APAN 5205 Final Project Report

Detecting Anti-Asian Sentiments on Twitter during COVID-19

Group 8: Brendan Ng, Eden Vacnich, Jenika Shah, Nipun Lohia, Sakshi Mehta, Yueming Song

Contents

1. Abstract

The COVID-19 pandemic has led to a troubling rise in anti-Asian hate and discrimination in the United States. Incidents of hate crimes and discriminatory acts against Asian Americans have surged, both in real-life and on social media platforms. Social media has also witnessed the spread of racism and hateful behavior towards Asian communities, with derogatory language and misinformation being circulated. In this research report, we aim to analyze the sentiments and emotions expressed in tweets related to anti-Asian hate and counterspeech during the COVID-19 crisis, as well as the most frequently used words and phrases in these tweets and how they vary over time. Through a comprehensive analysis of a large dataset of tweets, we aim to shed light on the content and sentiment of online discourse related to anti-Asian hate and counterspeech during the COVID-19 pandemic.

2. Background

The COVID-19 pandemic has had a significant impact on people's lives and economies worldwide, with governments implementing measures to limit the spread of the virus. However, along with the spread of the virus, there has been a sharp increase in hate crimes and discriminatory acts against Asian Americans in the United States. Social media platforms have also witnessed the spread of racism, fueled by derogatory language and misinformation, including the use of discriminatory terms such as "Chinese Virus" by public figures.

Amidst this backdrop, Bing He et al. (2020) conducted a study investigating the prevalence and spread of anti-Asian hate speech on Twitter, and the potential of counter-speech to mitigate its impact. The researchers collected a large dataset of tweets and used text classification methods to identify anti-Asian hate speech, counter-speech tweets, and neutral tweets in their dataset. Their findings revealed insights into the prevalence and nature of anti-Asian hate speech on social media during the COVID-19 crisis. This study highlights the significance of the issue and the role of social media in shaping online discourse related to anti-Asian hate during the COVID-19 crisis.

3. Research Questions

In this study, we seek to address the following research questions:

- *What are the most frequently used words and phrases in tweets related to anti-Asian hate during the COVID-19 crisis?*
- *How has the intensity of anti-Asian hate changed over time?*
- *What are the dominant sentiments and emotions expressed in tweets related to anti-Asian hate?*

4. Methodology

To answer these research questions, we collected and analyzed a large dataset of tweets related to anti-Asian hate and counterspeech during the COVID-19 crisis. We used text classification methods, including linguistic features, hashtag features, and BERT tweet embeddings, to identify the most frequently used words and phrases, as well as the dominant sentiments and emotions expressed in these tweets. The findings of this study provided insights into the content and sentiment of tweets related to anti-Asian hate and counterspeech during the COVID-19 pandemic and contributed to our understanding of the issue in the context of social media discourse.

5. Data Cleaning and Preparation

We use three datasets from the same source with the same measurements but with different time frames to analyze twitter speeches, including "Covid-19 Twitter Dataset (Apr-Jun 2020)", "Covid-19 Twitter Dataset (Aug-Sep 2020)", and "Covid-19 Twitter Dataset (Apr-Jun 2021)". These datasets were then combined into a single dataframe. To save memory, the individual datasets are removed from the R environment and only keep the combined dataset.

Duplicate tweets were identified and removed, keeping only the observed tweet with the maximum retweet count. Additionally, date reformatting was performed for the use of further analysis. In detail, we extracted the year, month, and day components from the original datetime column using regular expressions. Then we used these extracted components to create a reformatted date string in the format "YYYY/MM/DD " for each row of the data. At last, we convert the reformatted date into a date type.

To clean the contents of tweets, we removed the "RT @username:" substring from the original tweet texts, and the cleaned tweets were stored in a new column named "edited_tweet".

The tm package in R was used to create a "corpus" object, which represents a collection of text documents. We utilized the cleaned tweets without prefix generated from the previous step to create this corpus. Next, we converted all the text to lowercase, removed punctuation marks from the text, and eliminated stop words, which are common English words like "the", "and", etc., by using the removeWords function with a list of stopwords from the 'english' language. We stripped whitespace from the text and removed numbers. Additionally, we used regular expressions to remove specific punctuation marks like '“', '”', '’', and URLs from the text to further prepare the tweets’ contents for future sentiment analysis and text mining analysis by removing all potential no-meaning words. These steps helped us preprocess the text data in the "corpus" object by cleaning and normalizing the text for further analysis.

Lastly, a new column called "test" was added to the data frame for potential date filtration, which can be removed later. The "created_at" column was then converted to a date object with the format "%a %b %d %H:%M:%S %z %Y". This allowed for extraction of the day of the week component as a numeric value using the "$wday" attribute of the "created_at_n" object, which represents the date in POSIXlt format. Additionally, the hour component of the "created_at_n" object was extracted using the "$hour" attribute, representing the hour of the day. Finally, unnecessary columns from the data frame were removed using the subset function, selecting all columns except "created_at", "lang", "compound", "neg", "neu", and "pos". These steps facilitated the processing and transformation of date-related information in the data frame for further analysis.

6. Analysis Techniques & Key Findings

a. Data Exploration & Basic Text Mining

The analysis conducted in this study utilized relevant and suitable data to address the research questions on anti-Asian hate and counterspeech during the COVID-19 crisis. The dataset used in the analysis consisted of a large volume of tweets collected during the specified time period, providing a comprehensive sample of tweets related to the research topic. The inclusion of tweets from different geographical locations allowed for a diverse representation of sentiments and emotions expressed in different contexts, enhancing the robustness of the analysis.

The average number of characters in a tweet was 124, while the average number of words per tweet was roughly 18 words. The number of tweets with hashtags was 97692, which accounted for 23.74% of the overall tweets. The most frequent word that occurred in all tweets is "COVID", which is a neural term (appendix 1). We also explored common words by their sentiment, i.e. positive or negative. The positive common words identified were "trump", “support”, “positive”, “free”, “protect”, and “safe”. The negative common words included “death”, “crisis”, “breaking”, “emergency”, “symptoms”, and “risk” (appendix 2). This provides insights into the topics that were most commonly discussed in relation to anti-Asian hate and counterspeech during the COVID-19 crisis.

b. Sentiment Analysis

Sentiment analysis was employed using established sentiment lexicons, such as AFINN and NRC, to quantify and analyze sentiments and emotions in the tweets. The analysis revealed that the majority of the tweets were neutral, followed by positive and negative sentiments. The number and proportion of hateful COVID-related Tweets were also plotted over time, showing fluctuations in sentiment during different periods of the pandemic, which can help identify patterns and trends in sentiment expression related to anti-Asian hate and counterspeech (appendix 3 & 4).

The sentiment analysis was conducted using two different lexicons, namely AFINN and NRC, to examine sentiment trends in social media data.

For AFINN analysis, sentiment scores were calculated for each tweet by summing the sentiment scores for the individual terms in the tweet based on the AFINN lexicon. The sentiment intensity was then categorized into five categories: "Strongly negative", "Moderately negative", "Neutral", "Moderately positive", and "Strongly positive". Additionally, sentiment analysis allowed us to investigate the intensity of hate and anti-hate. Hateful tweets were identified by filtering the sentiment scores data frame for tweets containing racist terms or hashtags (e.g. chinavirus) and a negative total sentiment score. This will be elaborated upon in the time series analysis section.

For NRC analysis, the NRC lexicon was utilized to examine sentiment-associated word counts in the tweet data. The lexicon was filtered to exclude words with a numerical value of zero and then joined with the tokenized tweet data where tweets were separated into individual words. The overall word counts associated with different sentiments were calculated. The resulting counts were arranged in descending order. The analysis revealed that the most frequent sentiments in the tweet data were positive, negative, trust, fear, and anticipation, with positive being the most common and disgust being the least common (appendix 5).

The relationship between sentiment and retweet count was also examined, specifically focusing on the sentiments of disgust, anger, and fear. Correlation coefficients were calculated between the square root of retweet count and the frequency of each sentiment. The correlations were small, indicating a weak relationship between retweet count and the frequency of these sentiments in the tweet content. Linear regression analysis was also performed to further explore the relationship between retweet count and the frequency of disgust and anger sentiments. The regression models showed that the retweet count was not a significant predictor of the frequency of these sentiments in the tweet data, except for disgust, where there was a statistically significant effect, although the magnitude of the effect was small. These findings suggest that retweet count may have a limited influence on the occurrence of disgust and anger sentiments in tweets, while other factors may play a more prominent role in shaping sentiment expression on social media.

c. Time Series Analysis

In addition to sentiment analysis, time series analysis was used to examine temporal patterns and trends in the occurrence of tweets. This allowed for the identification of the number of tweets related to anti-Asian hate and counterspeech, and how they varied over time.

As mentioned above, we identified hateful tweets based on sentiment score and key words. The monthly hateful tweet count was calculated by grouping the data by month and summarizing the counts. The resulting data was then visualized using a line graph. The graph titled "Number of Hateful Tweets per Month" (appendix 3) and "Proportion of Hateful COVID-related Tweets" (appendix 4) shows the trend in negative tweets over time, with the x-axis representing the months and the y-axis representing the count and proportion of negative tweets. The graph helps in identifying patterns and trends in the occurrence of negative tweets over the specified time period, such as sharp increases, dips, and consistent rises in negative tweets at different time points. The proportion of hateful tweets seems to have peaked in August 2020.

We then calculated the mean sentiment score of all hateful tweets for each month. The sign of the result was flipped so that it would make more sense when visualized (e.g. upward trending line would mean higher hate intensity). Our findings showed that hate intensity had a general downward trend after the peak of the pandemic. This analysis sheds light on the trend of sentiment towards China or Asian people in the context of COVID-19.

Lastly, we checked whether hateful tweets would tend to have more retweets compared to all the other COVID-19 related tweets in the dataset. The resulting visualization shown in appendix 10 suggests that hateful tweets were sometimes more likely to be retweeted. Hate, just like COVID, is apparently viral.

Overall, the chosen techniques of sentiment analysis and time series analysis were effectively used to address the research questions related to anti-Asian hate and counterspeech during the COVID-19 crisis, providing valuable insights into the sentiment trends and temporal patterns in tweets related to this topic. These findings can be useful for researchers and policymakers in understanding public perception and sentiment towards the pandemic and vaccinations on social media and may inform public health messaging and communication strategies during pandemics.

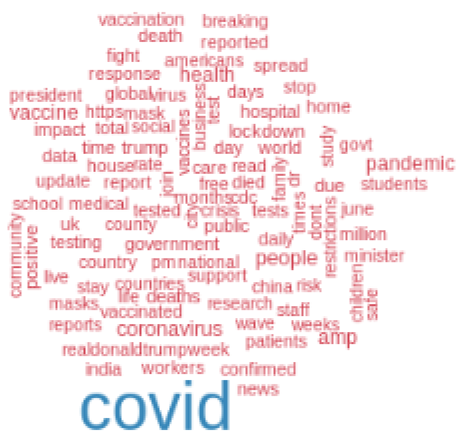
However, it is important to acknowledge the limitations of these techniques. Sentiment analysis may not capture the complexity and nuances of emotions expressed in tweets comprehensively, as it relies on predefined sentiment lexicons that may not encompass all possible sentiments. Additionally, time series analysis may be subject to limitations arising from the availability and quality of data, such as the fact that tweet data was only available for the periods April-June 2020, August-October 2020 and April-June 2021. Therefore, cautious interpretation of the findings and consideration of these limitations is necessary when drawing conclusions from the analysis results.

7. Conclusion

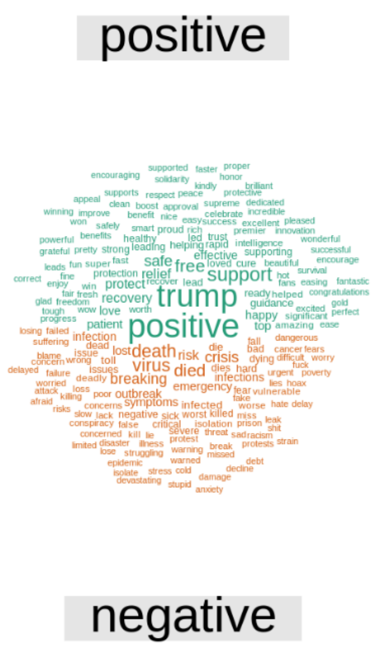
In conclusion, the anti-Asian hate intensity peaked during the first and second wave of COVID, and dropped significantly after March 2021. Some of the most frequently occurring positive words were “positive”, “guidance”, “recovery”, “support”, and “safe”, while frequently occurring negative words included “death”, “virus”, “breaking”, “crisis”, “risk”, and “emergency”. The four top dominant sentiments and emotions in tweets were positive, negative, trust and fear. The most frequent hashtags by retweets for anti-Asian Hate (Appendix 6) highlights the top five hashtags, including ChineseVirus, CCPVirus, ChinaLetPeopleDie, ChinaVirus, and wuhanvirus, which also shows the intensity of anti-Asian Hate, specifically China, during COVID.

Appendix

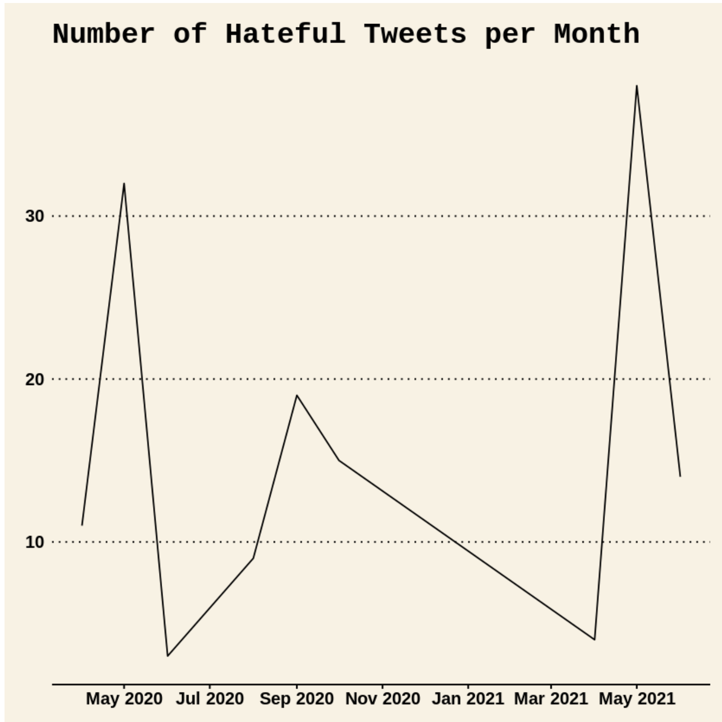
Appendix 1



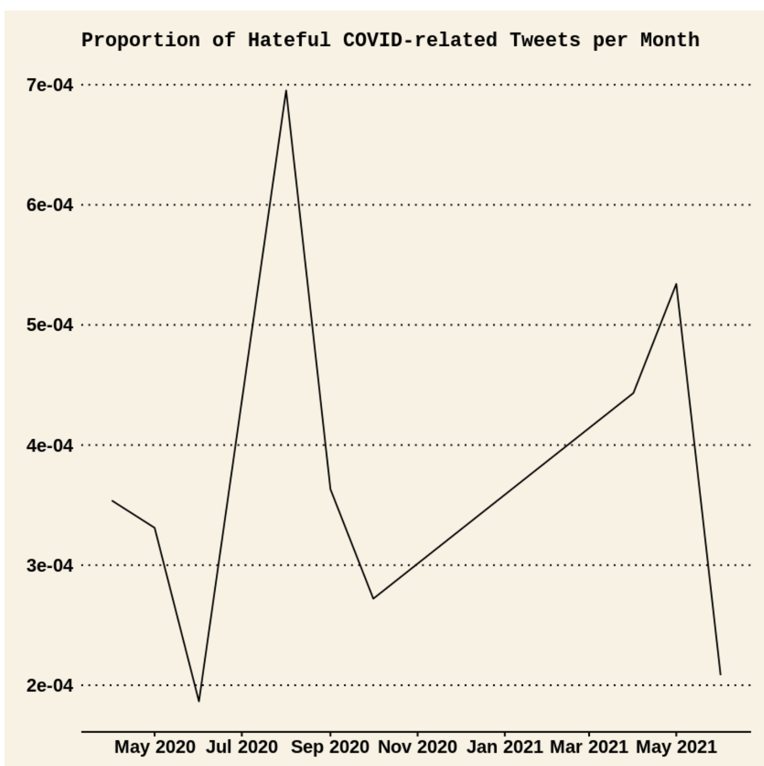
Appendix 2



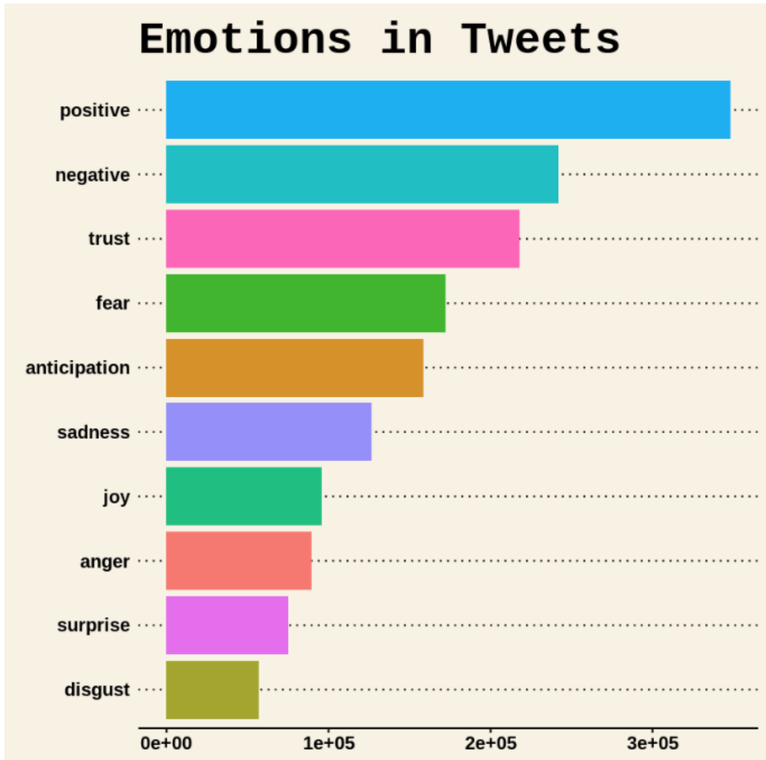
Appendix 3



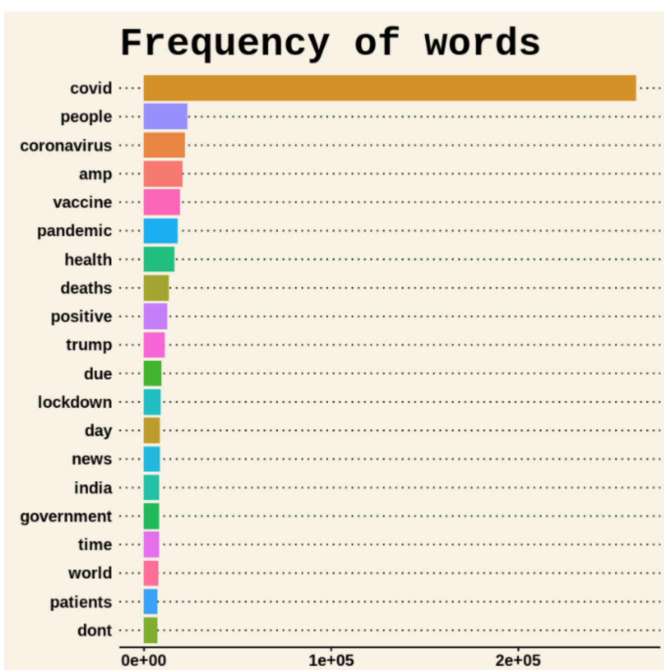
Appendix 4



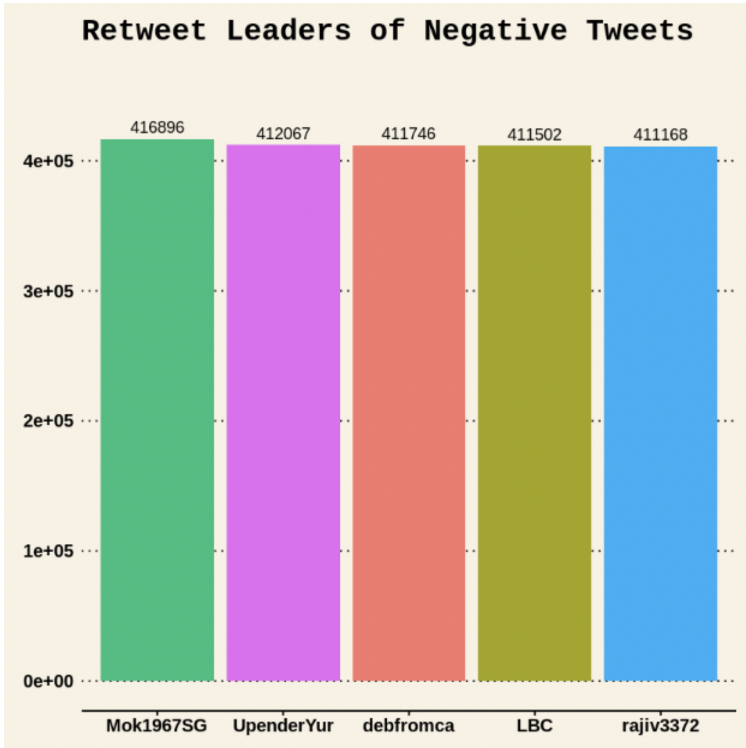
Appendix 5



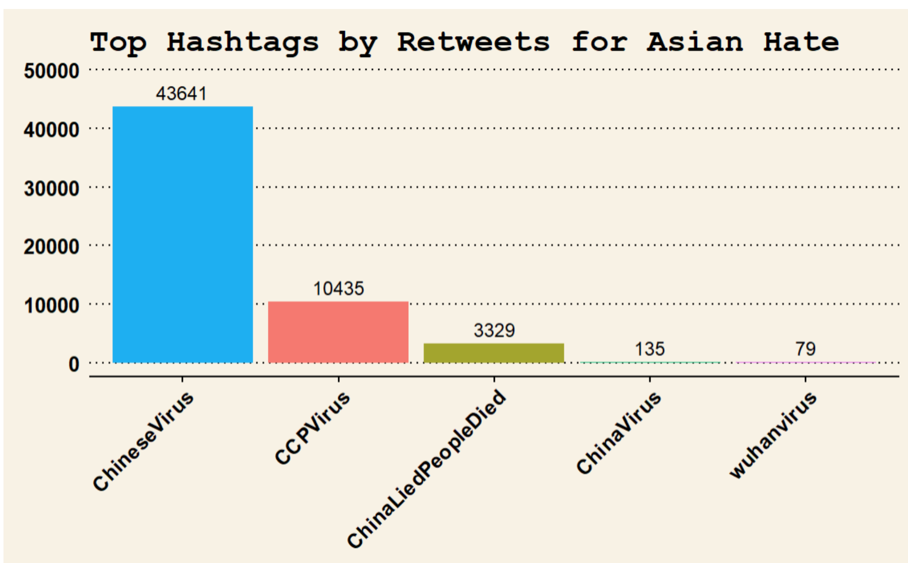
Appendix 6



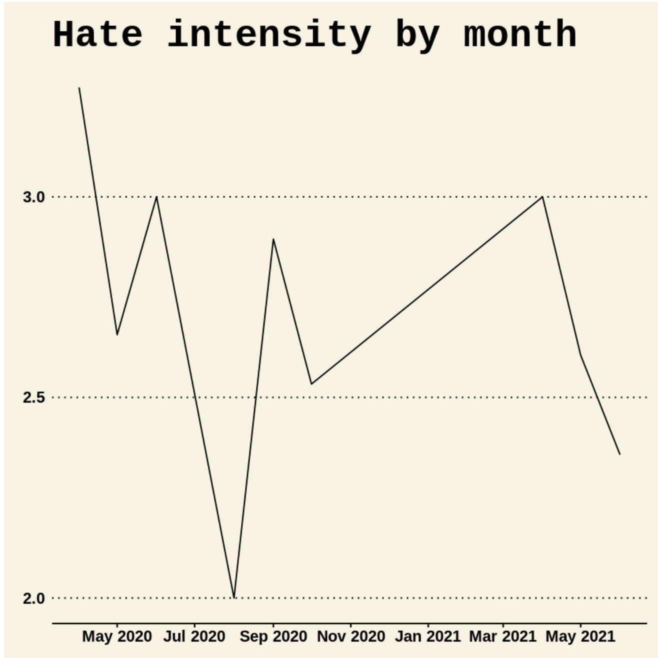
Appendix 7



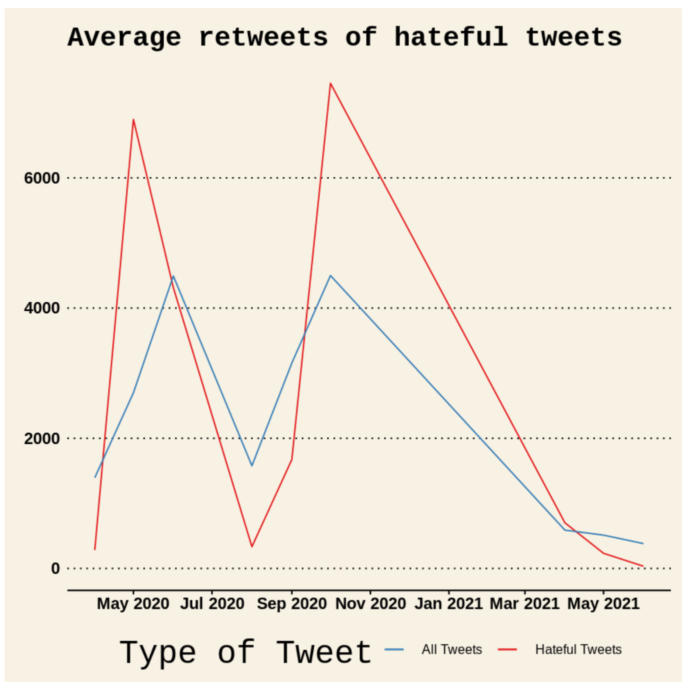
Appendix 8



Appendix 9



Appendix 10



Reference

1. Donaghue, E. (2020, July 2). *2120 hate incidents against Asian Americans reported during coronavirus pandemic*. CBS News. Retrieved February 28, 2023, from https://www.cbsnews.com/news/anti-asian-american-hate-incidents-up-racism/
2. Fallows, J. (2020, March 18). *Trump and the 'Chinese Virus'*. The Atlantic. Retrieved February 28, 2023, from https://www.theatlantic.com/ideas/archive/2020/03/2020-time-capsule-5-the-chinese-virus/622116/
3. He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus: Anti-asian hate and counterspeech in social media during the COVID-19 crisis. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 90–94. https://doi.org/10.1145/3487351.3488324
4. Zhang, Y., Zhang, L., & Benton, F. (2022). Hate Crimes against Asian Americans. *American Journal of Criminal Justice*, *47*(3), 441–461. https://doi.org/10.1007/s12103-020-09602-9
5. *Key facts about Asian Americans, a diverse and growing population*. (2021, April 29). Pew Research Center. Retrieved February 28, 2023, from https://www.pewresearch.org/fact-tank/2021/04/29/key-facts-about-asian-americans/
6. Kim, J. Y., & Kesari, A. (2021). Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic. *Journal of Online Trust and Safety*, *1*(1), Article 1. https://doi.org/10.54501/jots.v1i1.13
7. *2020 FBI Hate Crimes Statistics*. (n.d.). Department of Justice. Retrieved February 28, 2023, from https://www.justice.gov/crs/highlights/2020-hate-crimes-statistics