Multiplex Anti-Asian Sentiment before and during the Pandemic: Introducing New Datasets from Twitter Mining

Hao Lin[1], Pradeep Nalluri[1], Lantian Li[2], Yifan Sun[1], Yongjun Zhang[1]
[1]Stony Brook University
[2] Northwestern University
[1]{hao.lin, pradeepkumar.nalluri}@stonybrook.edu
[1]{yifan.sun,yongjun.zhang}@stonybrook.edu
[2]lantianli2014@u.northwestern.edu

Abstract

COVID-19 has disproportionately threatened minority communities in the U.S, not only in health but also in societal impact. However, social scientists and policymakers lack critical data to capture the dynamics of the anti-Asian hate trend and to evaluate its scale and scope. We introduce new datasets from Twitter related to anti-Asian hate sentiment before and during the pandemic. Relying on Twitter's academic API, we retrieve hateful and counter-hate tweets from the Twitter Historical Database. To build contextual understanding and collect related racial cues, we also collect instances of heated arguments, often political, but not necessarily hateful, discussing Chinese issues. We then use the state-of-the-art hate speech classifiers to discern whether these tweets express hatred. These datasets can be used to study hate speech, general anti-Asian or Chinese sentiment, and hate linguistics by social scientists as well as to evaluate and build hate speech or sentiment analysis classifiers by computational scholars.

Keywords: Hate speech, Sinophobia, COVID19, Anti-Asian, Anti-China, Twitter mining

1 Introduction

The COVID-19 pandemic has disproportionately threatened minority communities in the U.S. In particular, COVID-19 has brought sinophobia to the surface (Croucher et al., 2020; Zhang, 2021; Horton, 2020). Since the outbreak of COVID-19, there were over 4,000 hate incidents such as harassment and physical attacks reported to stopaapihate.org. The growing anti-Asian attacks have led to the recent passage of the anti-Asian Hate Crimes Bill by the U.S. House after the mass shootings in Atlanta. Despite the problematic surge in COVID-hate incidents and crimes targeting Asian American and Pacific Islander (AAPI) communities, social scientists and policymakers lack critical data and quantitative measures to capture the evolution of anti-AAPI trends in the U.S., and cannot evaluate the scale and scope of anti-AAPI hate incidents in the pandemic.

Recent scholars have used social media data with machine learning techniques to track online anti-Asian hate speech (Vidgen et al., 2020; Ziems et al., 2020; Jiang et al., 2020). For instance, Ziems et al. (2020) examined the evolution and spread of anti-Asian hate speech from 30 million tweets collected between January 15 and April 17, 2020. Cook et al. (2021) classified over 297 million tweets about China or COVID-19 between January 2017 and June 2020 by using a BERT model trained on 5000 labeled tweets and found that the awareness of COVID-19 has led to a sharp rise in anti-China sentiments in the U.S. Although these studies provide training datasets to build hate speech classifiers and have insights about the spread of anti-Asian hate at the early stage of the outbreak, little is known about the enduring evolution of anti-Asian hate or counter hate before and during the continuing pandemic.

In this paper, we report trends and patterns of anti-Asian sentiments and hate speech on Twitter by introducing new datasets. Twitter has been one of the most salient battlegrounds of both propagating and fighting against misinformation, fake news, hatred, and xenophobia during the COVID-19 pandemic. We use computational tools with natural language processing and machine learning methods to detect hate speech on Twitter before and during the pandemic. Our datasets contain 68.38 million tweets, and they fall in four categories:

- COVID-related anti-AAPI tweets, which are collected by using Covid-related keywords such as *'chinavirus'* and *'kung-flu'*
- Non-COVID-related hateful tweets, which are collected by using general Anti-AAPI key-

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 16 - 24
May 26, 2022 ©2022 Association for Computational Linguistics

words such as *'ching chong'* and *'chink'*

- Discussions that concern Chinese politics; The topics per se may not be hateful, but they often provoke hateful tweets such as discussions about Uyghers, Hong Kong protests, and Xi Jinping

- Counter-hate tweets, including keywords such as *'stopasianhate'* and *'racismisvirus'*

These datasets provide a comprehensive portrait of the dynamics of anti-Asian hate sentiments spanning from 2007 to 2021. Thus, we are able to address important questions related to the evolution of anti-Asian hate sentiments over time.

2 Background

In the past decade, computational scholars have made great efforts to detect hate speech on social media platforms (Davidson et al., 2017; Warner and Hirschberg, 2012; Del Vigna12 et al., 2017). Although there is no clear and formal definition of hate speech, scholars tend to define hate speech as language or speech "used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Davidson et al., 2017). Anti-Asian hate speech, especially Sinophobia, has attracted numerous attention from both computer scientists and social scientists (Vidgen et al., 2020; Cook et al., 2021; Lee, 2021; Ziems et al., 2020). Anti-Asian hate speech could be understood as any speech that targets Asians and people of Asian descendants in a way that elevates hatred, violence, or social disorder.

Scapegoating immigrants in public crises is not unique to Asians during the COVID-19 pandemic, but has a long tradition in American history (Daniels et al., 2021; Reny and Barreto, 2020; Zhang, 2021). For instance, tuberculosis was called "the Jewish disease" in the 1890s and Italian immigrants were blamed for the 1916 polio epidemic. Chinese immigrants were blamed for the spread of smallpox in San Francisco in the later 1870s and the spread of SARS in 2003 (Daniels et al., 2021). Now Asian immigrants, specifically Chinese immigrants, are scapegoated for the origin and spread of COVID-19. Since the pandemic, around 2 million Asian American adults have experienced various forms of anti-Asian hate incident such as being beaten, spit on, and harassed based on AAPI Data reports (Lee, 2021). Federal-level hate crime data also shows that anti-Asian hate crimes increased by 149% in 16 major cities while hate crimes, in general, decreased by 7%. When former President Trump singled out China as the origin of the pandemic and publicly used derogatory terms such as "Chinese Virus" and "Kung Flu," Asian Americans, particularly immigrants of Chinese descendent, became the primary target of anti-Asian hate crimes (Lee, 2021; Horton, 2020).

Anti-Asian bigotry, violence, and misogyny were ingrained in American history following the passage of the Page Act of 1875 and the 1882 Chinese Exclusion Act which prohibited the importation of Chinese laborers and women(Kim, 1999; Zhang, 2021). Asian Americans were often perceived as "disease, unassimilable aliens, and economic, cultural, and moral threats to a free White republic"(Lee, 2021). Even today, Asian Americans, particularly women of Asian descendants, are still the victims of anti-Asian misogyny and hatred (e.g., the Atlanta mass shooting in March 2021). [1] Even before the COVID-19 outbreak, hate crimes against Asians increased by 31% in 2016-2018 under the Trump administration (Lee, 2021).

3 Data

To understand the anti-Asian hate sentiments before and in the pandemic, we use COVID specific hate terms, general anti-Asian hate terms, anti-Chinese politics terms, and counter-hate terms to extract all relevant tweets from Twitter's historical database. This allows us to obtain a holistic view of multiplex anti-China or Asian hate sentiment. Here we briefly describe how we built our datasets and more information can be found in the Appendix. Since online hate speech mainly targets China, we focus on these anti-China related keywords in the present study. In future works, we wish to extend this study to other Asian countries.

COVID-Specific Hate Data First, we use keywords, such as *'chinese virus', 'china virus', 'wuhan virus', 'kungflu'*, and their variants, to extract all relevant tweets that target AAPI communities. Next, we use Ziems et al.'s classifier to identify and exclude all these counter-hate tweets.

General Anti-AAPI Hate Data We use keywords, such as *'bamboo coon', 'chigger', 'chinese*

[1] see NPR article: For Asian American Women, Misogyny And Racism Are Inseparable, Sociologist Says

wetback’,‘ching chong’, ‘chonky’, ‘chunky’, ‘slant eye’, ‘slopehead’, ‘bat eater’, ‘chink’, ‘ling ling’, and *‘commies’* to extract all non-COVID19 related but also hateful tweets that target AAPI communities. Some of the keywords can be used in a multitude of scenarios, in such cases we removed those keywords using a few filters to only collect tweets that target the AAPI community.

Anti-Chinese Politics Data We use keywords and hashtags, such as *‘BoycottChina’*, *‘MakeChinaPay’*, *‘StandWithHongKong’*, *‘FreeTibet’*, *‘FuckChina’*, *‘CCP_is_terrorist’*, and *‘Chinazi’* to extract all tweets that target china politically.

Counter-Hate Data We collect counter-hate data from two sources. First, we use counter hate terms askeywords, such as *‘Racisimisvirus’*, *‘StopAsianHate’*, and *‘StopAAPIHate’*. We then use Ziems et al’s (2020) classifier to extract those counter-hate tweets from the datasets collected with hateful terms.

Table 1 shows basic statistics for our four main datasets collected using the Twitter academic API. Note that these datasets span across different time periods. Researchers can use these datasets for different purposes. For instance, we can use these datasets to test the following hypotheses related to the overall anti-Asian or Chinese hate sentiments in the COVID-19 pandemic or the persistence of anti-Asian sentiment.

Hypothesis 1. The overall anti-Asian hate sentiments should be consistent before and after the outbreak of the COVID-19 pandemic.

Hypothesis 2. The pandemic threat has engendered the rise of COVID-19 specific hate sentiments on Twitter.

4 Results

In this section, we report the major trends and patterns from our four main datasets. We start with COVID19-Specific hate data.

4.1 COVID-Specific Hate Data

In the early stage of the COVID-19 pandemic, Twitter users used COVID-19 specific terms such as *‘chinese virus’, ‘china virus’, ‘wuhan virus’*, and *‘kungflu’* to describe the novel coronavirus. In February 2020, WHO named the disease caused by the novel coronavirus as COVID-19, but still these racial slurs remained popular on Twitter, especially after U.S. President Trump tweeted Chinese Virus multiple times in three consecutive days in mid March, 2020:

> *The United States will be powerfully supporting those industries, like Airlines and others, that are particularly affected by the Chinese Virus. We will be stronger than ever before!* Mar 16, 2020

> *Cuomo wants "all states to be treated the same." But all states aren’t the same. Some are being hit hard by the Chinese Virus, some are being hit practically not at all. New York is a very big "hotspot", West Virginia has, thus far, zero cases. Andrew, keep politics out of it....* Mar 17, 2020

> *I always treated the Chinese Virus very seriously, and have done a very good job from the beginning, including my very early decision to close the "borders" from China-against the wishes of almost all. Many lives were saved. The Fake News new narrative is disgraceful & false!* Mar 18, 2020

In Figure 1, the blue line shows the overall trend of tweets mentioning any COVID-19 related racial slurs, peaking around mid-March when Trump tweeted Chinese Virus. Note that Figure 7 is presented to normalize these patterns based on the estimated total number of tweets.

We also used state-of-the-art hate speech detection algorithms to classify whether these racial slurs count as hate speech. In general, all hate speech detectors have some degree of noise and subjectivity. For this reason, we provide our potential users with three sets of labels classified by algorithms of Ziems et al. (2020), Davidson et al. (2017), and Vidgen et al. (2020), whose aggregated counts are shown in Figure 1. Note that you can find more details regarding these classifiers in the appendix. The strong consistency between Ziems classifier and Vidgen classifier suggests that classification noise does not overwhelm the observed signal. We also notice that the Davidson classifier is less likely to classify tweets as hate speech, partly because it was initially trained on non-group-specific hate tweets.

4.2 General Anti-AAPI Hate Data

While Figure 1 clearly shows an increase in the volume of COVID-19 related hate tweets during

Table 1: **Summary of Twitter Data (in millions)**

	COVID-Hate	AAPI-Hate	Anti-Chinese Politics	Counter-Hate
# of tweets	12.93	12.92	32.6	9.93
# of unique tweets	3.29	7.24	6.36	2.14
# of retweets	9.64	5.68	26.29	7.79
# of Twitter users	3.15	4.58	2.85	3.39
Time range	2019.12-2021.3	2008.1-2021.3	2019.12-2020.12	2018.1-2021.12

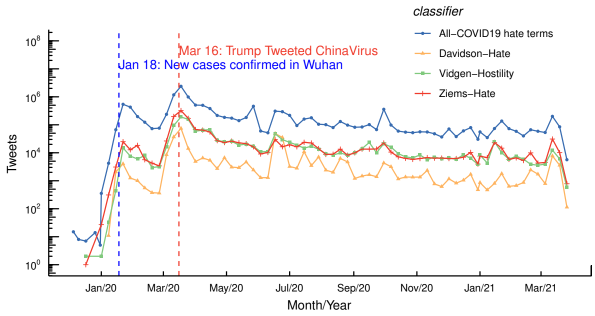

Figure 1: Weekly Trend of COVID-19 Hate Terms, as classified by three different hate-speech detectors.

the pandemic, it is unclear if this corresponds to an increase in hate or an increase in tweets about COVID-19. To provide a baseline, and to investigate anti-China or Asian hate sentiment *before* the COVID-19 outbreak, for comparison, we built the general anti-AAPI hate data using these anti-Chinese or Asian hate terms.

Figure 2 shows the monthly trends of different anti-AAPI hate terms in our database. We believe that these numbers significantly underestimate the true number of abusive tweets, since such slurs are easily identifiable and verifiable after reporting, and thus a large portion of them were removed by Twitter long ago.

The top blue line in Figure 2 shows the number of tweets containing any of the general hate terms between Jan 2008 and March 2021. We see a rapid increase in the number of tweets using anti-AAPI racial slurs from the founding of Twitter in 2007 to early 2013, and this growth may be attributed to the exponential growth of Twitter users at the same time. But after that, we see a decline pattern in the Obama administration before 2017. After Trump took over the Oval office, we see a clear increase in these hateful tweets. This could be attributed to a worsening of the US-China relations due to a growing trade war, or sentiments against the Chinese government due to its role in Taiwan and Hong Kong issues. We also present isolated counts of the major general hate terms used by Twitter users, including *'chink', 'coolie', 'sideways vagina', 'chinaman', 'chonk', etc.* One interesting pattern is that we see a huge increase in using *'Chunk' or 'chonk'* after 2018.

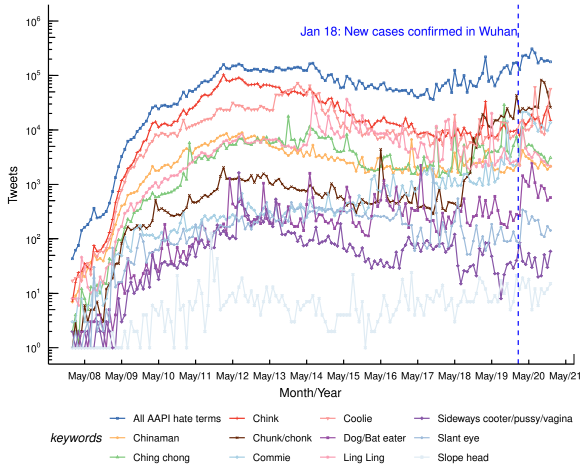

Figure 2: Monthly Trend for Anti-AAPI Hate Terms, as classified by three different hate-speech detectors.

4.3 Anti-Chinese Politics Data

While the general anti-AAPi dataset establishes a much clearer and COVID-agnostic metric of Asian hate, it also has faults as 1) these tweets include many outdated slurs that may not dominate the hateful users' vocabulary anymore, and 2) they are easy to detect by Twitter's own anti-hate software, and easily verified and removed and are thus likely undercounted. We therefore investigate a third AAPI hate dataset which covers a much grayer area, targeting subject matter that attracts hate speech: controversial Chinese politics. Here, our goal is not so much to argue that discussing Chinese politics in a negative way is in itself hateful, but that hateful users tend to use these subjects as an outlet to propagate anti-Asian sentiments. This can be measured, for example, by establishing significant overlap between users who post with obvious anti-Asian slurs and users who post in this dataset.

Figure 3 shows the weekly trend of these anti-Chinese politics terms on Twitter from Jan 2019 to December 2020. We observe an increase in the number of Tweets mentioning any anti-Chinese politics such as *'BoycottChina', 'MakeChinaPay',* and '*Uyghur'* before the outbreak of COVID-19. But since then, the total number of anti-Chinese politics tweets bounced back and fluctuated in the early stage of the pandemic. We suspect Twitter users' attention has shifted from anti-Chinese politics to these COVID-19 specific issues.

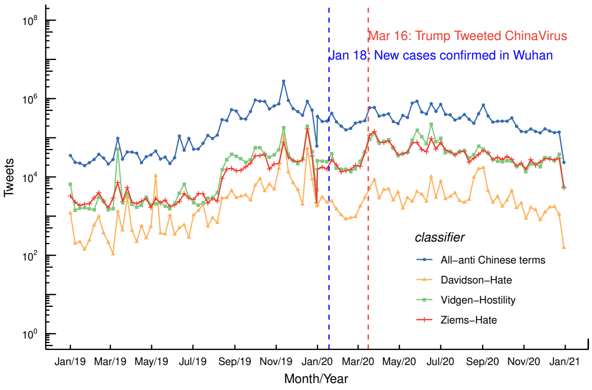

Figure 3: Weekly Trend of Anti-Chinese Politics tweets, as classified by three different hate-speech detectors.

4.4 COVID Counter-Hate Data

In addition to hate datasets, we also built a counter-hate dataset to assess the dynamics between pro- and anti-Chinese or Asian groups. Figure 4 shows the overall counter-hate weekly tweets after the outbreak of COVID-19.

Since the pandemic, we have seen a troublesome surge of anti-Asian attacks. This raises substantial concerns within the AAPI communities. We see a rapid increase in tweets that counter anti-AAPI hate speech such as *'RacismIsVirus'* and '*StopAAPIHate'*. The counter-hate tweets peaked after Trump tweets Chinese Virus on March and then declined. The StopAAPIHate movement took off after the early 2021 and peaked after the tragedy of Atlanta Spa mass shootings on March 16, 2021. Our dataset provides unique de-identified author IDs and conservation IDs which allow researchers to assess the interaction among Twitter users.

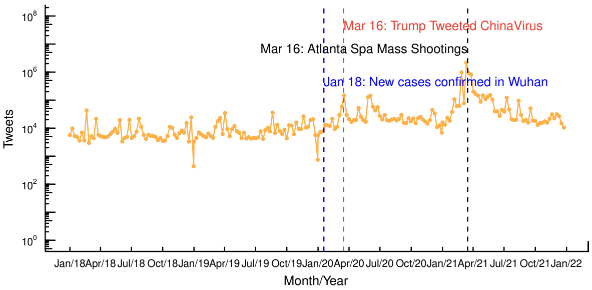

Figure 4: Weekly Trend of Counter-Hate Tweets

4.5 Hashtag Analysis

Here we provide some basic hashtag analysis in our four main datasets. What are the most popular hashtags in our datasets?

Figure 5 shows the hashtags used by Twitter users co-occurring with other keywords. Panel A shows that the most popular hashtags in anti-Chinese politics dataset are *#StandwithHongKong* and *#HongKong*. Panel B shows that the most popular hashtags used by counter hate users are related to *#StopAsianHate* and *#AsiansAreHuman*. Panel C shows that the most popular hashtags in COVID19-specific hate dataset are *#ChineseVirus*, *#CoronaVirus*, and *#WuhanVirus*. Panel D shows that the most popular hashtags used in general anti-AAPI hate dataset are *#boycottChina*, *#China*, and *#CCP*.

4.6 Overlapping Analysis

We conduct an extra analysis to examine the overlapping between COVID-Hate and AAPI-Hate data as well as between COVID-Hate and anti-Chinese politics data. We suspect that Twitter users who expressed general anti-AAPI hate and anti-Chinese politics were also more likely to show COVID-specific hatred in the pandemic. For those who posted COVID-19 specific hate terms in the pandemic, there are 741,802 Twitter users from the general anti-AAPI hate dataset and contributed 7.2 million tweets. These twitter users accounted for 23.57% of total users and 55.71% of total tweets in COVID-Hate dataset. There are also 864,287 Twitter users from anti-Chinese politics dataset overlapping with COVID specific hate data and contributed 7.86 million tweets. These twitter users accounted for 27.46% of total users and 60.77% of total tweets in COVID-Hate dataset. Figure 6 shows the monthly or weekly trends of these overlapping Twitter users.

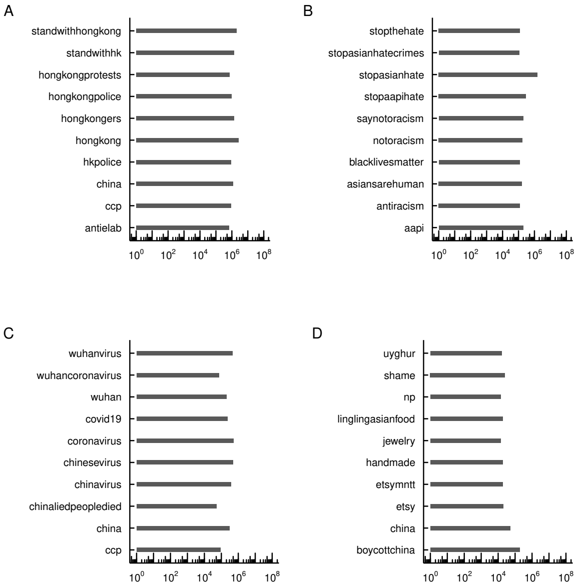

Figure 5: Top Hashtags in Four Main Datasets. Panel A: anti-Chinese politics; Panel B: Counter hate; Panel C: COVID-specific hate; Panel D: General Anti-AAPI Hate.

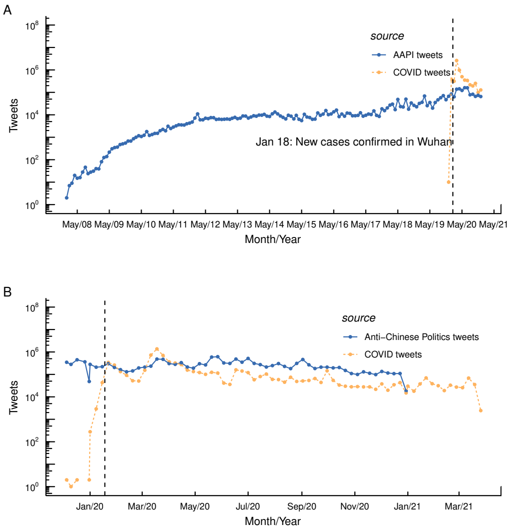

Figure 6: Monthly or Weekly Trend of Overlapping Twitter Users. Panel A: between COVID-Hate and AAPI-Hate; Panel B: between COVID-Hate and Anti-Chinese Politics

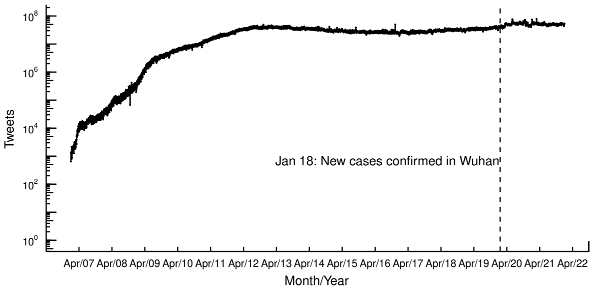

Figure 7: The Baseline Tweets for Normalizing Our Datasets Calculated by Counting the Number of Tweets including Common Words. (See appendix.)

5 Conclusion

This paper introduces new datasets to study anti-Asian hate speech and sentiment on Twitter before and during the pandemic. We show that the overall anti-Chinese/Asian hate sentiments were consistent before and after the outbreak of the COVID-19 pandemic, but the pandemic threat has engendered the rise of COVID-19 specific hate sentiments on Twitter.

Hate speech online is a multiplex phenomenon. We built our datasets using keywords related to COVID-19 specific hate terms, general anti-AAPI hate terms, and anti-Chinese politics terms as well as counter hate terms. As shown in our main analysis, we demonstrate that we can use these datasets to illustrate the overall trends and patterns of anti-Asian hate speech online, and use aggregate statistics to demonstrate and describe the rise in anti-AAPI hate speech during the COVID era.

Researchers can also use these datasets to study how Twitter users are radicalized by engaging into controversial conversations or what the linguistic features of hate speech on Twitter are. We also provide the baseline tweets for researchers to normalize the trend of our datasets as shown in Figure 7. Researchers can also use the de-identified author IDs and conversation IDs (which is unique IDs for all Tweets within the same reply thread and reply threads that are created from earlier reply threads) to conduct conversation network analysis. Future users should be aware of possible underreporting due to many blatantly abusive tweets already being removed. Still, our novel datasets can contribute to research in the areas of computational social science, machine learning, and hate speech detection.

Acknowledgements

We wish to thank the Institute for Advanced Computational Science at Stony Brook University for the access to high performance computing systems and its generous seed grant support for Yongjun Zhang and Yifan Sun.

References

Gavin Cook, Junming Huang, and Yu Xie. 2021. How COVID-19 has impacted american attitudes toward china: A study on twitter. *CoRR*, abs/2108.11040.

Stephen M. Croucher, Thao Nguyen, and Diyako Rahmani. 2020. Prejudice toward asian americans in the covid-19 pandemic: The effects of social media use in the united states. *Frontiers in Communication*, 5.

Chelsea Daniels, Paul DiMaggio, G. Cristina Mora, and Hana Shepherd. 2021. Has pandemic threat stoked xenophobia? how covid-19 influences california voters' attitudes toward diversity and immigration*. volume 36, pages 889–915.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Richard Horton. 2020. Offline: Covid-19 and the dangers of sinophobia. *Lancet (London, England)*, 396(10245):154.

Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. 2020. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211.

Claire Jean Kim. 1999. The racial triangulation of asian americans. *Politics & society*, 27(1):105–138.

Jennifer Lee. 2021. Reckoning with asian america and the new culture war on affirmative action. In *Sociological Forum*. Wiley Online Library.

Tyler T Reny and Matt A Barreto. 2020. Xenophobia in the time of pandemic: othering, anti-asian attitudes, and covid-19. *Politics, Groups, and Identities*, pages 1–24.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Dennis Zhang. 2021. Sinophobic epidemics in america: Historical discontinuity in disease-related yellow peril imaginaries of the past and present. *Journal of Medical Humanities*, 42(1):63–80.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.

A Appendix

A.1 Code and Data Availability

Codes and aggregated data used to replicate main figures are available via `https://osf.io/xtw4c/`. Dis-aggregated and de-identified data are available for academic use upon request. You can email the corresponding author for the data sharing agreement form.

A.2 Key words

Here we provide a detailed list of keywords and hashtags we used to extract all tweets.

COVID-Specific Hate Data. We use keywords, including *'chinese virus', 'china virus', 'wuhan virus', 'wuhan coronavirus', 'kungflu', 'china coronavirus', 'chinese coronavirus', 'chinavirus', 'chinesevirus', 'wuhanvirus', 'Kung flu'*.

General Anti-AAPI Hate Data. *'bamboo coon', 'bamboo coons', 'celestial', 'celestials', 'chigger', 'chiggers', 'chinese wetback', 'chinese wetbacks', 'ching chong', 'ching chongs', 'chinig', 'chinigs', 'chink a billies', 'chink a billy', 'chonkies', 'chonky', 'chunkies', 'chunky', 'coolie', 'coolies', 'sideways cooter', 'sideways cooters', 'sideways pussies', 'sideways pussy', 'sideways vagina', 'sideways vaginas', 'slant eye', 'slant eyes', 'slopehead', 'slopeheads', 'aseng', 'bat eater', 'boycottchina', 'chinadidthis', 'chinaman', 'chinamen', 'chink', 'chinky', 'cokin', 'dog eater', 'fuckchina', 'ling ling', 'makechinapay', 'niakoué', 'pastel de flango', 'slant-eye', 'ting tong', 'idiot chink', 'chinky bat', 'commie', 'commies'*.

Anti-Chinese Politics Data. *'BoycottChina', 'MakeChinaPay', 'BoycottChineseProducts', 'boycottchina', 'StandWithHongKong', 'BoyCottChina', 'Uyghur', 'ReplaceIT', 'BoycottMadeInChina', 'FreeUyghur', 'boycottChina', 'antichinazi', 'CCPChina', 'BoycottChineseProduct', 'FreeTibet', 'FuckChina', 'CCP_is_terrorist', 'FreeHongKong', 'StopChina', 'BOYCOTTChina', 'StandwithHK', 'fuckchina', 'Chinazi', 'Tibet', 'Genocide', 'AnywherebutChina', 'ABC_challenge', 'Uyghurs', 'China_is_terrorist', 'HongKongers', 'BOYCOTTCHINA', 'XiJinping', 'MadeInChina', 'Boycottchina', 'TakeDownTheCCP', 'AntiChinazi', 'FreeHK', 'ChineseProductsInDustbin', 'SOSHK', 'BoycottChineseApp', 'FUCKCHINA', 'SanctionChina', 'RemoveChinaApps', 'chinazism', 'fuckchina', 'SaveUygur', 'Chinamustfall', 'HKPoliceState', 'HoldChinaAccountable', 'StandWithHK', 'Xitler', 'CCPChina', 'HongKongPolice', 'Communist', 'BoycottCh', 'antitotalitarianism', 'ChinaBacksTerror', 'antiELAB', 'FreedomHK', 'TaiwanIsNotChina', 'Hongkongprotest', 'boycottchinaproducts', 'fuckChina'.*

Counter Hate Data. *'StopAsianHate', 'AsiansAreHuman', 'StopAAPIHate', 'stopasianhate', 'NOtoracistMedia', 'RacismIsNotComedy', 'NOSilence', 'StopAsianHateCrimes', 'AsianAmericans', 'PROTECTASIANLIVES', 'AsianLivesMatter', 'RacismIsAVirus', 'RacismIsNotAnOpinion', 'AAPI', 'RacismIsntComedy', 'StopAsianHa', 'STOPASIANHATE', 'NoRacismInMedia', 'SayNOtoRacism', 'STOPASIANRACISM', 'AsiansAreHu', 'IamNotAVirus', 'racismisavirus', 'stopaapihate', 'HATEISAVIRUS', 'ProtectOurElders', 'StopRacism', 'EndAntiAsianViolence', 'StopWhiteTerrorism', 'StopWhiteSupremacy', 'AsianAreHuman', 'stopracism', 'RacismInAmerica', 'RacismIsNotJoke', 'StandForAsians', 'StopTheHate', 'StopTheAttacks', 'stopasianhatecrimes', 'AAPIFightBack', 'FightRacism', 'NoToRacism', 'ProtectAsianWomen', 'AAPIHate', 'WorldAgainstRacism', 'WeCantBeSilenced', 'EndWhiteSupremacy', 'StandWithAsians', 'NoChanceForRacism', 'ProtectAsianLives', 'antiracism', 'EndViolenceAgainstWomen', 'IAmNotAVirus', 'WashTheHate', 'RacismIsAVirus', 'IAmNotCovid19', 'BeCool2Asians', 'StopAAPIHate', 'ActToChange', 'HateIsAVirus'.*

Common (Non-stopwords) Words for Normalization Plot. Note that Twitter API does not accept stop words in the query string to get an estimate of total number of tweets containg the word.

'ask', 'be', 'become', 'begin', 'call', 'can', 'come', 'could', 'do', 'feel', 'find', 'get', 'give', 'go', 'have', 'hear', 'help', 'keep', 'know', 'leave', 'let', 'like', 'live', 'look','make', 'may', 'mean', 'might', 'move', 'need', 'play', 'put', 'run', 'say', 'see', 'seem', 'should', 'show', 'start', 'take', 'talk', 'tell', 'think', 'try', 'turn', 'use', 'want', 'will', 'work', 'would'.

A.3 Information on Classifiers

Davidson et al. (2017)'s Model: The dataset used in constructing the model was scraped from Twitter with the help of hate speech lexicon available on hatespeech.org. The total dataset size is around 25k which were manually labeled into one of the following classes Hate (5.7%), Offensive (77.4%), and Neither (16.7%).

After text-preprocessing like lowercase and stemming, uni-gram, bi-gram and tri-grams were constructed and weighted by their TF-IDF scores. Along with this the authors included binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet.

The authors experimented with different models and finally chose to use Logistic Regression with L2 regularization. They claim that the best performing model has an overall precision of 0.91, recall of 0.90, and F1 score of 0.90. The caveat they mention regarding this classifier is that it tends to classify tweets as less hateful or offensive than the human coders.

Ziems et al. (2020)'s Model: The training data used to build the classifier was a sampled version of a large text corpus scraped from Twitter. The authors manually labeled a set of 3,255 tweets. They then considered a set of 2,290 where all the annotators agreed with the same tag. The dataset is categorized into Hate Speech (18.7%), Counter hate speech (22.5%) and Neutral (58.6%).

The authors constructed three sets of features for classification: Linguistic, Hashtag, and BERT embeddings. Linguistic features are a set of 90 features which span across stylistic, metadata, and psycholinguistic categories. Hashtag feature is a vector representation of the number of occurrences of a hashtag or a keyword from the list the authors had compiled. BERT Embeddings are the embeddings constructed using a BERT model with a classification head that was fine tuned to label the tweets into one of the above mentioned classes.

Along with the BERT classification model authors also, separately trained two feed forward neural networks to classify the tweets using Linguistic and Hashtag features. They concluded that the BERT classification model outperformed these feed forward neural networks with better F1, recall, and precision metrics.

Vidgen et al. (2020)'s Model: A dataset of 20k is scrapped from Twitter using hashtags that relate to East Asian Hate and Virus, some of which express anti-East Asian sentiments. The data is segregated into 6 categories and the distribution is as follows: Hostility (19.5%), Criticism (7.2%), Counter Speech (0.6%), Discussion on East Asian Prejudice (5.1%), and Neutral (67.6%).

The authors combined Counter Speech and Discussion on East Asian Prejudice due to low prevalence and conceptual similarity. They replaced all the hashtags present in the data with suitable thematic words which they constructed during annotation setup or a generic hashtag token. Post this a large language model RoBERTa with a classification head is fine tuned for the task of this classification and they claim that they observed best results with this setup with an F1 score of 0.83 as opposed to their LSTM baseline with F1 score of 0.76.