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, but not in the year 2021. Notably, counterspeech messages discourage users from turning hateful, potentially suggesting a solution to curb hate on web and social media platforms.**

Introduction

The global outbreak of coronavirus disease 2019 or COVID-19 caused widespread disruption in people's lives. Following the identified origin of COVID-19 in China, racially motivated hate crime incidents have increasingly targeted the Chinese and the broader Asian communities. The attacks in Atlanta, Georgia on March 16, 2021, which led to the death of six Asian women, show the grim reality of racial hate [1].

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, especially their longitudinal pattern. 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 [2, 3]. 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.

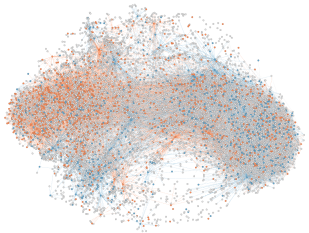


Fig. 1: The COVID-HATE social network with hate nodes (orange), counterspeech nodes (blue), and neutral nodes (gray).

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 are available on http://claws.cc.gatech.edu/covid.
- We 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. A subgraph of user social network with nodes annotated with hate, counterspeech, and neutral labels is shown in Figure 1.
- 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. This effect is more pronounced in 2021 than 2020.

Property	Statistic
Duration	Jan 15, 2020–Mar 26, 2021
Number of tweets	206,348,565
Number of (frac.) hate tweets	1,337,116 (0.64%)
Number of (frac.) counterspeech tweets	1,154,289 (0.55%)
Number of (frac.) neutral tweets	203,857,160 (98.81%)
Number of users	23,895,911
Number of (frac.) hate users	697,098 (2.91%)
Number of (frac.) counterspeech users	629,029 (2.63%)
Number of (frac.) neutral users	22,477,616 (94.06%)
Number of nodes in the social network	127,831,666
Number of edges in the social network	910,630,334

TABLE I: Statistics of `COVID-HATE` dataset, containing anti-Asian hate and counterspeech tweets and social network in the context of COVID-19.

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

In this section, we describe `COVID-HATE`, a Twitter dataset containing COVID-19 anti-Asian hate and counterspeech tweets and social network. Table I shows the data statistics.

Tweet Dataset

We adopted a keyword-based approach to collect relevant COVID-19 tweets through Twitter's official APIs. Specifically, we used a collection of keywords and hashtags belonging to three sets: (a) `covid-19 keywords` are terms referring to COVID-19 which are used to collect tweets related to the pandemic, (b) `hate keywords` are keywords and hashtags indicating anti-Asian hate amidst COVID-19. To compile this list, we first took the hate keywords from existing papers and news articles [4]. We then expanded this list by including co-occurring hate hashtags observed in an initial tweet crawl. We also included Asian slurs listed in Hatebase.[1] Finally, (c) `counterspeech keywords` are keywords and hashtags that were used to organize efforts to counter hate speech and support Asians. These keywords were listed in news articles covering counterspeech efforts during the initial phases of the data collection setup [5]. In total, we used 42 keywords as shown in Table II. During the process, we intentionally created a broad list of keywords to ensure high recall. This may result in collection of borderline-relevant tweets as well, which can later be identified and removed in the filtering step via a classifier, which we describe later. After getting the keywords, we utilized Twitter's Streaming API and Twitter's Search API to collect the data. Finally, we collected 206,348,565 English-language tweets made by 23,895,911 users between January 15, 2020 and March 26, 2021, which do not contain retweets

Twitter Network Construction: In addition to the tweets, we crawled the ego-network (i.e., the followers and followees) of a randomly-sampled subset of users who made at least one COVID-19 tweet by Twitter's GET API, as shown in Tab. I.

Annotating Anti-Asian COVID-19 Hate and Counterspeech

To identify tweets relevant to our study of hate and counterspeech, we hand-label a subset of tweets and create a textual classifier to label the rest. Even though tweets may have explicitly hateful hashtags, categorizing tweets simply based on the presence (or absence) of a hashtag and keyword is insufficient because hashtags can be added to gain visibility and promote tweets. Conversely, a tweet can be hateful even without having a hateful hashtag. The same is true for counterspeech tweets. Thus, we developed a rigorous annotation process to establish the ground truth categories of tweets based on the tweet content.

[1]https://hatebase.org/

Category	Keywords
COVID-19	coronavirus, covid 19, covid-19, covid19, corona virus
Hate keywords	#CCPVirus, #ChinaDidThis, #ChinaLiedPeopleDied, #ChinaVirus, #ChineseVirus, chinese virus, #ChineseBioterrorism, #FuckChina, #KungFlu, #MakeChinaPay, #wuhanflu, #wuhanvirus, wuhan virus, chink, chinky, chonky, churka, cina, cokin, communistvirus, coolie, dink, niakoué, pastel de flango, slant, slant eye, slopehead, ting tong, yokel
Counterspeech keywords	#IAmNotAVirus, #WashTheHate, #RacismIsAVirus, #IAmNotCovid19, #BeCool2Asians, #StopAAPIHate, #ActToChange, #HateIsAVirus

TABLE II: The list of keywords and hashtags used for comprehensive data collection.

We labeled the tweets into the following three broad categories, as we define below.

Anti-Asian COVID-19 Hate Tweets: We build on previous studies of racial hate literature to define anti-Asian hate [3]. Building on this, we define anti-Asian COVID-19 hate as antagonistic speech that is directed towards an Asian entity (individual person, organization, or country), and *others* the Asian outgroup through intentional opposition or hostility in the context of COVID-19. One overt example of anti-Asian hate we considered is (censorship ours):

> *F*ck Chinese scums of the Earth disgusting pieces of sh*t learn how to not kill off your whole population of pigs, chickens, and humans. coronavirus #wuhanflu #ccp #africaswine #pigs #chickenflu nasty nasty China clean your f*****g country.*

COVID-19 Counterspeech Tweets: This category of COVID-19-related tweets either: (a) explicitly identify, call out, criticize, condemn, challenge, or oppose racism, hate, or violence towards an Asian entity or (b) explicitly support, express solidarity towards, or defend an Asian entity. These tweets can either be direct replies to hateful tweets or be stand-alone tweets, but they must be explicit. An example of a tweet in this category is as follows:

> *The virus did inherently come from China but you can't just call it the Chinese virus because that's racist. or KungFlu because 1. It's not a f*****g flu it is a Coronavirus which is a type of virus. And 2. That's also racist.*

Neutral and Irrelevant Tweets: These tweets neither explicitly nor implicitly convey hate, nor counterspeech, but are related to COVID-19. Tweets in this category also include news, advertisements, or outright spam. One example of a tweet in this category is:

> *COVID-19: #WhiteHouse Asks Congress For $2.5 Bn To Fight #Coronavirus: Reports #worldpowers #climatesecurity #disobedientdss #senate #politics #news*

Feature set	Precision	Recall	F1 score
Anti-Asian hate tweet detection			
Linguistic	0.541	0.233	0.323
Hashtag	0.100	0.002	0.005
BERT	0.765	0.760	0.762
Counterspeech tweet detection			
Linguistic	0.483	0.189	0.267
Hashtag	0.800	0.029	0.056
BERT	0.839	0.868	0.853
Neutral tweet detection			
Linguistic	0.632	0.891	0.739
Hashtag	0.591	0.999	0.743
BERT	0.886	0.874	0.880

TABLE III: Tweet classification performance of different feature sets with a neural network classifier. The BERT model has the best classification performance in all three tasks.

#unsc #breaking #breakingnews #wuhan #wuhanvirus https://t.co/XipNDc

Annotation process: We trained two undergraduate annotators to recognize anti-Asian COVID-19 hate tweets, COVID-19 counterspeech tweets, and neutral/irrelevant tweets using the above definitions. Both annotators are of Asian descent (one Chinese and one Indian). One co-author supervised the annotation process. After practicing on a set of 100 tweets and discussing disagreements with the supervising co-author, the annotators each independently labeled the same set of 3,255 tweets, which were randomly sampled from the collected dataset. Since the majority of tweets were expected to be neutral, we over-sampled tweets that contained anti-Asian hate, and counterspeech terms. This ensured our labeling process yielded sufficient hate and counterspeech tweets to train a classifier. The annotation process took six weeks.

The two annotators agreed on 68% of the data, with Cohen's Kappa score of 0.448 for hate and 0.590 for counterspeech, indicating a moderate inter-rater agreement that is typical of hate speech annotation [3, 6]. We removed the tweets where the two annotators disagreed and were left with 429 hate, 517 counterspeech, and 1,344 neutral tweets. The annotators also identified 110 tweets containing hatefulness or aggression towards non-Asian groups. Since our goal is to study anti-Asian COVID-19 hate, we drop the latter set of tweets. We focus only on anti-Asian hate, counterspeech, and neutral tweets in the remainder of this paper.

Anti-Asian Hate and Counterspeech Text Classifier

We use the annotated tweets to train a text-based machine learning classifier to label tweets as anti-Asian hate, counterspeech, or neutral by three features separately: 1) **Linguistic Features.** This set contains a total of 90 features including stylistic and psycholinguistic patterns [7]. 2) **Hashtag features.** These features represent the number of occurrences of each hashtag and keyword listed in Table II. 3) **Bert Tweet Embeddings.** We embed each tweet using the BERT base uncased text embedding model and use a feed-forward layer for classification [8].

Model training. Similar to the BERT classifier, one-layer feed-forward neural network classifiers are trained using linguistic features and hashtag features. We conducted five-fold cross validation and reported the performance in Table III, finding BERT has the superior performance. Thus, we use the BERT model to label the rest of the tweets in the dataset for downstream analysis.

Longitudinal Characterization of COVID-19 Hate and Counterspeech

In this section, we use the COVID-HATE dataset to analyze the patterns of hate and counterspeech in the Twitter ecosystem. We focus our analysis on the evolution and spread of hate and counterspeech and the characteristics of the users. To characterize the temporal changes in trends, we will compare the statistics from the year 2020 (from January 15, 2020 to December 31, 2020) and the year 2021 (from January 1, 2021 to March 26, 2021).

The Ebb and Flow of Hate and Counterspeech

Here we consider the longitudinal spread of hate and counterspeech tweets in the Twitter ecosystem.

Hate tweets were more frequent than counterspeech tweets in the year 2020. Figure 3 shows the daily distribution of hate and counterspeech tweets. First, we note that hate tweets outnumber counterspeech tweets throughout the timeline during 2020. Next, the number of hate and counterspeech tweets was negligible-to-low during the early phases of the pandemic in January, 2020 and February, 2020. We observe the spike in hate speech between March 16, 2020 and March 19, 2020. These spikes appear to closely follow President Trump's use of the phrase "Chinese Virus" in his tweet on March 16.[2] There were several major spikes in daily hate tweets throughout 2020, which exceeded the daily counterspeech tweets.

Counterspeech tweets increased dramatically after the 2021 Atlanta shooting. Counterspeech messages typically had lower volume throughout 2020 compared to hate tweets. However, after the Atlanta Spa shooting on March 16, 2021 [1], there was a dramatic increase in the number of counterspeech tweets in March, 2021. Counterspeech tweets increased by 401.2% within one week, while we observed that hateful tweets also surprisingly rose by 17.9%. The spike in counterspeech signals the Twittersphere expressing sympathy and solidarity towards the Asian community.

Please note that even though the keywords and hashtags used for data collection were selected during the early phase of the pandemic (March 2020), the dataset reveals spikes in hate and counterspeech throughout the 14 month period.

User Activity and Interaction Behavior

We analyze the properties of the users who produce hate and counterspeech tweets. Following the tweet categorization labels, we categorize users, based on their tweets, into one of the following: *hate*, *counterspeech*, *dual*, or *neutral* Hate users make at least one hate tweet but no counterspeech tweets. Similarly, counterspeech users make at least one counterspeech tweet but no hate tweet. Users who tweet from both categories

[2]https://twitter.com/realDonaldTrump/status/1239685852093169664

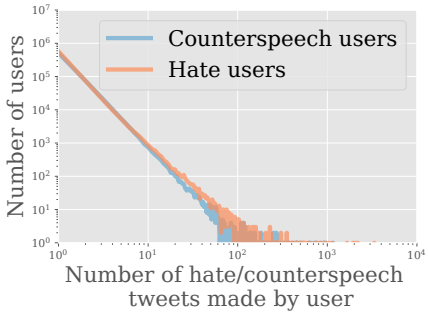


Fig. 2: Distribution of the number of hate and counterspeech tweets made by users shows a long tail pattern.

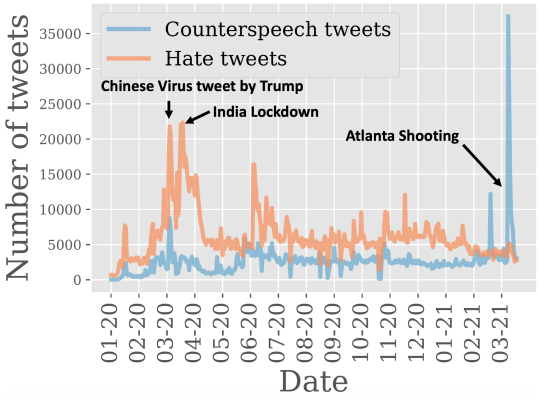


Fig. 3: The number of hate and counterspeech tweets from January 15, 2020–March 26, 2021.

are categorized as dual users. Finally, users who make at least one COVID-19 tweet (and thus, are part of our dataset), but no hate or counterspeech tweets, are labeled as neutral. Among the 23,895,911 users in the dataset, most of the users (94.06%) are neutral, 697,098 (2.92%) are hateful, 629,029 (2.63%) are counterspeech users, and a very small fraction of users (0.39%) are dual. This distribution mimics the category-wise tweet distribution. Our following analysis focuses on hate, counterspeech, and neutral user categories. Due to low volume, we do not emphasize on the dual users, which can be worth exploring in future studies.

Figure 2 shows the distribution of the number of hate tweets (counterspeech tweets, respectively) made by hate users (counterspeech users, respectively). We observe that both distributions exhibit a long tail, showing that most users make few relevant tweets and only a handful of users are responsible for spreading most of the hate propaganda and counterspeech messages.

Social Network Connectivity Structure

In this section, we examine the user-user social connectivity in the hate and counterspeech ecosystem. As described in the dataset section, we crawled the social network containing over 127 million nodes and 910 million edges. Out of these, 1,380,613 nodes have made at least one COVID-19-related tweet. The rest of the nodes are part of the network as they are neighbors of these nodes. Figure 1 shows a subgraph of this network, with nodes colored according to their category (hate, counterspeech, or neutral).

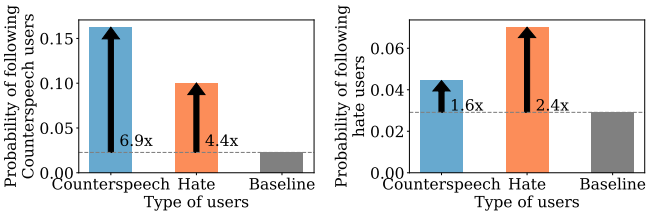


Fig. 4: Social network of hate and counterspeech users: Hate and counterspeech users are highly interconnected and exhibit homophily.

To understand the differences in how hate and counterspeech users behave, we compare their ego-networks. We find that on average, counterspeech users are better connected than hate users—counterspeech users follow more users compared to hate users (1201.84 vs. 828.40; $p < 0.001$) and are followed more by other users (1249.42 vs. 759.96; $p < 0.001$).

Intragroup and intergroup connectivity. We analyze the connectivity of users within and across the different groups to establish if nodes express homophily or form echo chambers. Simply comparing their probability of creating edges to nodes of a certain group is not sufficient as it is confounded by the node degrees and node distribution across categories. Thus, we create a network baseline preserving the node property to model the expected behavior of nodes and compare against this baseline.

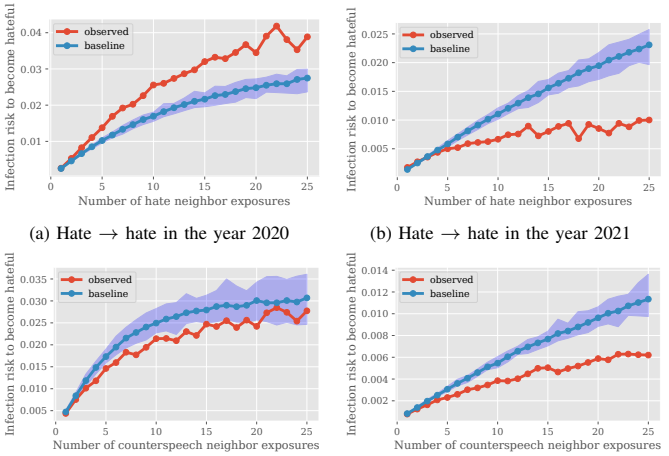
We compare the observed and the baseline behavior using the probability of connecting to hate, counterspeech, and neutral nodes. Figure 4 presents the results.

Nodes exhibit homophily. First, we examine the propensity for hate and counterspeech nodes to connect with nodes within their own group. In Figure 4 (left), we show that counterspeech users are 6.92× more likely to connect to other counterspeech users compared to the baseline behavior. Similarly, the right figure shows that hateful users connect with other hateful users 2.42× more than expectation. Thus, nodes are preferentially connected to other nodes in the same group.

Do hateful and counterspeech users form polarized communities? Echo chambers and polarization are commonly-observed phenomena in social media, which are responsible for the spread of propaganda and misinformation. However, it is not known whether echo-chambers exist in the hate network too. Given that nodes preferentially connect to similar nodes, four scenarios are possible. (1) Hate and counterspeech users live in isolated echo-chambers, where these groups do not interact with one another. (2) On the other extreme, the two groups interact highly with each other, possibly exhibiting conflict. The remaining two possibilities are that the out-group connections are one-sided.

Figure 4 illustrates the empirically-observed behavior. Both hate and counterspeech nodes are more likely to connect with one another than expected. Precisely, hateful users follow counterspeech users 4.45× more than expected (left figure) and counterspeech users are 1.62× more likely to follow hateful users compared to the baseline (right figure).

Altogether, these indicate that hateful and counterspeech



(a) Hate → hate in the year 2020

(b) Hate → hate in the year 2021

(c) Counterspeech → hate in the year 2020

(d) Counterspeech → hate in the year 2021

Fig. 5: The impact of hate speech and counterspeech on the spread of hate in year 2020 (left) and year 2021 (right).

users are highly engaged and closely interact with each other.

Influence of Counterspeech on the Spread of Hate

In this section, we investigate the within-group and across-group influence on the diffusion of hate messages. Specifically, we quantify influence as the likelihood of a user to become hateful (i.e., writing an anti-Asian hate tweet for the first time) after a user is exposed to any number of hate or counterspeech tweets from his or her neighbors. Following the techinques by [9], we get ther result in Figure 5. As the contagion patterns can change over time due to changes in social dynamics, we separately analyse the patterns in year 2020 (when hate exceeded counterspeech) and in year 2021 (when counterspeech overpowered hate).

Figure 5(a) shows that exposure to hate speech increased the likelihood of adopting hate speech, compared to the baseline in year 2020. Moreover, the likelihood of hate adoption increased with the number of exposures. The pattern changed in year 2021, as shown in Figure 5(b), when hate speech became less contagious than baseline. This is likely due to support towards the Asian community after the Atlanta shooting in 2021.

Furthermore, in the year 2021, counterspeech significantly deterred the spread of hate speech compared to the baseline, as shown in Figure 5(d). This is stronger compared to the pattern in the year 2020 (Figure 5(c)) when counterspeech's effect on hate speech was slightly lower than the baseline, thus showing low social inhibition effect. The change shows a positive trend towards counterspeech potentially mitigating hate speech.

Related Work

Due to the long-lasting societal effect of COVID-19 pandemic and infodemic, some researchers study hate speech [7] analyzed its pattern in the context of COVID-19 [2]. But, counterspeech is ignored in those research, which is the gap we address. Meanwhile, while there are some works regarding counterspeech [10], they are quite generic and not placed in a pandemic. Furthermore, contemporaneous work by [3] released a large hand-labeled dataset of hatespeech and counterspeech. However, they do not conduct any analysis of the hate and counterspeech Twittersphere, which we present in this work, in addition to creating a complementary hand-labeled dataset. More importantly, we present longitudinal analysis of tweets and users to give a comprehensive view of hate speech and counterspeech.

Conclusions

Our findings shed light on societal problems caused by the COVID-19 pandemic. Notably, we observe that counterspeech reduced the probability of neighbors becoming hateful. It paves the way towards the use of public counterspeech messaging campaigns as a potential solution against hate speech.

Acknowledgements

This research is supported in part by NSF (Expeditions CCF-1918770, NRT DGE-1545362, IIS-2027689), Adobe, Facebook, Microsoft, Georgia Institute of Technology, Russell Sage Foundation and the Institute for Data Engineering and Science (IDEAS) at Georgia Tech.

References

[1] New York Times, "8 dead in atlanta spa shootings, with fears of anti-asian bias," 2021, [Online; accessed 14-May-2021]. [Online]. Available: https://www.nytimes.com/live/2021/03/17/us/shooting-atlanta-acworth

[2] N. Vishwamitra, R. R. Hu, F. Luo, L. Cheng, M. Costello, and Y. Yang, "On analyzing covid-19-related hate speech using bert attention," in *ICMLA*. IEEE, 2020.

[3] B. Vidgen, A. Botelho, D. Broniatowski, E. Guest, M. Hall, H. Margetts, R. Tromble, Z. Waseem, and S. Hale, "Detecting east asian prejudice on social media," *arXiv preprint arXiv: 2005.03909*, 2020.

[4] E. Chen, K. Lerman, and E. Ferrara, "Covid-19: The first public coronavirus twitter dataset," *arXiv preprint arXiv:2003.07372*, 2020.

[5] Steve Barrett, PR Week, "Racism is a virus, not asians: stopaapihate," 2020, [Online; accessed 14-May-2021]. [Online]. Available: https://www.prweek.com/article/1711232/racism-virus-not-asians-stopaapihate

[6] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," *arXiv preprint arXiv:1701.08118*, 2017.

[7] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM CSUR*, vol. 51, no. 4, pp. 1–30, 2018.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *ACM SIGKDD*, 2008.

[10] B. Mathew, N. Kumar, P. Goyal, A. Mukherjee *et al.*, "Analyzing the hate and counter speech accounts on twitter," *arXiv preprint arXiv:1812.02712*, 2018.