2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis

Bing He[1], Caleb Ziems[1], Sandeep Soni[1], Naren Ramakrishnan[2], Diyi Yang[1], Srijan Kumar[1]
[1] Georgia Institute of Technology, [2] Virginia Tech
[1]{bhe46, cziems, sandeepsoni, diyi.yang, srijan}@gatech.edu, [2] naren@cs.vt.edu

***Abstract*—The spread of COVID-19 has sparked racism and hate on social media targeted towards Asian communities. However, little is known about how racial hate spreads during a pandemic and the role of counterspeech in mitigating this spread. In this work, we study the evolution and spread of anti-Asian hate speech through the lens of Twitter. We create COVID-HATE, the largest dataset of anti-Asian hate and counterspeech spanning 14 months, containing over 206 million tweets, and a social network with over 127 million nodes. By creating a novel hand-labeled dataset of 3,355 tweets, we train a text classifier to identify hateful and counterspeech tweets that achieves an average macro-F1 score of 0.832. Using this dataset, we conduct longitudinal analysis of tweets and users. Analysis of the social network reveals that hateful and counterspeech users interact and engage extensively with one another, instead of living in isolated polarized communities. We find that nodes were highly likely to become hateful after being exposed to hateful content in the year 2020. Notably, counterspeech messages discourage users from turning hateful, potentially suggesting a solution to curb hate on web and social media platforms. Data and code is available at http://claws.cc.gatech.edu/covid.**

I. Introduction

Hateful incidents throughout the world, such as acts of microaggression, physical and verbal abuse, and online harassment have increased during the COVID-19 pandemic [1]. Following the identified origin of COVID-19 in China, racially motivated hate crime incidents have increasingly targeted the Chinese and the broader Asian communities, resulting in over 6,603 racially-motivated hateful incidents in a year [2].

While there is mounting evidence of offline discriminatory acts and racism during COVID-19, the extent of such overtly hateful content on the web and social media is not widely known. Online hate speech has severe negative impact on the victims [3] and can lead to real-world crimes [4]. Meanwhile, while efforts to educate about, curb, and counter hate have been made via social media campaigns (e.g. the #RacismIsAVirus campaign), the success, effectiveness, and reach of counterspeech messages remain unclear. Thus, it is crucial to detect online hate speech to curb both online and physical harm, and monitor counterspeech messages to quantify their effectiveness, and inform future strategies to counter hate.

Recent research has been conducted on COVID-19-related hate online posts against Asians [5]–[10]. Building on these concurrent research works, we contribute several novel aspects to the understanding of this phenomenon. First, we conduct a long-term longitudinal study of the hate and counterspeech ecosystem on Twitter to monitor the changes in social perception and stance towards the Asian community as the pandemic progressed. Second, we study the combined ecosystem of hate and counterspeech messages on Twitter, as opposed to studying them in isolation. This is important because both coexist on the platform and influence each other simultaneously. Studying only one type of message (hate or counterspeech) is unable to uncover the influence they have on each other.

Our contributions. In this paper, we present COVID-HATE, the largest dataset of anti-Asian hate and counterspeech on Twitter in the context of the COVID-19 pandemic, along with a 14 month-long longitudinal analysis of the Twittersphere. We make the following key contributions:

- We create a dataset of COVID-19-related tweets, containing over 206 million tweets made between January 15, 2020 and March 26, 2021, and the social network of users, having over 127 million nodes and 910 million edges. The data and code is available at http://claws.cc.gatech.edu/covid.
- We hand-annotate 3,355 tweets based on their hatefulness towards Asians as hate, counterspeech, or neutral tweets to build highly accurate text classifier to identify hate and counterspeech tweets, finally identifying 1,227,116 hate and 1,154,289 counterspeech tweets.
- We conduct statistical, linguistic, and network analysis of tweets and users to reveal characteristic patterns of hate and counterspeech, and find counterspeech tweets lower the probability of neighboring nodes becoming hateful.

II. COVID-HATE: An Anti-Asian Hate and Counterspeech Dataset during COVID-19

We describe the COVID-HATE dataset. Table I shows the data statistics.

ASONAM '21, November 8-11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11…$15.00
https://doi.org/10.1145/3487351.3488324