason DePatie, Manitoba's Top Doctor Says He's Been Target of Online Threats during Pandemic, *CTV News Winnipeg*, July 12, 2021.

- Many interviewees had received threats of future punishment, including vague comments about "getting what's coming to you," as well as calls for them to face criminal trials or international criminal trials reminiscent of the post-WW2 trials in Nuremberg. Such threats were also mentioned in news accounts,[26,27] including a report that an individual in a livestreamed video stated that Bonnie Henry should be "tried in a court – given a fair trial and then hung."[28] Some interviewees dismissed such comments as empty hyperbole, while others found them indicative of intense hostility that could lead to violent actions.

- Three interviewees received violent threats online that explicitly referenced their non-white racial or ethnic identity. More broadly, vulnerability to physical threats due to health communication are intertwined with additional feelings of vulnerability due to an individual's race, ethnicity or sexual orientation. As one interviewee told us, while he hasn't received explicit threats of violence online, he remains alert to the possibility that he might be targeted for his health communication or for his ethnic identity. "Part of this is being conditioned as a Sikh man living in Canada…I'm pretty well protected, I'm in a community that's generally safe, but at any point, I could get attacked anyway, and that's my reality."

Potential consequences of online abuse and hate. Our interviews with health communicators, similar to our research on the impacts of online abuse for Canadian politicians,[29] revealed that online abuse can result in harm along multiple dimensions.

- *Psychological health and well-being.* Most interviewees explained anxiety, distress, grief and other consequences of online abuse. For instance, one described nightmares about being attacked at their clinic by some of the individuals who sent hostile messages online.

- *Professional effectiveness*. For a variety of reasons, online abuse can undermine people's ability to be effective health communicators. They describe spending extensive time assessing – and sometimes responding to or blocking – negative comments. They expressed concern that their public health information was diluted because of the toxic or false comments that quickly followed their posts. Some health communicators, particularly those who were contract workers or precariously employed, expressed concern that being harassed online could lead to future risks to

[26] André Picard, Opinion: The Troubling Nazi-Fication of COVID-19 Discourse, *The Globe and Mail*, August 16, 2021.
[27] Aaron D'Andrea, 'Exhausted and Scared': Canada's Doctors Call for Help to Stop Online Hate, *Global News*, November 10, 2021.
[28] Kelvin Gawley, Conspiracists Wishing for Dr. Henry's Execution Terrible but Unsurprising, Experts Say, *CityNews Vancouver*, February 24, 2021.
[29] Chris Tenove and Heidi Tworek, Trolled on the Campaign Trail: Online Incivility and Abuse in Canadian Politics (Vancouver: Centre for the Study of Democratic Institutions, University of British Columbia, 2020).

employment, either because false claims were being posted about them or because they were being labeled as too provocative.

- *Chilling effect.* One further concern is that abuse may drive communicators from active engagement online, a phenomenon often called a "chilling effect." Research has shown that such chilling effects more commonly affect women and members of marginalized ethnic, racial and gender minorities.[30,31]

- *Corroding debate and exacerbating polarization.* Many health communicators expressed a concern that the abuse and hostility they elicited, originating from a small but highly active set of hostile commenters online, was preventing the public from having rational and inclusive discussions of important public health issues. Some were concerned that the hateful or racist comments that they elicited online would encourage further expressions of bigotry. More broadly, some worried that social media discussions of health issues were becoming motors of widening and entrenching social cleavages. This included the stigmatization or vitriol directed at individuals who are vaccine hesitant or critical of public health measures. Interviewees saw the hostility directed against individuals holding such positions to be corrosive to social solidarity and to public health communication.

A.5 Responses to online hate and hostility

When faced with hate and hostility, health communicators have developed a range of personal practices and collective or institutional support. We will briefly summarize these and highlight gaps that ought to be addressed.

Responding on platforms. Social media platforms offer a range of tools to address hate, harassment or other problematic content. Users can block or mute accounts and, in some cases, can delete or filter content directed at them. Furthermore, individuals can report communication to the platforms that they believe violates the platforms' terms of service. Health communicators regularly use these tools, but our interviewees identify many inadequacies, including slow or non-existent responses from platforms.

Institutional support. The collective or organizational support an individual receives is critical to mitigating online abuse.[32,33] When individuals shoulder the burden of managing online abuse, they are more likely to self-censor or withdraw from public communication, or experience

[30] Jon Penney, Online Abuse, Chilling Effects, and Human Rights, in *Citizenship in a Connected Canada: A Research and Policy Agenda*, ed. Elizabeth Dubois and Florian Martin-Bariteau (University of Ottawa Press, 2020).
[31] Julie Posetti et al., The Chilling: Global Trends in Online Violence Against Women Journalists, (Paris: UNESCO, April 2021).
[32] Shandell Houlden et al., Support for Scholars Coping with Online Harassment: An Ecological Framework, *Feminist Media Studies*, Advance online publication (February 10, 2021): 1–19.
[33] Alex Ketchum, Report on the State of Resources Provided to Support Scholars Against Harassment, Trolling, and Doxxing While Doing Public Media Work, *Medium*, July 14, 2020.

burnout, emotional exhaustion and other mental health issues.[34] The institutional support described by interviewees was very uneven.

Legal protection. As protests against hospitals began to impede care and caused health care workers to fear for their safety, calls grew in fall 2021 for the federal government to pass legislation to protect health care workers.[35] This culminated in the passage of Bill C-3 in December 2021, which amended the Criminal Code and made intimidating a health professional or restricting access to their place of work an offence.[36] It is unclear, however, how this might affect online abuse of health communicators. It also remains to be seen how these new protections are implemented and enforced.

A.6 Conclusion

This short study of online abuse of health communicators helps shed light on the extent and impacts of online abuse and hate during the pandemic. Our interviewees and news reports reveal the frequent hostility faced by individuals who chose to engage different communities in B.C. and Canada on pandemic-related health issues. A very small proportion of this hostile communication rose to the level of private or public hate speech, which disparages, threatens or deeply insults people according to their identity or social group affiliations. However, the majority of racialized individuals we interviewed had experienced at least one instance, and some faced it more regularly. Much more frequently, individuals faced insults and ambiguous threats. These messages reached health communicators via the same online channels (social media and email), from similar types of sources and had similar effects on people's psychological health and professional effectiveness. However, the instances of hate speech – most often racist – promote broader social harms, exacerbating deep-rooted inequities and inter-group conflicts, and therefore deserve particular attention.

Our interviews have also uncovered the range of offline effects of such online abuse. This case study suggests the need to think about broader policies to protect groups who become central communicators during future crises, such as climate change, or ongoing crises, such as the opioid overdose epidemic. While focused legal protection for health care workers is important, the systemic issues of online abuse will not be addressed by attempting to solve these problems group by group. Rather, our case study has uncovered trends that will need broad policy solutions to mitigate the online and offline harm of online abuse.

---

[34] George Veletsianos et al., Women Scholars' Experiences with Online Harassment and Abuse: Self-Protection, Resistance, Acceptance, and Self-Blame, *New Media & Society* 20, no. 12 (2018): 4689–4708.
[35] Adam Miller, 'Under Attack': Canadian Health-Care Workers Call for More Protection from Harassment and Threats, CBC News, November 13, 2021.
[36] Government of Canada, "An Act to Amend the Criminal Code and the Canada Labour Code," Pub. L. No. C–3 (2021).

CASE STUDY B: Hate and the COVID-19 pandemic - An analysis of B.C. Twitter discourse

Authors: Matt Canute,[1] Hannah Holtzclaw,[2] Alberto Lusoli[3] and Wendy Hui Kyong Chun[4]

This case study emerged from a larger project at the Digital Democracies Institute. The Institute's From Hate to Agonism Project, funded by a UK-Canada Responsible Artificial Intelligence (A.I.) grant, is developing innovative and responsible machine learning approaches to support healthy democratic dialogue online.

B.1 Main findings

- During the pandemic we saw an increase in tweets classified under the anti-Asian hate topic:
    1. Natural language processing (NLP) text-model results showed an increase in hate speech in March 2020, when B.C. declared a provincial state of emergency.
    2. The increase in hate speech was accompanied by an even larger increase in tweets classified as counterspeech. This finding is meaningful as it shows how the proliferation of hateful and harmful speech triggered an oppositional, and larger, response. However, reactionary counterspeech developing within highly toxic environments can further polarization rather than contribute to constructive dialogue over differences and conflict.
    3. The conversation about anti-Asian hate in B.C. was highly susceptible to events taking place outside of the province and country, particularly events in the U.S. Specifically, we saw a dramatic increase of tweets classified as counterspeech in the wake of the tragic Atlanta, GA, spa shooting in 2021, as well as an increase in tweets attacking specific identities when notable and contentious events occurred in the United States (e.g., George Floyd murder, U.S. Capitol riot).
- Data also show an increase in toxicity in general conversations about COVID-19 in B.C. and government management of the crisis (COVID-19 topic). Tweets within this topic expressed frustrations directed towards restrictions and vaccine mandates, political leaders and health officials, as well as individuals defying lockdown orders or public health order restrictions such as wearing a mask.

[1] Data Scientist, Digital Democracies Institute, Simon Fraser University
[2] PhD researcher, Digital Democracies Institute, Simon Fraser University
[3] Postdoctoral researcher, Digital Democracies Institute, Simon Fraser University
[4] Canada 150 Research Chair, Director Digital Democracies Institute, Simon Fraser University

- The effectiveness of text models decreased when these were applied to novel contexts (e.g., Wikipedia trained model used to analyze tweets, anti-Asian trained model used to analyze COVID-19). This limitation represents a challenge for researchers as well as for social media platforms, whose algorithms similarly struggle to contextualize language use across platforms, communities, cultures and subcultures.

B.2 Introduction

In this case study, we examine anti-Asian hate speech on Twitter in British Columbia during the COVID-19 pandemic. We analyze how the anti-Asian rhetoric developed in relation to conspiracy theories about the origin of the virus and within broader online conversations about COVID-19. We consider hate speech as "a form of online public communication disparaging, threatening, or deeply insulting people according to their identity or social group affiliations."[5] One crucial question for understanding online hate during the pandemic is whether online hate has increased. We take readers through the steps required to find and analyze hate speech, showing why it is very complicated and difficult to understand whether hate speech has actually increased during the pandemic. We also consider the role and the relevance of counterspeech, meaning posts and speech that resist, combat or try to push back against racism and hate.

As we show in this case study, complexities emerge at every stage of trying to understand the amount of hate speech online, from collecting to analyzing data. This means that researchers cannot offer a simple answer to the question about the quantity of hate speech online. Instead, our findings reveal a complex scenario. On the one hand, more posts were classified as anti-Asian hate speech in 2020 than in 2019. On the other hand, counterspeech also increased during the pandemic, especially in 2021.

Counterspeech is understood as speech that seeks to oppose, refute or undermine hateful speech. Though tactics of counterspeech vary across historical and cultural settings, examples of contemporary digital counterspeech include organized group counter-messaging campaigns as well as organic and individual responses to hateful contents.[6] Since hate speech and counterspeech share many similarities in tone, computational approaches to content moderation, such as AI and machine learning algorithms for abuse detection, often fail to tell the difference. In our case study, then, we combine computational methods with a qualitative and manual analysis of tweets.

This case study emerged from a larger project at the Digital Democracies Institute. The Institute's *From Hate to Agonism Project*, funded by a UK-Canada *Responsible A.I.* grant, is developing innovative and responsible machine learning approaches to support healthy democratic dialogue online.[7]

---

[5] Chris Tenove and Heidi Tworek. *Online Hate in the Pandemic*, 2022.
[6] For more information on counterspeech, see Counterspeech, Dangerous Speech Project. Accessed February 16, 2022.
[7] For more information about the Hate to Agonism project, visit https://digitaldemocracies.org/.

B.3 Data collection

The first challenge for measuring online hate in B.C. is determining which online communication counts as occurring within the province. The most straightforward way to do so is to investigate social media posts by users located in B.C. Independent researchers do not have access to data that would enable them to measure hate on the most commonly used platforms by British Columbians, including Facebook and Instagram. Retrieving data such as Facebook and Instagram posts, pages and profile information, while technically feasible, could violate platforms' terms of services and users' privacy. Therefore, we focused on Twitter, which does make such data publicly available and which was used by approximately one out of four B.C. residents in 2021.[8] Even though tweets are publicly available, we have removed all usernames to protect users' privacy.

Investigating Twitter has certain benefits. The platform is an efficient means for sampling public discourses developing around emerging issues. Moreover, historical data are available which allow researchers to conduct retrospective analyses on current issues. To build the dataset, we relied on Twitter's academic application programming interface[9] (API), an initiative launched in July 2020 by Twitter. Academic API allows research institutions to retrieve real-time and historical Twitter data. Our study focused on all tweets published between January 2019 and December 2021 and geolocated in British Columbia. Overall, we were able to identify and retrieve approximately 6 million tweets. Filtering and grouping tweets by keywords, our research team was able to capture user participation in ongoing public conversations about COVID-19, anti-Asian hate and conspiracy theories during the pandemic.

The decision to study Twitter also comes with limitations. First, the demographics of Twitter users might differ from the broader B.C. population. For example, Twitter users in Canada tend to be male, urban and over 35 years old.[10] Second, it is not possible to identify all communications by Twitter users in B.C. Users have to opt-in to have their tweets associated with a location (also known as geocoding or geolocating one's tweets). Previous studies suggest that only 0.7% of tweets contain geographic information and that factors such as socioeconomic status, location and digital literacy are likely to influence the decision to geolocate one's tweets.[11] While we examined the activity of 103,421 geolocated users tweeting from B.C., we do not know whether this subset of users conducts themselves differently from those who do not geolocate their tweets. Lastly, since we are conducting this case study retroactively, we are unable to collect tweets that Twitter removed because they violated its terms of service. It is not possible to know how much hateful or hostile content has been removed or how long that content remained public before its removal.

---

[8] Canadian Internet Registration Authority. Canada's Internet Factbook 2021 - Full Survey Results. 2021.
[9] API can be understood as interfaces allowing third-party software applications to access the data and functionalities of popular online services.
[10] Statista. Social networks: Twitter in Canada 2021 Brand Report. November 2021.
[11] Graham, Mark, Scott A. Hale, and Devin Gaffney. Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer 66*, no. 4 (October 2, 2014): 568–78.

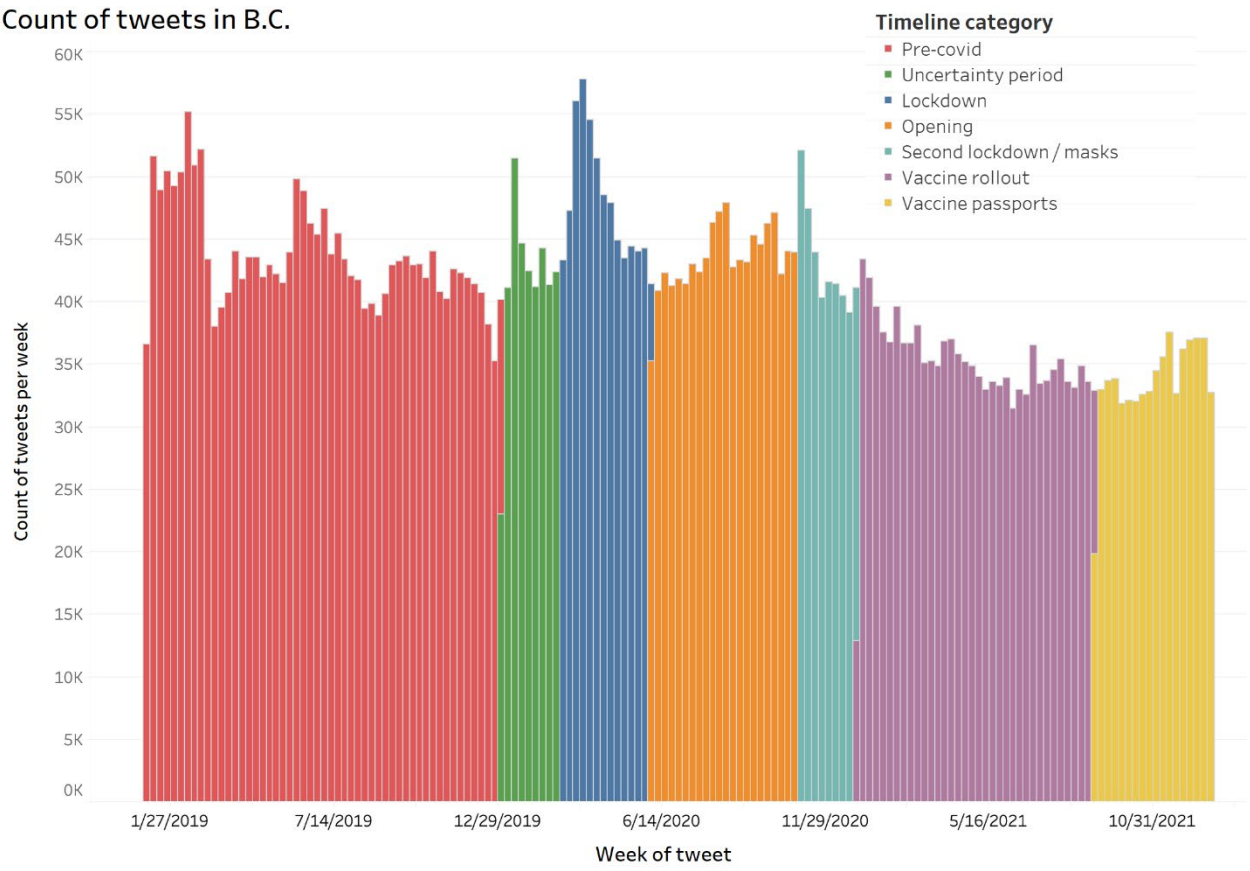
To analyze our data and understand whether the pandemic affected how British Columbians debated public issues online, we relied on a combination of quantitative and qualitative methods. As discussed in greater detail in Section B.4, we first analyzed the B.C. Twittersphere through a combination of three open-source natural language processing (NLP) text models.[12] The creators of those models developed them to identify patterns and common semantic structures of hate speech as well as toxic language over time. Toxic language ranges from rude and obscene tweets, to disrespectful and inflammatory comments to direct attacks targeting individuals or groups. Hate speech, instead, is a more specific form of online communication and, in this context, refers to tweets threatening or insulting people according to their identity or social group affiliations. We then drew upon qualitative research methods and conducted a close textual analysis of subsets of tweets about anti-Asian hate, conspiracy theories and COVID-19 to identify discursive and linguistic nuances that the above-mentioned NLP text models were unable to recognize.

The following sections provide a detailed overview of the data collection and sampling techniques. Please note that we do provide some samples of toxic tweets. However, due to the sensitive nature of some of the contents analyzed (for example, extremely toxic tweets about race, politics, religion, etc.), we have not provided the most offensive examples.

B.3.1 Sampling hate speech on Twitter

Through the Twitter Academic API we were able to retrieve all tweets (excluding retweets) geolocated in B.C. and published before and during the pandemic. Specifically, we collected all B.C. tweets published between January 1, 2019 and December 31, 2021. Overall, we collected 6,034,975 tweets, with an average of around 5,500 tweets per day. The following figure shows the volume of tweets each week, coloured to illustrate different time periods during the pandemic in B.C.

[12] NLP is a branch of computational science for analyzing and developing rule-based models of human language.



As expected, there appears to have been a brief spike in online activity during the few weeks after a provincial state of emergency was declared in B.C. in March 2020.

B.3.2 Narrowing the analysis

After collecting all available tweets published in B.C. from 2019 to 2021, we focused the analysis on tweets about anti-Asian hate (n=3,369). In addition, we analyzed tweets about conspiracy theories (n=12,499) and about COVID-19 (n=69,771). The inclusion of these two additional topics allowed us to better frame our analysis about anti-Asian hate and to understand how conspiratorial rhetoric circulating on Twitter before and during the COVID-19 pandemic intersected and supported forms of hate speech directed toward Asian people. We created the three topics by searching within our dataset for tweets containing specific keywords or combinations of keywords. We built on previous studies to define the lists of keywords used to

delimit each topic.[13] We briefly describe those three topics and give an illustration before providing an analysis.

Anti-Asian hate: This topic contains all tweets expressing hatred or aggression towards Asian people. It was identified by searching for all tweets containing words such as *China*, *Wuhan*, *CCP*, *virus*, as well as common racial slurs targeting Asian people. Examples of tweets found in this topic include:

> That describes the current ruling establishment in Communist China, and their #CCP princelings #satelliteFamilies in Vancouver. Time to hold all of them liable and use #MaginskyAct to seize their #vanre and assets to pay for #covid19 havoc. (May 2020)

Conspiracy theories: This subset contains all tweets dealing with conspiracy theories, not only those related to COVID-19. This topic was identified by filtering tweets containing words such as *deep state*, *plandemic*, *soros*, *nwo (New World Order)*, etc. Examples of tweets found in this topic include:

> China responsible for Covid-19. Documents obtained by the Daily Mail confirm that Communist China had been harvesting, developing, and testing novel coronaviruses on mammals using grant money from the U.S. government under former President Obama. (April 2020)

COVID-19–related discussions: This topic includes all tweets about the pandemic and related issues such as government management of the crisis, mask mandate and vaccines. This topic was identified by filtering tweets containing words such as *covid*, *pandemic*, *endemic*, *vaccine*, *mask*, etc. Examples of tweets found in this topic include:

> I work in a COVID hospital and I've just had it with people's ignorant fucking bullshit about this being fake/exaggerated/planned (whatever the fuck term these idiots are using). I'm just fucking done with idiots. DONE I tell you. (December 2020)

To make sure that the three topics were relevant within the B.C. Twittersphere and that we were not omitting other significant topics, we validated them using topic modeling. Topic modeling is a machine learning technique routinely employed to identify salient topics within large datasets. By analyzing our entire dataset through a popular topic modeling algorithm,[14] we were able to confirm that our three topics were relevant issues in the B.C. Twittersphere in the period of analysis.

---

[13] Specifically, we relied on: He, Bing, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 90–94. Virtual Event Netherlands: ACM, 2021; Moonshot. Antisemitism within Anti-Vaccination Conspiracy Theories. June 2021.

[14] Specifically, we relied on BERTopic, a topic clustering and modeling technique.

B.4 Data analysis

To analyze the discourses developing around the three topics of interest (anti-Asian hate, conspiracy theories and COVID-19), we relied on a combination of qualitative and quantitative methods.

To begin, we analyzed the tweets composing three topics defined above (anti-Asian hate, conspiracy theories and COVID-19–related discussions) through three different text-based machine learning classifiers. As the name suggests, text classifiers are algorithms employed to categorize texts into pre-determined categories. We relied on text classifiers to categorize tweets as hate speech or counterspeech, as well as to assess their toxicity. In addition, we conducted a human-based, qualitative analysis on a subset of tweets. The findings emerging from the qualitative analysis (discussed in Section B.5.3) allowed us to develop a more nuanced understanding of how the anti-Asian rhetoric developed on Twitter in relation to conspiracy theories about the origin of the virus and within broader online conversations about COVID-19.

In addition, the qualitative findings helped us highlight some of the limitations of quantitative methods using automated text models. For instance, text classifiers in general are unable to discern hate speech from counterspeech due to the similar tones that these messages sometimes share. In the following two sections, we describe the tools employed in this research in greater detail.

B.4.1 Identifying hate speech and counterspeech: A quantitative approach

The first text classifier we used to analyze tweets was the *GaTech* hate speech/counterspeech model.[15] This is a text classifier algorithm developed at the Georgia Institute of Technology and explicitly designed to detect anti-Asian hate on Twitter. In previous studies, the model achieved good performance in classifying tweets as hate speech, counterspeech or neutral. [16] Applying the *GaTech* model to our dataset, we analyzed hate speech and counterspeech in the three topics of interest.

Due to the *GaTech* focus on anti-Asian hate speech, we analyzed our dataset through two additional text models to capture hate speech more broadly. Specifically, we relied on the *Detoxify*[17] and the *GateNLP*[18] models to assess the levels of toxicity across our topics. Both models measure toxicity across five categories, or dimensions: toxic, severely toxic, obscene, threat, insult and identity hate.[19] For each dimension, the models return a score from 0 to 1. The higher the score, the greater the probability is that a reader would perceive the tweet as

[15] For more information on *GaTech*, see He, Bing, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 90–94. Virtual Event Netherlands: ACM, 2021.
[16] The model achieved an average macro-F1 score of 0.832 on a training dataset of 3,355 tweets.
[17] More information is available on the Detoxify's Github repository.
[18] More information is available on the GateNLP's Github repository.
[19] For a definition of each category, please refer to the Perspective API documentation.

pertaining to the given dimension or category. To illustrate the useful nature of viewing toxicity through these different dimensions, here are some example tweets from our dataset:

Tweet with high insult and low identity hate:

> Dear idiotic fandom person. You are so many levels of hypocrisy that my brain cannot even compute it. You campaign for inclusion but are misogynistic and bigoted when it involves anything but your own ship. I actually loathe you. Thanks for coming to my Ted talk. Now off you fuck.

Tweet with high insult and high identity hate:

> Filipinos LIBERALS, WOKES or YELLOWS are all FUCKIN RACIST!!!!

Tweet with high threat and low insult:

> I hope they all die a slow painful death. Someone revives them and they die again.

In evaluating the results of the two toxicity models, it is important to consider that both the *Detoxify* and the *GateNLP* algorithms were originally trained on data collected from Wikipedia comments.[20] Therefore, these models might not be as effective in identifying and assessing toxicity levels on Twitter as they are on Wikipedia. However, when used in conjunction with the *GaTech* model, the aggregated patterns could provide meaningful insights.

As discussed in Section B.5, the three models struggled to differentiate hate speech from counterspeech due to the similar tones the two forms of speech share. For this reason, we conducted an additional qualitative analysis on a subset of tweets extracted from the three topics of interest. Besides highlighting the idiosyncrasies of each text model, our qualitative analysis points to a larger issue concerning the computational detection of hate speech and toxic content online. It shows the limits of machine learning when it comes to understanding the larger conflicts from which hate speech emerges.

Current AI models used across social media platforms for abuse detection likewise struggle to reliably identify and address problematic communications. This can lead to instances of discrimination against people within particular groups (defined, for example, by sex, gender, religion, etc.) whose vernaculars feature word and phrasing choices that happen to be targeted by moderation algorithms as sensitive or problematic.[21]

---

[20] For more on this, see the Kaggle Toxic Comment Classification Challenge dataset.
[21] On the limits of computational hate speech detection, see Sap, Maarten, et al. The risk of racial bias in hate speech detection. *ACL*. 2019.

B.4.2 Qualitative content analysis

For the qualitative analysis, we examined the top 100 tweets by hate speech likelihood, the top 100 tweets by counterspeech likelihood (both from the *GaTech* model) and the top 100 tweets by toxicity likelihood for each of the selected topics of anti-Asian hate, conspiracy theories and COVID-19. In addition, we examined a random sample of non-top 100 tweets.

Qualitative, manual analysis of the text within these subsets was then used to assess the accuracy of the *GaTech, GateNLP,* and *Detoxify* models when applied to our set of tweets, as well as to identify other trends within public conversation surrounding the three topics of interest. Such trends include more subtle forms of cultural frustrations, discrimination and tension not captured by the models.

Additionally, we manually examined a random sample of tweets from each of these categories from January 2019 through January 2020 to compare pre-pandemic tweets to tweets sent during COVID-19. We drew upon this as a comparison period for the *GaTech* classifier to determine the accuracy of this text model outside of its originally trained context (discussions during COVID-19). This comparison was particularly helpful in determining the accuracy of the *GaTech* model, which was originally set up to analyze discussions during COVID-19 in particular, and in so doing identifying limitations in the hate and counterspeech categorizations. It was necessary to check for these limitations to fairly compare the prevalence of hate speech in 2019 versus 2020-2021, in order to ensure that the model would not be biased towards the linguistic context of the pandemic (a phenomenon called "model over-fitting" in machine learning).

B.5 Findings

In this section, we look at how the three text models classified tweets as hate speech/counterspeech and organized them by the above-described categories of toxicity. Then, we present the findings of our close reading of a subset of tweets.

B.5.1 Quantitative analysis of anti-Asian hate speech and counterspeech

Overall, tweets in the anti-Asian topic increased from 288 tweets in 2019 to 2,238 in 2020. The tweets count declined in 2021 (843) but remained above pre-pandemic levels.[22] Please refer to Table 1 on page 47 for more details.

Figure 2 on the following page shows how the *GaTech* text model classified tweets in the anti-Asian hate topic as neutral (the bottom-left quadrant), hate speech (top-left quadrant) or counterspeech (bottom-right quadrant).

[22] The relatively small number of tweets composing the anti-Asian topic depends on the very specific list of keywords and combinations of keywords used to retrieve tweets from the initial dataset of 6 million tweets.

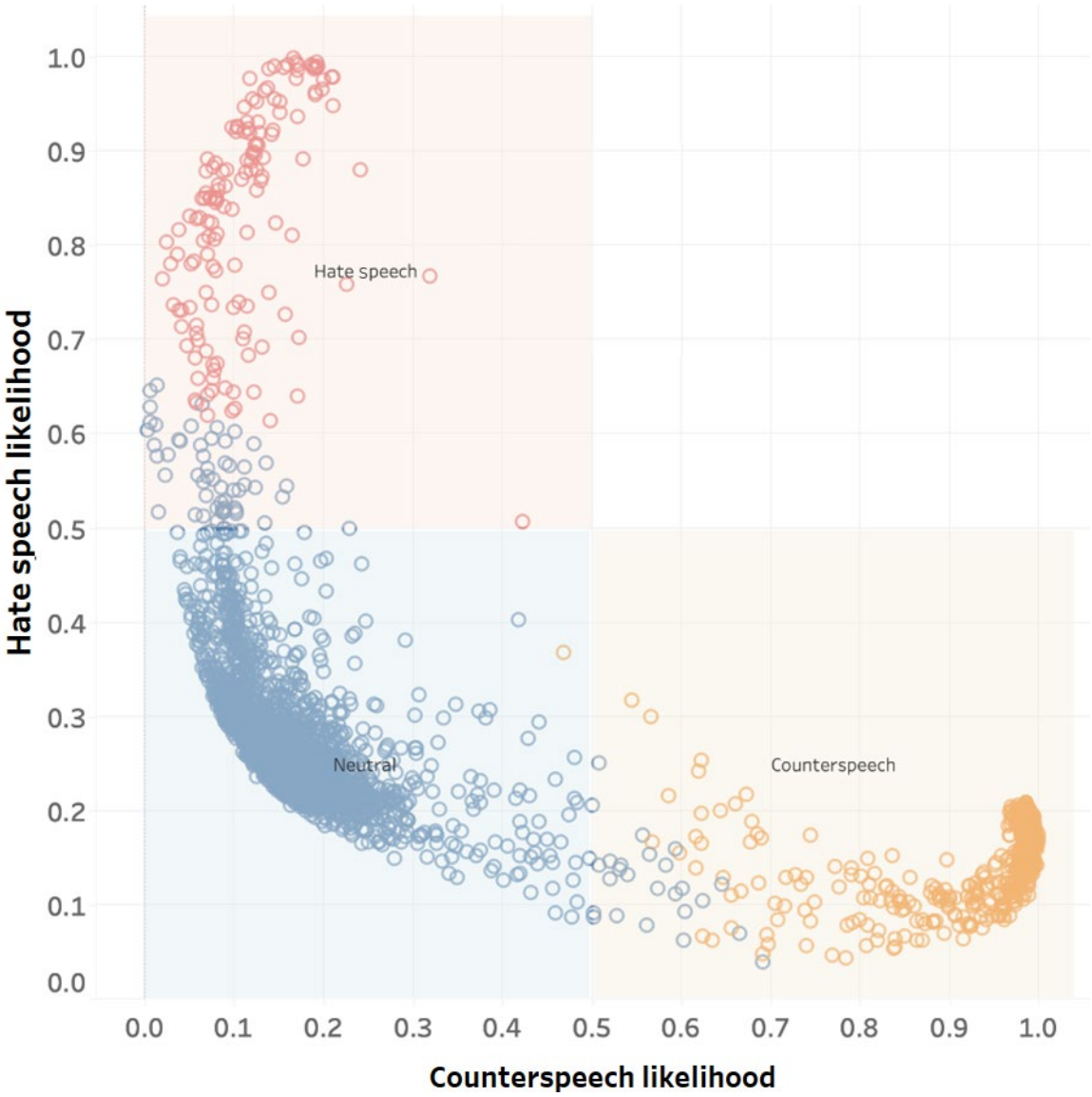


Figure 2: Counterspeech likelihood vs. hate speech likelihood in the anti-Asian hate topic

Although most of the tweets fall in the neutral category (79.3%), the clusters of tweets in the hate speech (16.3%) and counterspeech categories (4.4%) point to the extreme tones characterizing anti-Asian hate discourses on Twitter. The chart also reflects the *GaTech* model's ability to unambiguously interpret the content of the tweets and classify them accordingly. This is expected since the model was specifically trained to detect anti-Asian hate on Twitter.

The anti-Asian hate topic has changed over time, generally tied to current events. Both hate speech and counterspeech started to rise during the first few months of 2020, peaking at around March and April. This is when B.C. declared a provincial state of emergency and provincial and federal restrictions to travel and gatherings were imposed. Both anti-Asian hate speech and counterspeech declined in the following months.

Counterspeech increased dramatically again one year later, in March 2021. This is when the racially motivated shooting in two spas in Atlanta, Georgia (in the United States) left eight people dead, six of whom were of Asian descent. Figure 3 (below) shows how the tragic events of Atlanta impacted the B.C. discourse about anti-Asian hate, with counterspeech tweets peaking in March and reverberating throughout April 2021. The second spike of counterspeech in May 2021 corresponds to Asian Heritage Month.

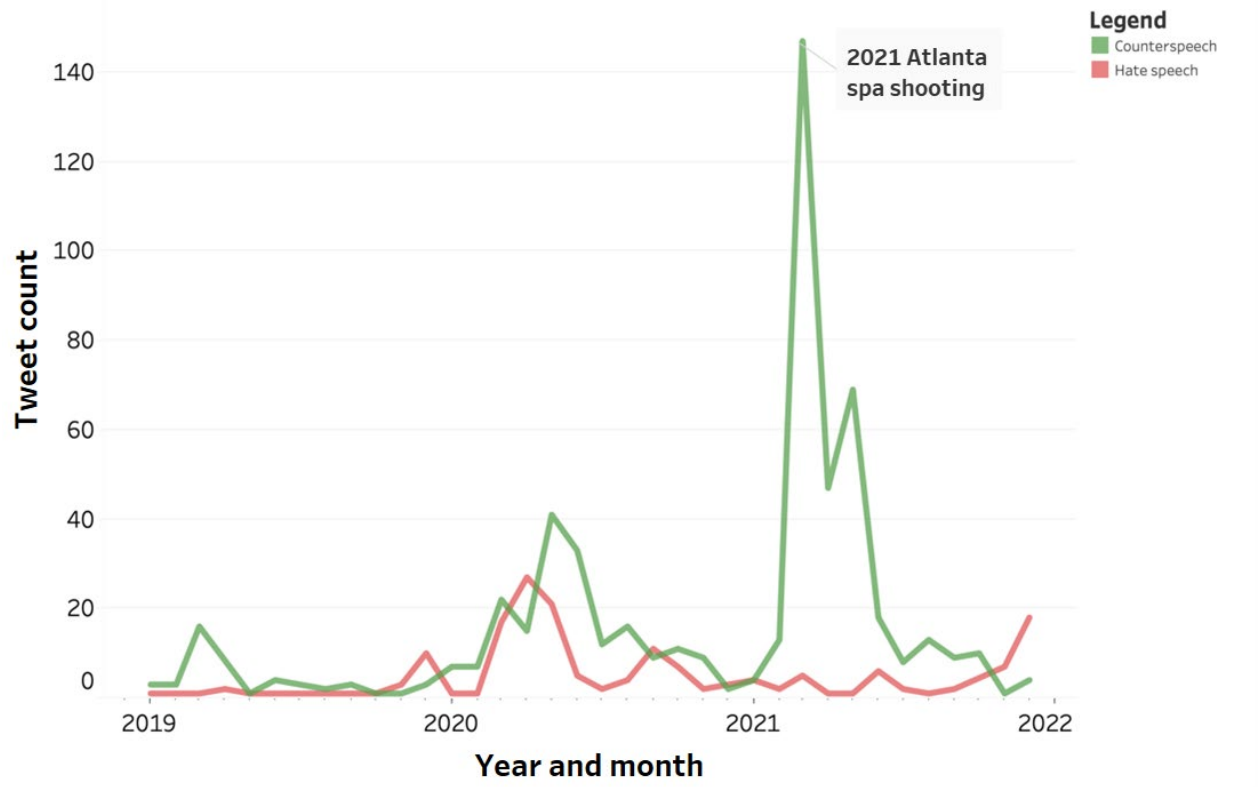


Figure 3: Anti-Asian hate speech and counterspeech timeline

This finding aligns with a trend highlighted in the report accompanying this case study: the influence of U.S. events on the Canadian Twittersphere.[23] This trend does not seem limited to the anti-Asian hate topic either. We analyzed our entire dataset of 6 million B.C. geolocated tweets using the already mentioned *GateNLP* and *Detoxify* text models for toxic content detection. The results show spikes in tweets classified as identity attacks, defined as "negative or hateful comments targeting someone because of their identity,"[24] when notable and contentious events occurred in the United States. Examples include the murder of a Black man, George Floyd, by police officers in 2020, the U.S. presidential elections in 2020 and the January 6, 2021 attack on the U.S. Capitol in response to the electoral defeat of President Donald Trump (see Figure 4 on the following page).

[23] Chris Tenove and Heidi Tworek. *Online Hate in the Pandemic*, 2022.
[24] "Attributes & Languages," Perspective. Accessed February 16, 2022.

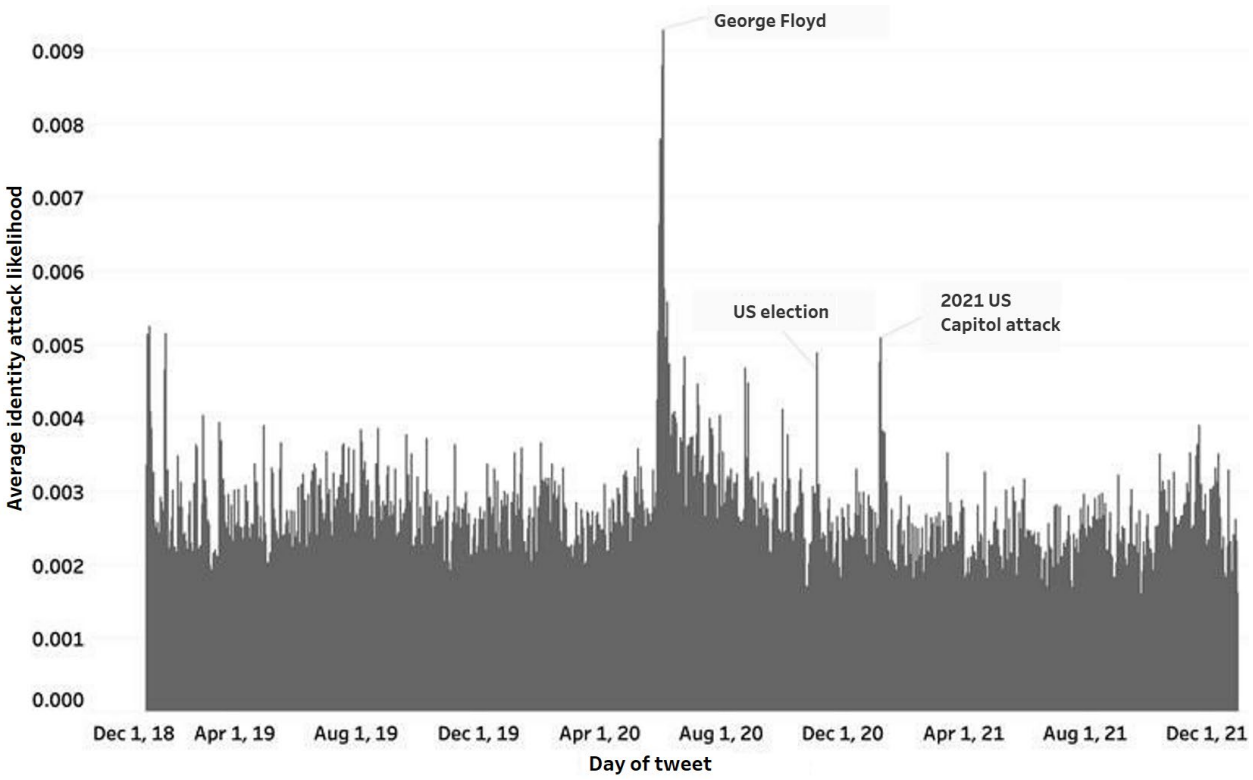


Figure 4: Average identity attack likelihood over time

B.5.2 Quantitative analysis of COVID-19 and conspiracy theories

We applied the same text models and techniques to analyze COVID-19 and conspiracy theories topics on Twitter. As expected, the *GaTech* model was not as effective in categorizing tweets in these two topics as it was in the anti-Asian hate topic. As shown in Figure 5 (on the following page), the data points appear much more spread out, indicating more uncertainty with the model's output. These results underline the importance of carefully evaluating the effectiveness of hate detection algorithms, especially when these are applied in contexts that differ from the ones they were trained on.

Looking at the evolution of these topics over time, we noticed an 860% increase in tweets (excluding retweets) classified as COVID-19 in 2020 compared to 2019 (4,195 in 2019 vs. 36,088 in 2020 – see Tables 2 and 3 for a breakdown of tweets by topic and category). This is expected due to the prominence of COVID-19 and related news (travel restrictions, vaccines, variants, etc.), throughout 2020. The number of tweets in this topic decreased slightly in 2021 (29,488 tweets, 19% decrease year over year). However, the share of tweets within this topic that were classified as toxic by the GateNLP and the Detoxify text models increased steadily, from 1.69% in 2019 to 3.79% in 2021. Similarly, the number of obscene tweets increased from 0.83% in 2019 to 2.01% in 2021.

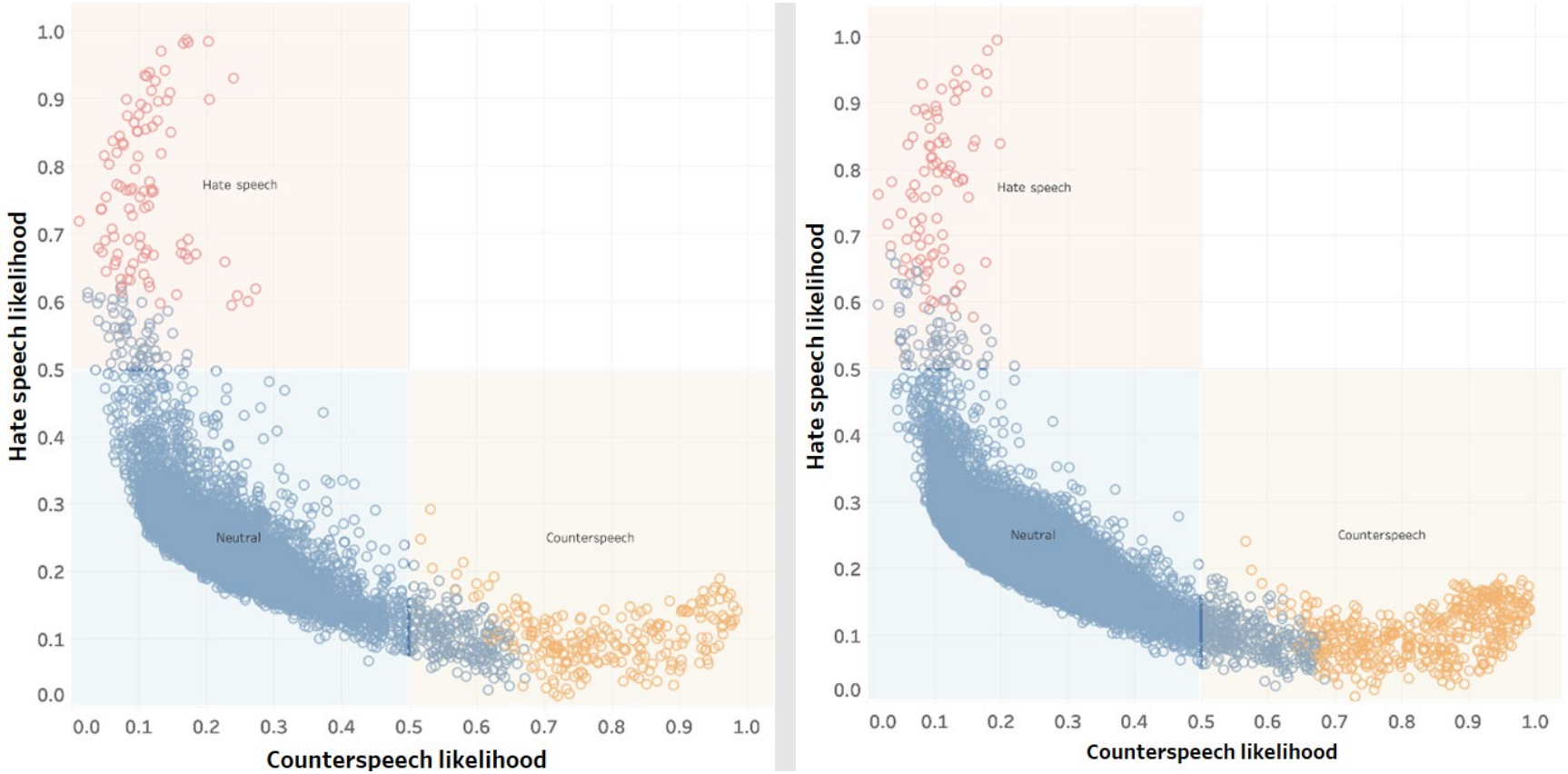


Figure 5: Counterspeech likelihood vs. hate speech likelihood in the conspiracy (left) and COVID-19 (right) topics

Table 1: Hate speech, counterspeech and toxicity dimensions in the anti-Asian hate topic

Year	Tweet count	% Hate speech	% Counter-speech	% Insults	% Identity attacks	% Obscene tweets	% Threats	% Toxic tweets	Hate speech	Counter-speech	Insults	Identity attacks	Obscene tweets	Threats
2019	288	7.29%	9.38%	0.69%	0.69%	2.08%	0.00%	3.13%	21	27	2	2	6	0
2020	2,238	3.93%	8.09%	1.34%	0.04%	2.19%	0.00%	4.65%	88	181	30	1	49	0
2021	843	4.74%	40.33%	1.07%	0.36%	1.66%	0.00%	2.85%	40	340	9	3	14	0

Table 2: Hate speech, counterspeech and toxicity dimensions in the COVID-19 topic

Year	Tweet count	% Hate speech	% Counter-speech	% Insults	% Identity attacks	% Obscene tweets	% Threats	% Toxic tweets	Hate speech	Counter-speech	Insults	Identity attacks	Obscene tweets	Threats
2019	4,195	0.12%	0.33%	0.19%	0.00%	0.83%	0.00%	1.69%	5	14	8	0	35	0
2020	36,088	0.16%	0.78%	0.70%	0.00%	1.96%	0.00%	3.76%	57	280	252	0	707	1
2021	29,488	0.10%	0.70%	0.65%	0.00%	2.01%	0.01%	3.79%	30	206	192	0	593	2

Table 3: Hate speech, counterspeech and toxicity dimensions in the conspiracy topic

Year	Tweet count	% Hate speech	% Counter-speech	% Insults	% Identity attacks	% Obscene tweets	% Threats	% Toxic tweets	Hate speech	Counter-speech	Insults	Identity attacks	Obscene Tweets	Threats
2019	4,093	0.37%	2.05%	2.08%	0.05%	4.18%	0.00%	9.68%	15	84	85	2	171	0
2020	5,398	1.39%	1.80%	1.70%	0.00%	3.83%	0.00%	9.61%	75	97	92	0	207	0
2021	3,008	0.70%	1.73%	1.43%	0.03%	3.52%	0.00%	7.28%	21	52	43	1	106	0

Conspiracy theory tweets increased only slightly in 2020 compared to 2019: +31% (4,093 tweets in 2019 vs 5,398 in 2020). As highlighted in the main report accompanying this case study, discussions about COVID-19 often involved conspiracy theories, for example about the role of China in the development and spread of the virus. The conspiracy topic, interestingly, had the highest proportion of toxic and obscene tweets, although the volume of tweets did not appear to shift as dramatically as with the other topics when comparing pre- versus COVID-19 pandemic years.

B.5.3 Qualitative analysis of toxicity, hate speech and counterspeech

Close reading of the COVID-19, anti-Asian hate and conspiracy topics showed several trends of interest. Analysis of the COVID-19 subset showed a general level of toxicity that reflects the heightened affective states of individuals in B.C. within the 2019–2021 period. Many of the tweets in this subset that contained high toxicity scores were expressions of various frustrations with the pandemic. These frustrations also reverberated through the conspiracy theories topic, which had the largest share of toxic tweets before and during the pandemic. Here, COVID-19 frustrations intersected with anti-Asian hate discourses, resulting in highly toxic and offensive tweets.

Within the COVID-19 topic, toxic content ranged from general animosity directed towards broader COVID-19 subjects themselves, such as the virus itself or quarantine, restrictions and lockdown orders, to resentment directed at political leaders or representatives, as well as other members of the public. The latter expressions vary but appear to stem from frustrations over people not following health orders, like wearing masks or social distancing, to condemnation over Canadians travelling to international locations and defying lockdown orders or restrictions. For example:

> At the vet and this woman is here to get her dog vaccinated, vaccinated for what? INTERNATIONAL TRAVEL. this woman is not only going to the Caribbean for Christmas, but she's bringing her DOG because she's not afraid of covid. I am gonna punch someone. (December 2020)

And finally, within the vaccine rollout period, another common pattern that we observed within the set of toxic tweets had to do with inoculation and the vaccine passports. On the subject of inoculation, we tended to see toxic remarks directed at the general public, whereas with vaccine passports we see them directed toward larger government responses.

> Hey @DrBonnieHenry, you @jjhorgan & @adriandix can all go fuck yourselves. Good luck with enforcement. Parallel society already going and there's not a damn thing you can do about it. I haven't seen any of you with the balls to enforce your totalitarian agenda yet. (August 2021)

Analysis of the anti-Asian hate subset revealed that tweets intersecting with anti-Asian hate and COVID-19 categories tended to have lower toxicity scores despite the fact that certain conspiratorial narratives and inflammatory rhetoric stemming from this intersection amplify hateful sentiments. For example, those tweets that direct accusations and acrimonious

dispositions toward the Chinese Communist Party (CCP) and the Chinese government tend to contain lower toxicity scores despite the ways these attacks can be used to advance discriminatory positions. Included in this category are tweets containing divisive rhetoric like "end trade with China" or assertions of CCP as a "criminal conspiracy" or "communist thugs."

B.5.4 Counterspeech before and during the pandemic

Analysis of the anti-Asian hate topic showed that tweets containing expletives or otherwise acrimonious responses of outrage or exasperation at hate incidents such as "fuck racism," "holy fuck," "what/why the fuck" were also flagged as toxic by classification models. Some of these examples also include insult-based responses that are antagonistic as well, such as "you racist piece of shit," "racist fucks," etc.[25] Though these responses could be considered forms of counterspeech, the algorithm identified them as toxic because of the language used, even if they also expressed opposition to racism. These examples also carry a notably different tenor than counterspeech that seeks to conversationally undermine hate such as the following example tweet from our dataset:

> This is hate speech as you well know, any use of "Canadians First" is a dog whistle to racists. We ruined other countries through weapons sales, mining, sending garbage & etc, the least we can do is help a bit (Canada was built by #refugees & #immigrants). #cdnpoli #bcpoli

However, when examining tweets within the *GaTech* data, including the top 100 counterspeech tweets, we found that tweets contained more generalized statements rather than conversational replies to individual users' statements. This is significant because the top 100 tweets are an indication of those tweets most likely to be classified as belonging to the counterspeech class. Examples of this include:

> We Canadians act all smug about racism. But today I learned that Vancouver alone reported more anti Asian hate crimes last year than the worst 10 US cities COMBINED. And history isn't any more kind. This shit HAS to stop. If you see it, speak up! #StopAsianHate

> [L]ead by example. But if your leaders and Police Force are closet racists, well that's a flawed system. Why trust them with your safety, let alone, your life!? It's time WE lead by example, so that WE can lead, and make a difference. #STOPASIANHATE

While this indicates that hate speech or racism are not going unopposed within this context, we find these insights introduce interesting questions about what tonalities and forms of counterspeech serve to replicate toxicity or exacerbate polarization rather than contribute to constructive dialogue. Or rather, what forms of counterspeech are truly *responsive* to hate speech rather than simply reactive? This is important because research has shown that typically within

[25] It should be noted that a good portion of these responses were directed at former President Donald Trump and other right-wing political representatives.

online exchanges the proportion of toxic speech to counterspeech, as well as the intensity and tonality of opinion within these settings, has a direct effect on how internet users take cues from one another and participate in conversations. Any comment sufficiently toxic or wielded as a stance of moral correctness can receive an equal and opposite reaction despite good intentions.[26] This forces conversation into a polarized stalemate which can sharpen discriminatory positions rather than defuse or effectively undermine discrimination.

B.6 Conclusion

In summary, our investigation found tweets under the anti-Asian hate topic surging in 2020, with a high concentration of counterspeech in 2021. The COVID-19 topic had a high volume of toxic tweets, but a comparably smaller proportion were classified as either hate speech or counterspeech by the *GaTech* model. Tweets under the conspiracy topic had the highest concentration of toxic tweets, but again a smaller proportion of hate speech or counterspeech than the anti-Asian topic. Preliminary findings presented in this case study seem to suggest a strong influence of U.S. events, Twitter users and contents over the B.C. Twittersphere.

The investigation also highlighted the challenges and limitations of analyzing public online communications. In addition to sampling issues highlighted in Section B.3.1, we found it difficult to estimate how many harmful tweets had already been deleted by Twitter when we first built our dataset. These technical limitations should be taken into consideration when generalizing some of the findings reported in this case study.

Our findings also align with previous work showing how algorithmic methods for detecting hate speech do not fully capture and contextualize ever-evolving online public discourse. As illustrated in Section B.5.3, we found the line between counterspeech and toxic speech to be blurred in certain cases. This means that any automated content moderation by platforms may inadvertently silence voices speaking up against hate and racism.

[26] For a comprehensive list of such research examples, see: Rösner, Leonie, Stephan Winter, and Nicole C. Krämer. Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior 58* (2016): 461–470; Seering, Joseph, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on twitch through moderation and example-setting. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017; Schieb, Carla, and Mike Preuss. Governing hate speech by means of counterspeech on Facebook. *66th ica annual conference*, at Fukuoka, Japan. 2016.