**Final Report**

Shrey Jambudi, Jacky Pan, Esther Park

Data Curation and Management 04:547:221:01

Dr. Shagun Jhaver

Rutgers University

## Introduction, Background, and Significance

The problem our group wanted to focus on is the rise in anti-Asian hate due to COVID-19. Since this pandemic, there has been a significant rise in verbal and physical attacks against Asians and members of the Asian American and Pacific Islander (AAPI) community. More prominent media outlets frequently covered this issue in the past, but now the general public does not see more prominent news about anti-Asian hate because it seems like the trend has "died." However, the attacks have never stopped, and people, whether young or old, continue to be targets and victims. We wanted to focus on how people in the AAPI community are affected online, mainly on social media. We decided on using Twitter as our social media platform of choice, and in order to see how much of an impact the pandemic had on the insurgence of anti-Asian sentiment, we conducted research with two datasets of curated Tweets related to COVID-19. There are two time periods that are represented by these datasets are right at the beginning of the pandemic and a few months into the lockdown.

Asian hate has been on the rise since the beginning of the pandemic in 2020. Whether it is to direct their anger towards Asians, as the first cases of Covid were found in China, or it is to find a scapegoat to blame. Even before the start of the pandemic, Asians faced unprecedented racism from society, and the pandemic only added fuel to the fire. Asians faced uncalled-for hate speech on social platforms and in public as well, sometimes even leading to physical altercations. Since the start of the pandemic, Asian hate has risen by over 500% in some major cities in the US, and that does not account for the outside of the US (Gover et al., 2020). Our study on anti-Asian hate is an important problem in our modern society. It is crucial to respect each other

regardless of their ethnicity and skin color. Sometimes we forget that we are all human at the end of the day and are vulnerable to what others say and do.

As mentioned earlier, studies and extensive research have been done on this topic. In our research across many academic journals, we have encountered many claims of rising Asian hate, regardless of whether it was in person or online. Some studies were more focused on different occasions with masks and Asians. It concluded that about 61% of perpetrators were white, and 39% were black (Ren & Feagin, 2020). While another study compares three features of hate crimes against Asians to hate crimes against African Americans and Hispanics, including the victim, offender, and incident-related factors, using these models as a guide. But since this research is focused online on Twitter, we followed a journal that focuses on online hate speech. In the article, they examined Twitter as well and placed tweets in classifications. In the tweet classification labels, users are classified into one of four categories depending on their tweets: hatred, counter-speech, dual, or neutral. Hate users send at least one hate tweet but no counter-propaganda tweets.

On the other hand, counter-speech users send at least one counter-speech tweet but no hate tweet. Dual users are those who tweet from both categories at the same time. Dual users are those who tweet from both categories at the same time. Users that send at least one COVID-19 tweet but no hate or counter-speech tweets are classified as neutral.

## Related Work

### *The Anxiety of Being Asian American: Hate Crimes and Negative Biases During the COVID-19 Pandemic*

The article starts with the reports of physical and psychological Asian hate crimes from vandalism to stabbings. Physical assaults and bias statements that were reported were covered by the article. This article covers terms of Asian hate speech that were used such as “Chinese virus, diseased, go back to China” and many other phrases that are offensive and inappropriate. The limitation of this article is that it was a study done when most people were still in quarantine and the vaccine for Covid-19 was not developed yet. As fear was instilled into many due to the fact that there was no vaccine for it and many people were still unemployed at the time.

### *Hate Crimes against Asian Americans*

The paper has done an analysis and two theoretical models are provided to guide the analysis: minority-general and minority-specific models. The analysis compares three features of hate crimes against Asians to hate crimes against African Americans and Hispanics, including the victim, offender, and incident-related factors, using these models as a guide. The study demonstrates the value of conducting a comparative analysis of hate crimes against different racial/ethnic groups to identify incidence similarities and differences. This study highlights the need of comparing hate crimes against different racial/ethnic groups to uncover parallels and differences in prevalence

### *Face mask symbolism in anti-Asian hate crimes*

This article gives a study of perpetrators that committed Asian hate crimes that are based on their complexion. While the article was more focused on different occasions with masks and Asians it concluded that about 61% of perpetrators were white and 39% were black. While this

study was more about race and mental disorders and Asian immigrants and Asian Americans. The limitation to the work is that it is only focused on a mental health issue while we are focusing more on social media hate and anti-hate.

***Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality***

The article uses othering theory to explore and analyze how hate crimes are rooted at the individual and institutional level. Racism and xenophobia have been affecting the other Asian Americans as emphasized in the article which is creating inequality. To critically examine this dynamic of hate, a mix of media and empirical sources was used. Anti-Asian hate crime during the COVID-19 pandemic is discussed historically and in relation to racist and xenophobic othering in this research paper. Finally, the research aims to broaden knowledge by examining Asian Americans' experiences with pandemic-driven hate crimes in order to more precisely define the idea of othering.

***Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis***

The characteristics of users who post hate and counter-speech tweets are examined in this article. Following the tweet classification labels, users are classified into one of four categories depending on their tweets: hatred, counter-speech, dual, or neutral. Hate users send at least one hate tweet but no counter-propaganda tweets. Counter-speech users, on the other hand, send at least one counter-speech tweet but no hate tweet. Dual users are those who tweet from both categories at the same time. Dual users are those who tweet from both categories at the same time. Users that send at least one COVID-19 tweet but no hate or counter-speech tweets are

classified as neutral. We will be following along the lines of this article as it gives us a great foundation to build our own research.

## Objectives, Goals, and Outcomes

Our original objective from the proposal stage was to collect data from both hashtags related to hate speech and counterspeech from February 2021 to February 2022 and examine the trends during this one year period. We also wanted to specifically look at the time frames of major events associated with the Asian community, such as the first spike of the Omicron variant, the Lunar New Year, and the 2022 Beijing Olympics. Halfway through our project, we were given feedback on how to identify more relevant hashtags utilizing an iterative technique based on co-occurrence. First, we would find all tweets with the phrase "#ChinaVirus," for example. Then, we would look over these tweets for any hashtags (words that begin with a '#') and count how many times each one appears. Next, we would have to sort the hashtags by how often they're used. Then we would examine the top few hashtags manually to determine if there are any others for which you should collect data. We would then repeat this process to gather more relevant hashtags for us.

However, due to problems that we faced during the data collection stage, we chose to analyze pre-existing datasets and search for keywords, instead of only looking at hashtags, associated with anti-Asian hate in the Tweets that were posted during December 2019 to February 2020 (which represents the time period when it was the very beginning of the COVID-19 pandemic) and July 2020 to August 2020 (which represents the time period of the

lockdown was in place). Our overall goal remained the same and was to raise awareness to a broader audience about the hate against the Asian community on social media has not changed.

From the pre-existing datasets, we predicted that the data collected would show an increase in time of the lockdown and the beginning of the quarantine worldwide. The number of hate speech incidents would be higher in the months of the shut down as compared to before the shut down. We made the prediction due to the fact that people were starting to get frustrated being stuck at home constantly with the same people and only going out for necessities. Other factors were because of the origins of where the virus came from and comments from President Trump and China going back and forward over many issues. In addition, many countries and people started to blame China for not listening to its doctor calling for a national crisis and causing massive outbreaks around the world. We also predicted that the amount of hate would rise and the amount of anti-hate will try to counterbalance it as well.

The outcomes and findings did not meet our expectations because the data showed an overall decrease in anti-Asian sentiment during the time between before the lockdown and after the lockdown, which is the opposite of what we predicted. We compared the number of tweets that had keywords pertaining to anti-Asian hate from the December 2019 to February 2020 dataset with the dataset from July 2020 to August 2020. When we conducted a sentiment analysis, the number of tweets that had a more positive sentiment slightly increased during this time period; the number of tweets that were neutral decreased; and the tweets that had a negative sentiment slightly increased. The differences in expectations and outputs are most likely due to the pre-existing datasets that we had to resort to using for this project. The datasets included tweets that were specifically related to COVID-19 and one of the datasets included an equal number of positive and negative tweets. These datasets were also small to begin with and in

order to work with a consistent number of tweets for each dataset, the number of rows were decreased, further reducing the size of the datasets. If we were able to work with more data of tweets from a wider time frame, we potentially could have had different results that would better match our predictions.

## Description of Work Accomplished: Data

One challenge that our group faced was gaining access to the Twitter API. We had trouble accessing the data that we needed through the Twitter API, and unfortunately our research group did not meet the criteria to get academic research access. We tried many available methods online from using tweepy and following online forums and videos, but they were not usable for our project. As we requested Twitter to use their API, for scholarly use and with our Rutgers emails, we were denied access to it still. And it was not just once but multiple times as we used different approaches to try to get access to their API.

After failing to secure permission to use Twitter's API, we consulted with the professor and decided to search for open source datasets online to continue our research. We narrowed our search to tweets specifically related to COVID-19, since we wanted to see how the pandemic was affecting hate against the Asian community. Fortunately, we were able to find a few datasets that were aligned with our research. Some of the datasets that we found had critical information that is needed for our research such as the time it was tweeted and the tweet content. We decided to use two datasets that were relevant to our research: one with around 60,000 rows of tweets related to COVID-19 from December 2019 to February 2020 and the other with about 179,000 rows of data from July 2020 to August 2020.

After loading these two datasets into a Jupyter Notebook file, we first cleaned up the data by removing any duplicate values and null values. These datasets were also formatted differently (different number of rows and columns) so the next step we took was to ensure that the dataframes created from these datasets were using the same format. One dataset had more rows than the other so in order to reduce sampling bias we took a random sample from the larger dataset and made sure that each dataset had the same amount of data. Next, we removed unnecessary columns and renamed them so that we could work with two dataframes of the same format.

## Description of Work Accomplished: Approach

For one of our approaches, we decided to run a sentiment analysis on both sets of data to get a better understanding of the sentiment values of each data set. This is appropriate because we are trying to see if there was an increase or decrease in the negativity in the tweets between the two time periods, which was before and after the shut down. The strengths of using sentiment analysis is that it allows us to see how much of the data is negative and positive and neutral and helps us get an idea of what kind of data set it is. The limitations of this approach is that this is only the analysis of the tweet content, and it does not have the best ability to tell the difference in the meaning of the text it reads; it would be difficult to tell if the tweets labeled as negative were because it contained anti-Asian hate or because it had a general negative comment on the pandemic. And it is limited to only the English language, making it difficult to scan the tweets that are in other languages.

The next approach was finding certain keywords related to anti-Asian hate and comparing the number of times the keywords appeared in both datasets. One of the advantages of making these comparisons using the keywords is that we can control what the definition of negativity is. This is useful for our case since we are labeling negative tweets as those with the keywords that show anti-Asian hate. The limitations of this approach is that the keywords may not be present in all of the tweets. The absence of certain keywords does not necessarily mean that there was a decrease in hate; it could possibly indicate that the keyword is irrelevant or that there needs to be more data to work with. Another disadvantage of using this approach is that this is more suitable for larger datasets.

## Description of Work Accomplished: Results

https://github.com/sjambudi/Data-Curation

By using the neattext.function, we were able to remove unnecessary emojis and symbols from the tweet content so that we could work with cleaner datasets. We then created new columns called ‘clean_tweets’ and ‘sentiment_results’ by using textblob to show the newly cleaned tweet content and the sentiment analysis results for each tweet. The results are shown in the following screenshots below.

| | Tweet Posted Time (UTC) | Tweet Content | Tweet_Language | Verification | clean_tweets | sentiment_results |
|---|---|---|---|---|---|---|
| **0** | 28 Feb 2020 15:44:49 | Also the entire Swiss Football League is on ho... | English | Verified | also the entire swiss football league is on ho... | {'polarity': -0.049999999999999996, 'subjectiv... |
| **1** | 28 Feb 2020 15:44:40 | World Health Org Official: Trump's press confe... | English | Verified | world health org official trumps press confere... | {'polarity': 0.21904761904761905, 'subjectivit... |
| **2** | 28 Feb 2020 15:44:39 | I mean, Liberals are cheer-leading this #Coron... | English | Verified | i mean liberals are cheerleading this coronavi... | {'polarity': -0.07625, 'subjectivity': 0.61375... |
| **3** | 28 Feb 2020 15:44:29 | Under repeated questioning, Pompeo refuses to ... | English | Verified | under repeated questioning pompeo refuses to s... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **4** | 28 Feb 2020 15:44:23 | #coronavirus comments now from @larry_kudlow h... | English | Verified | coronavirus comments now from kudlow here | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **...** | ... | ... | ... | ... | ... | ... |
| **60154** | 10 Dec 2019 11:49:59 | RT @timhquotes: It's my party, you're invited!... | English | Non-Verified | its my party youre invited ps this is my life... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **60156** | 10 Dec 2019 01:33:34 | RT @timhquotes: It's my party, you're invited!... | English | Non-Verified | its my party youre invited ps this is my life... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **60157** | 10 Dec 2019 01:10:03 | It's my party, you're invited!\n\nPS, this is ... | English | Non-Verified | its my party youre invited ps this is my life ... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **60158** | 03 Dec 2019 22:57:36 | Amy's a survivor! #bariclab #pnnl #movingon #c... | English | Non-Verified | amys a survivor bariclab pnnl movingon coronav... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **60159** | 01 Dec 2019 03:17:00 | A review of asymptomatic and sub-clinical Midd... | English | Non-Verified | a review of asymptomatic and subclinical middl... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |

33174 rows × 6 columns

**Figure 1: Screenshot displaying Clean data for Dataset 1 (Timeframe 2019-2020)**

| | Tweet Posted Time (UTC) | Tweet Content | Verified or Non-verified | clean_tweets | sentiment_results |
|---|---|---|---|---|---|
| **0** | 2020-08-04 07:54:21 | #Nigeria #COVID19 #Update Tuesday, August 4th ... | False | nigeria covid19 update tuesday august 4th 2020... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **1** | 2020-07-27 03:29:54 | #NOVACYT #COVID19 #ATOTHIS #NCYT #ALNOV #PRIME... | False | novacyt covid19 atothis ncyt alnov primerdesig... | {'polarity': 0.0, 'subjectivity': 1.0, 'sentim... |
| **2** | 2020-07-27 05:08:44 | #onlineeducation is here to stay: to make the ... | False | onlineeducation is here to stay to make the fo... | {'polarity': 0.0, 'subjectivity': 0.125, 'sent... |
| **3** | 2020-08-08 13:45:40 | More DEATH from #COVID19 then in all of WW2 \n... | False | more death from covid19 then in all of ww2 | {'polarity': 0.5, 'subjectivity': 0.5, 'sentim... |
| **4** | 2020-08-06 15:55:08 | The #COVID19 pandemic has created opportunity ... | False | the covid19 pandemic has created opportunity f... | {'polarity': 0.02500000000000001, 'subjectivit... |
| **...** | ... | ... | ... | ... | ... |
| **33172** | 2020-08-29 22:45:26 | #COVID19 exposures have been reported on anoth... | True | covid19 exposures have been reported on anothe... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **33173** | 2020-07-31 18:00:00 | Black people are dying from #COVID19 at 2.5 ti... | False | black people are dying from covid19 at 25 time... | {'polarity': -0.08333333333333333, 'subjectivi... |
| **33174** | 2020-08-14 05:47:46 | #CoronaInfoCH #COVID19 #corona #us \nMost Reco... | False | coronainfoch covid19 corona us most recovered ... | {'polarity': 0.25, 'subjectivity': 0.25, 'sent... |
| **33175** | 2020-07-25 12:08:51 | 10,000+ #healthworkers have been infected with... | False | 10000 healthworkers have been infected with co... | {'polarity': 0.0, 'subjectivity': 0.0, 'sentim... |
| **33176** | 2020-08-22 10:35:26 | Funny that the narrative is slowly moving towa... | False | funny that the narrative is slowly moving towa... | {'polarity': 0.21666666666666665, 'subjectivit... |

33175 rows × 5 columns

**Figure 2: Screenshot displaying Clean data for Dataset 2 (Timeframe Mid 2020)**

By normalizing the ‘sentiment_results’ column using json, we obtained the polarity, subjectivity, and sentiment description in a new dataframe. With the values of the sentiment description, we were able to get the results for negative, positive, and neutral shown in the charts below.

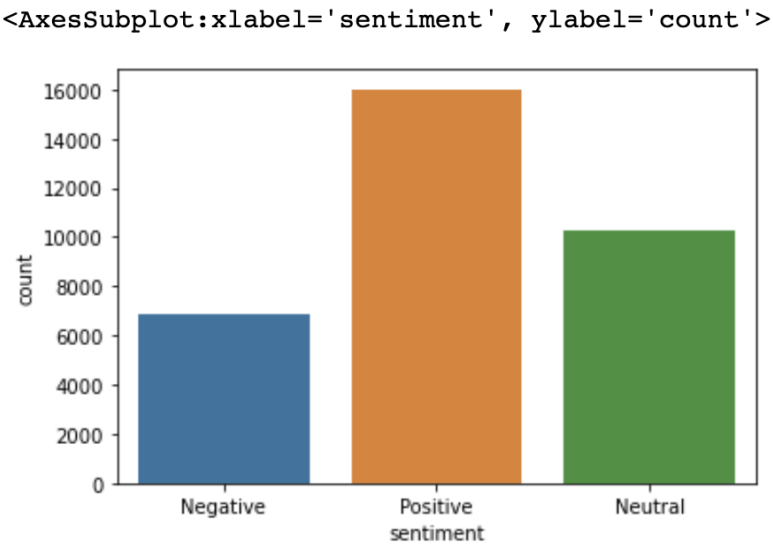


**Figure 3: Sentiment analysis results for Dataset 1 (Timeframe 2019-2020)**

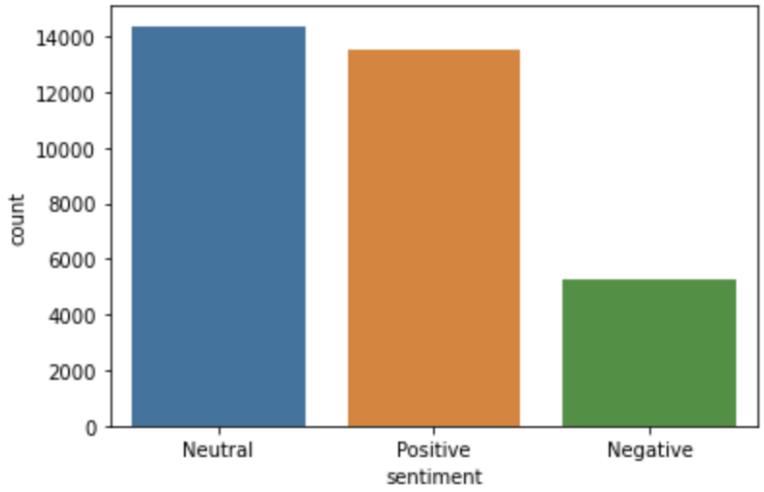


**Figure 4: Sentiment analysis results for Dataset 2 (Timeframe Mid 2020)**

As we did our sentiment analysis we realized that the results were different from what we had initially expected. It was almost opposite of what we predicted as we initially expected it to be mostly negative rather than positive. But a larger portion of the analysis showed that it was mostly neutral, and there was only a constant flow of positive sentiment and only a slight increase in the amount of negative sentiment. As the results of the two datasets analysis had only slight changes and mostly towards being neutral. As this could mean that the analysis may not have been able to see the change in sentiment or it was not able to tell if the tweets were actually negative or positive. But regardless it can be said that maybe it is right and the possibility that there was only a slight change in sentiment and only a constant influx of positive sentiments and negative sentiments.

We analyzed the properties of the users who produce hate and counterspeech tweets. Following the tweet categorization labels, we categorized users, based on their tweets, into one of the following: positive, negative, or neutral. Users who tweet from both categories 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 33,175 users in the dataset 1, the users (48.25%) are Positive, 10,297 (31.03%) are Neutral, 6,870 (20.72%) are Negative. For Dataset 2 the users 13,506 (40.73%) are Positive, 14,381 (43.34%) are Neutral, 5,288 (15.93%) are Negative. This distribution mimics the category-wise tweet distribution.

| | Keyword | 2019-2020 | 2020 |
|---|---|---|---|
| 0 | china virus | 138 | 64 |
| 1 | chinesevirus | 112 | 56 |
| 2 | kungflu | 90 | 34 |
| 3 | ccpvirus | 100 | 69 |
| 4 | whuhanvirus | 314 | 210 |

**Figure 5: Table showing comparison between the number of keywords found in each dataset**

<AxesSubplot:>

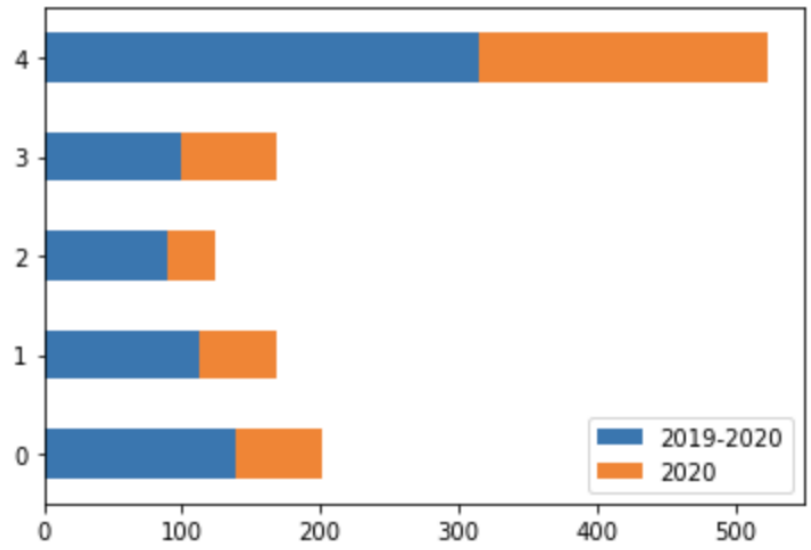


**Figure 6: Graph showing comparison between the number of keywords found in each dataset**

Selecting certain keywords in the data sets and comparing them allows us to see if the number of hate and COVID is correlated to each other. As we selected five keywords from the data sets and they both have multiple results and the amount of times they appeared varied in both data sets. As in the later data set there was actually less number of keywords found than the earlier data set. In retrospect it may be due to the data set that we obtained not having as many keywords in the data and the limited amount of data available. As at the start of the spread of the COVID there were more keywords than during the quarantine period.

## Discussion of Outcomes, Implications, and Conclusion

The finding allowed us to see the rise in Covid related issues and how many people are frustrated and unhappy with the situation that is happening around them. The amount of times people express their anger online and wanting to find a solution or a scapegoat for their problems. Especially after the shutdown and people losing their jobs and homes the rise in hate and tweets increased over the period of time. Our findings align with some of the articles of reading that we have experienced reading and it is not surprising as many people have nothing to do but to stay at home and unable to go places due to the restrictions that were implemented. Even though many people were home they were scared of visiting others or others visiting them as they can potentially be carriers for Covid and if they have people with high risk of dying of Covid it was preferable for them to stay away as it was not safe for them. Holidays and gatherings were different because of Covid as travel restrictions and the fear of spread of Covid. The limitations of the work is it was only for a period of time as we were not able to get our original data to see the difference in hate when Asians are shown because many hold a grudge

against Asians as the virus came from Asia whether it is intentional or not. But the research will be around for at least the next few years as people are still trying to recover from their experiences caused by COVID-19 and people being ignorant and biased. The only promising direction of the work would be to research the decrease in Asian hate over the next few years as everyday life is slowly changing back to normal but will it be the same as before the pandemic? That is something hard to say as COVID did change many aspects of life for better and worse. But the research is still very important to continue as it will show human behavior and other problems that are connected.

## List of Teammate Names

Shrey Jambudi, Jacky Pan, Esther Park

**References**

Tessler, H., Choi, M., & Kao, G. (2020, June 10). *The anxiety of being Asian American: Hate crimes and negative biases during the COVID-19 pandemic - American Journal of Criminal Justice*. SpringerLink. Retrieved February 10, 2022, from https://link.springer.com/article/10.1007/s12103-020-09541-5

Gover, A. R., Harper, S. B., & Langton, L. (2020, July 7). *Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of Inequality - American Journal of Criminal Justice*. SpringerLink. Retrieved February 10, 2022, from https://link.springer.com/article/10.1007/s12103-020-09545-1

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. https://doi.org/10.1145/3487351.3488324

Zhang, Y., Zhang, L., & Benton, F. (2021, January 7). *Hate crimes against asian Americans - american journal of criminal justice*. SpringerLink. Retrieved February 10, 2022, from https://link.springer.com/article/10.1007/s12103-020-09602-9

Ren, J., & Feagin, J. (2020). Face mask symbolism in anti-Asian hate crimes. *Ethnic and Racial Studies*, *44*(5), 746–758. https://doi.org/10.1080/01419870.2020.1826553