Vanessa Hu

Professor Kevin Ryan
Quantitative Methods in Linguistics
Spring 2020

Analyzing Sentiment in COVID-19 and China-Related Tweets

1. Abstract

This paper investigates the sentiments of the Twitter community from February to April 2020 in the context of the global coronavirus pandemic, increased discussion of China-related topics with the nation as the earliest epicenter of the outbreak, and resulting anti-Chinese sentiments with blame and controversy surrounding China and ethnically Chinese people. Aggregating a dataset of tweets both related and unrelated to coronavirus and China, the findings indicate a link to more negative sentiment surrounding online dialogue regarding COVID-19 and China-related topics, as well as possible links to how sentiment towards COVID-19 or China may have respectively increased and decreased in positive sentiment over the three-months. This research hopes to better illuminate the shifts in our emotions and attitudes during this global crisis.

2. Introduction

The 2019 novel coronavirus outbreak (COVID-19), declared by the World Health Organization to be a pandemic in early March of 2020, has drastically impacted economies and societies across the world. From governments locking down city, state and national borders, non-essential businesses shutting down, to citizens self-isolating for the foreseeable future, our daily lives have transformed as we face this unprecedented global crisis.

An unfortunate consequence of the resulting fears and anger surrounding COVID-19, which was first identified in Wuhan, China, is rising anti-Chinese rhetoric, linking the blame for the pandemic to ethnic and national identity [1]. This has resulted in the spread of derogatory and misleading terms via social media such as "kungflu" or "Chinese virus" and "over 1700 incident reports of verbal harassment, shunning, and physical assaults" against Asian-Americans in the United States as the wider Asian population become subject to resentment and discrimination [2, 3].

Moreover in its COVID-19 and The Need for Action on Mental Health, the United Nations has emphasized the increased risk for anxiety, depression, and other mental health issues exacerbated by the pandemic [4]. Because with social distancing, people are turning to digital communities more than ever to participate in global dialogue and express uncertainties, fears and anger, this paper thus aims to investigate the sentiment of Tweets having to do with COVID-19, as well as China and Asian-Americans, to better understand the attitudes of online communities in this global crisis.

3. Data

a. COVID-19-TweetIDs Dataset

To find Tweets regarding COVID-19, I utilized the COVID-19-TweetIDs Dataset [5], collecting 100 Tweets per day from February 1, 2020 to April 30, 2020 by randomly selecting one of multiple day files and randomly selecting 100 Tweet IDs from each file. I retrieved the Tweets' metadata in a JSONL file using the Hydrator tool [6].

b. GetOldTweets3

I utilized the GetOldTweets3 Python3 library [7], iterating through February 1 to April 30 to collect 50 Tweets with English as the only search parameter, and 50 Tweets related to China by querying the following keywords*: china, chinese, asia, asian, asian american, chinese american, asian america, chinese america, yellow peril, aapi, chink, chinatown.*

I use the COVID-19-TweetIDs Dataset to supplement GetOldTweets3, which has limited querying functionalities -- and I include Asia-related terms as China-related due to the reality, where "Asians" are equated to a Chinese monolith, which has resulted in the targeting of Asian-Americans of many backgrounds due to anti-Chinese sentiment. The collection starts at February 1, the day the United States confirms its eighth case and the first death outside mainland China is reported in the Philippines [8].

c. Data Processing

i. Sentiment Analysis with VADER

After cleaning all the Tweets by removing hashtags and links using regular expressions,
I use Natural Language Toolkit's VADER [9], a "lexicon and rule-based sentiment analysis tool" for social media, that, for instance, factors in emojis, punctuation, and capitalization for emphasis. In particular, I retrieve its compound score, normalized on a scale of -1.0 to 1.0 from negative to positive sentiment, and scaled up the scores by a factor of 100 for readability.

ii. Tagging COVID-19 and China Keywords

I tag whether Tweets are about COVID-19 and China, or both using regular expressions based on these lists of keywords, of which the COVID-19 list is more strict than the provided Tweet dataset where some collected Tweets were ostensibly not related to COVID. Less commonly seen, and more specific keywords are tagged if there are multiple words found.

Category	**Keyword List**
covid_word	kungflu, chinkvirus, china virus, chinese virus, sars-cov-2, ncov, covid-19, covid, coronavirus, quarantine, corona, pandemic, virus, social distancing, flatten the curve
china_word	yellow-peril, kungflu, chinkvirus, chink, china virus, chinese virus, wuhan, chinese america, chinatown, china, chinese, asian america, asian, asia

Table 1. List of keywords used with regular expressions to tag tweets as related to COVID and/or China; tagged in this order.

After filtering out the 11% of tweets that had been deleted and also non-English tweets from the Tweet ID Dataset, the aggregate Tweet dataset was n = 13,860 tweets.

1. Results

a. Summary Statistics

Of all n = 13,860 tweets, the mean sentiment was 0.64 (s = 45.71). For tweets only related to China, the sentiment averaged -2.84 (s = 47.92). The mean sentiment for *about_covid* tweets was -2.43. For tweets about both COVID-19 and China, the mean sentiment was -16.28, and the mean sentiment for all other tweets was 8.22.

When comparing solely COVID-19 versus non-COVID-19 tweets, the mean sentiment of COVID-19 tweets was -7.97 (s = 45.54), with non-COVID tweets at an average sentiment of 4.21 (s = 44.97). The mean sentiment of tweets having China-related keywords was -7.07 (s = 47.77), whereas the mean of non-China-related tweets was 5.23 (s = 43.805).

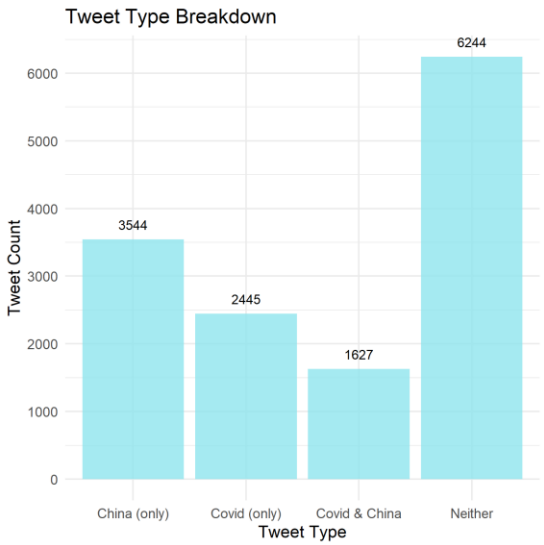

Figure 1. The breakdown of tweet types in the collected dataset of n = 13,860 tweets.

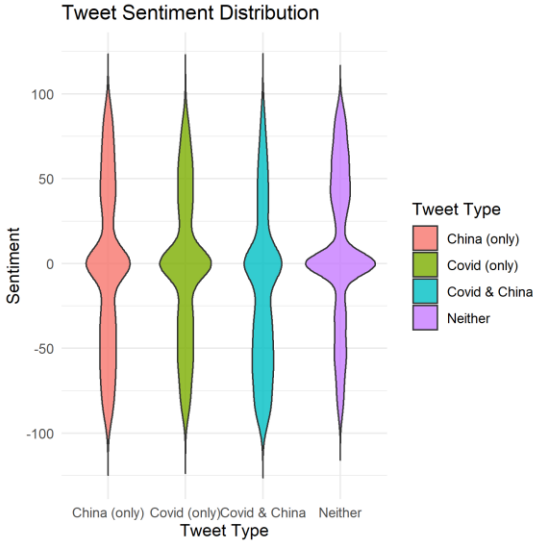

Figure 2. The distributions of VADER compound sentiment scores scaled by 100, broken into four tweet categories.

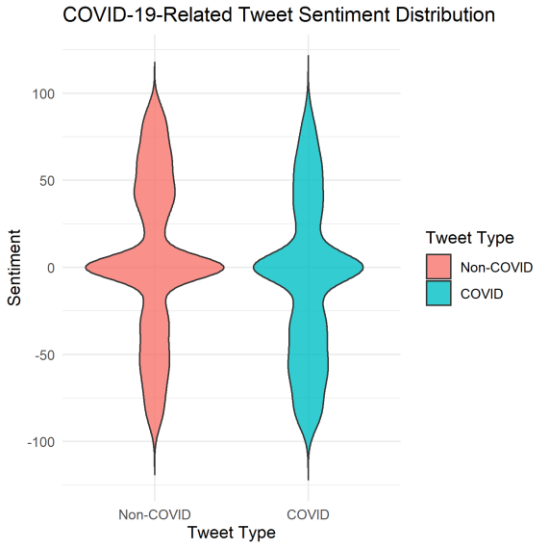

Figure 3. The distributions of sentiment scores based on non-COVID or COVID-related tweets.

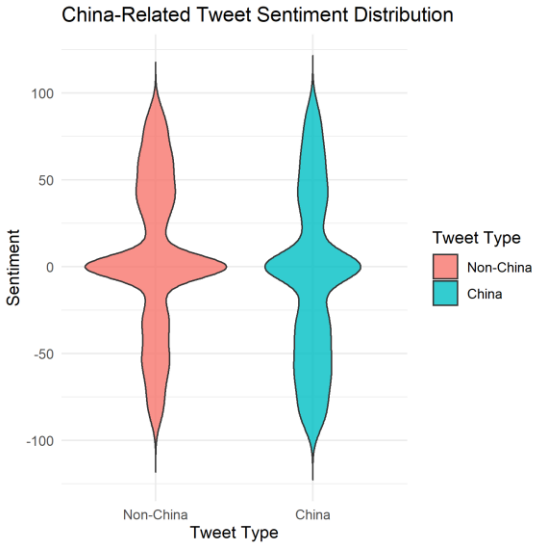

Figure 4. The distributions of sentiment scores based on non-China or China-related tweets.

b. Significance Testing

The data was reshaped to 15,247 observations by merging the covid_word and china_word into one word column and duplicating observations with both, removing repeated entries. Significance tests are based on $\alpha = 0.05$.

i. Simple Regression Models

In three two-sided t-tests for simple linear regression of sentiment dependent on one predictor, all models were found to be significant. China-related tweets had an estimated drop by 0.141 in sentiment *($p < 2e\text{-}16$);* COVID-related tweets had an estimated drop by 0.142 *($p < 2e\text{-}16$)*, and date had a slight increase by 0.0004 *($p = 0.001$)*.

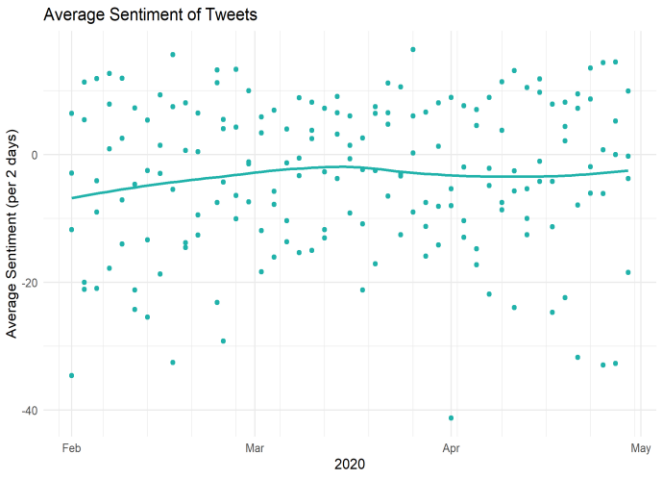

Figure 5. A scatterplot of all tweet sentiments (February 1 to April 30, 2020).

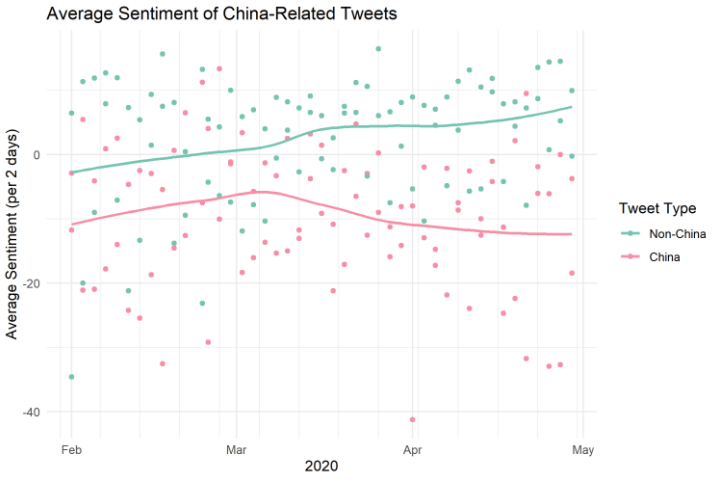

Figure 6. A scatterplot of tweet sentiments of China-related vs. non-China-related tweets.

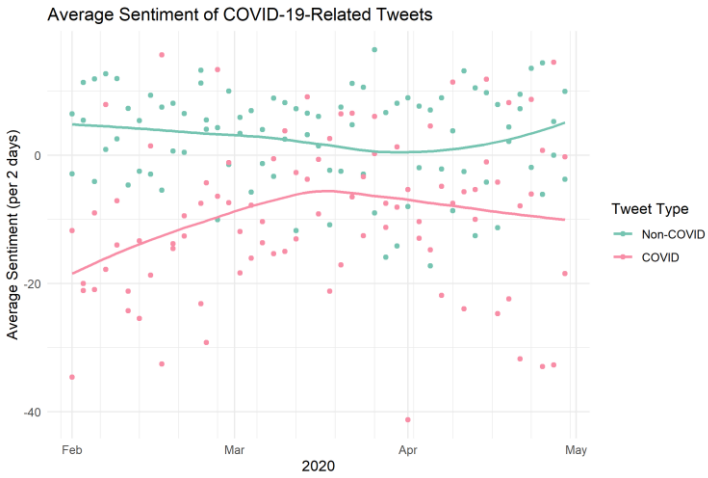

Figure 7. A scatterplot of tweet sentiments of COVID-19-related vs. non-COVID-19-related tweets.

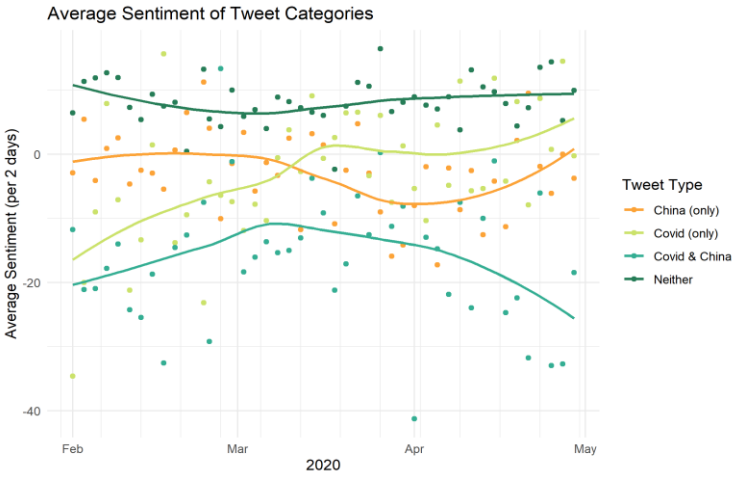

Figure 8. A scatterplot of tweet sentiments, separated by tweet category.

ii. Final Model

1. Individual Predictors

I fitted the data to a linear mixed-effects model using Satterthwaite's method, with sentiment as the response variable contingent on predictors: date, binary variables about_covid or about_china, as well as random intercepts for the corresponding word. The initial intercept estimate is -1.716, essentially neutral based on the scaled -100 to 100 VADER sentiment score range. The date predictor did not significantly influence sentiment ($p = 0.654$), but a Tweet being about COVID-19 (about_covid = 1) significantly decreased sentiment by an estimated -26.17 ($p = 0.002$). Being about China was also significant but seemingly less so, with a p-value close to the 0.05 value ($p = 0.044$) and an estimated sentiment increase by 13.20.

2. Interactions

While date was not itself significant, its interaction with about_covid was, with a slight estimated increase in sentiment by 0.14 (p = 0.002). Like the individual predictors, date's interaction with about_china was significant but less so, an estimated decrease in sentiment by 0.07 (p = 0.042). However, the interactions of both about_covid and about_china (p = 0.159) and of the two along with date (p = 0.158) , did not significantly influence tweet sentiment.

3. Random Effects

By the likelihood ratio test refitting with maximum likelihood, the model with random effects significantly improves upon the simple regression without the intercepts for the keywords ($p = 0.1124, \chi^2 = 6.4276$).

Keyword	Random Intercept	Keyword	Random Intercept
asia	6.829	*corona*	-0.261
covid	5.214	*wuhan*	-0.321
chinatown	2.283	*flatten-the-curve*	-0.651
quarantine	2.072	*yellow-peril*	-0.921
social-distancing	2.035	*kungflu*	-1.309
ncov	1.975	*chinese-america*	-2.216
asia	1.906	*coronavirus*	-2.309
chinese-virus	1.805	*china*	-2.524
covid-19	1.554	*pandemic*	-2.950
chinese	0.791	*virus*	-4.007
sars-cov-2	0.937	*china-virus*	-4.494
chink	-0.230	*asian-america*	-5.206

Table 2. The final list of keywords the random intercepts were based, in descending order.

2. Discussion

a. Sentiments Over Time

The isolated effects of about_china and about_covid with estimated decreases in sentiment, as mentioned in the simple linear regression results and shown in Figures 6 and 7, are echoed in the state of world affairs, as social media users exchange information about increased fears and dire situations such as medical equipment shortages or coronavirus deaths. And, with the increase in anti-Chinese sentiments, and ensuing discussions of such unfounded attitudes and the increased attacks against Asians, China-related tweets corresponding to more negative tweets make sense as well.

However, in the final mixed effects model, while a COVID-related tweet also had an estimated negative effect, China-related tweets had a reported significant positive effect on sentiment. This unexpected result may be a consequence of how about_covid and about_china, while not synonymous, largely overlap. The two are interconnected, for instance, because of the first COVID-19 outbreak occurred in China, and with the increasing criticisms since February until now about China's purportedly inaccurate death toll and suspicions of the virus originating from a Chinese lab [10]. Our significant interaction confirms this – with the interaction of date and about_china leading to an estimated drop in positive sentiment. On the other side, the estimated slight increase in positive sentiment of the date, about_covid interaction may be a signal of how people are becoming accustomed to their new lives under quarantine, and as societies are rallying to overcome the first waves of the virus. So even though time itself is not significant, it can reveal subtle shifts in sentiments in Twitter discussions when put in the context of COVID-19 or anti-Chinese dialogue from February to April 2020.

b. Potential Corresponding Events

While most daily mean sentiments, split by the four categories in Figure 8, ranged between ±20, some were unexpectedly negative. I briefly explore some world events that may be reflected in these numbers [8].

i. <u>February 1:</u> The US confirms its 8th case, Philippines reports first death outside mainland China (COVID tweets, mean -41.54)

ii. <u>February 18:</u> the CDC announces members aboard the Diamond Princess cruise must wait at least fourteen days to return to the U.S.; the director of the Wuchang hospital in Wuhan dies of COVID-19. (COVID/China tweets, mean -45.53).

iii. <u>March 29:</u> The U.S. surpasses 140,000 cases with the most cases of any nation.

iv. <u>April 2:</u> 6.6 million Americans file for unemployment; 91% of Americans must stay at home due to government lockdowns (COVID/China tweets, mean -51.17).

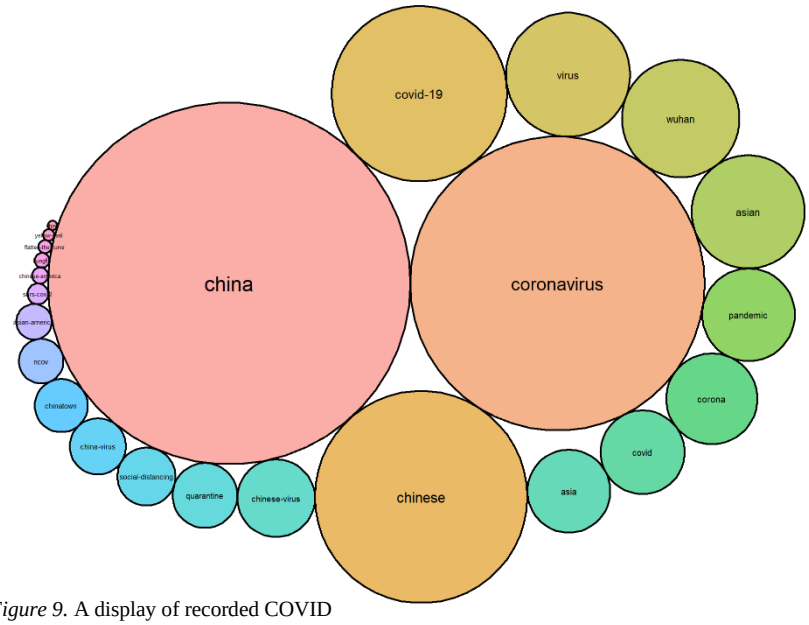

Figure 9. A display of recorded COVID and China keyword frequencies in the tweets.

c. Keywords

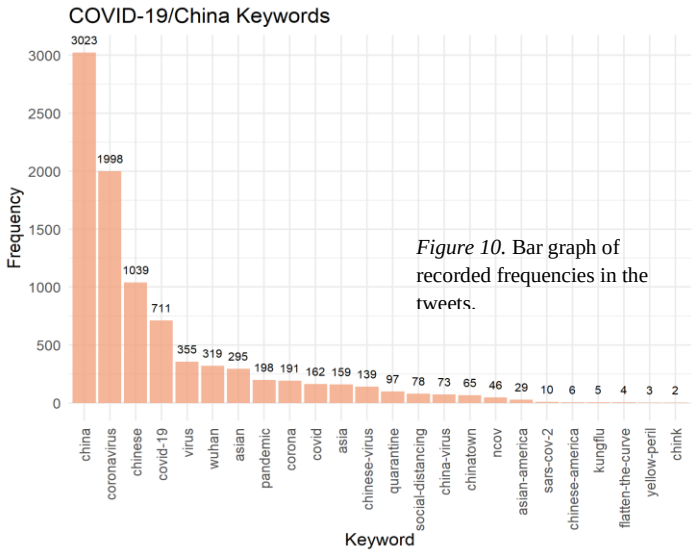

Figure 10. Bar graph of recorded frequencies in the tweets.

Because the mixed model significantly improves the original model, we can briefly explore the random effects of the keywords (Table 2, Figures 9-10*)*. It is interesting to note more specific terms like *asian-america*, and derogatory or less frequent terms such as *kungflu*, *yellow peril*, or *china virus* begin with more negative intercept scores, revealing how such language may link to more negative dialogue. This contrasts words with positive intercepts: more neutral terms such as *quarantine*, *social distancing*, which may be used to explore more positive aspects of the crisis related to at-home activities or social responsibility. The higher intercepts of more general words such as *asia* or *asian*, may also point at how its broader usage is more conducive to less negative dialogue. Finally, amongst the distribution of coronavirus "synonyms," the more official *covid*, *ncov*, *covid-19*, *sars-cov-2* seem to have more less negative intercepts than the more frequent and colloquial *coronavirus*, *corona*, *virus*, which may reveal a higher intensity of negative emotion when using the latter, more frequently used terms.

d. Limitations

i. Tweet Categorization

While I used regular expressions to search for keywords to categorize tweets as COVID and China-related, this does not account for implied contexts; for instance, one tweet states "Didn't realize President Trump was spokesperson for all of Asian America," which would only be tagged as China-related, even though the broader context refers to Trump's comment referring to coronavirus [11]. Like mentioned, about_covid and about_china have a sizeable overlap, and so a more detailed analysis may call for manual, or a machine-learning based topic classification to categorize Tweets to as strictly non-COVID but about China, for instance.

`

ii. Keyword Tagging

In this analysis, tweets were tagged with at most one China keyword and COVID keyword in a cascading fashion based on even though tweets may mention multiple of each, such as in one tweet, "What does COVID-19 'social distancing' look like? ✅ Stay home unless absolutely necessary…" [12]. With more rigorous word-tagging of all keywords, the random intercepts for keywords may then represent the data more accurately and better reveal what specific language corresponds to more negative or positive sentiments surrounding COVID-19 and China.

iii. VADER

While VADER is suited for social media text analysis, accounting for degree modifiers, acronyms and other features, its rule-based setup may still be limited, as some machine learning models (LSTM) have performed better than VADER [13]. Another possibility is using a mixed ML approach to improve accuracy [14].

iv. Future Explorations

Along with the above mentioned routes, a more focused analysis to detect language correlated with risk for mental health issues, or specifically focusing on language surrounding anti-Asian hate speech and crimes, can further reveal the social and emotional impacts of COVID-19 reflected in our online exchanges.

3. Conclusion

This paper aims to explore the sentiments of the Twitter community over a three-month period from February to April 2020, with the escalation of the global COVID-19 pandemic. The results reveal how online discussion during this period may correlate to more negative sentiment surrounding online dialogue regarding COVID-19 crisis and China-related topics. This finding can be an indicator of the reverberations of the global pandemic on our emotional states, with a reported 45% increased prevalence of distress in the United States and higher risk for worsened mental health [4]. It can also reveal the more negative attitudes towards China, which has been criticized for mishandling and concealing information about the pandemic.

Furthermore, it can reveal the alarming increase in xenophobic, anti-Asian harassment and assaults, with a reported 50% rise in news articles regarding COVID-19 and such discrimination from February to March 2020 [12]. The results also point to more specific terms used that can play into our attitudes discussing COVID-19, while also revealing the difficult task of separating China-related and COVID-related tweets because of how linked the two topics are. Thus, by exploring the language of social media users on Twitter, this investigation can hopefully serve as launching point to gain insight into more lasting impacts on societal sentiments and perspectives during this global crisis.

4. Acknowledgements

Thank you to Professor Kevin Ryan and my Teaching Fellow Ethan Wilcox, for guiding me through my first quantitative research project!

5. References

[1] C. Timberg and A. Chiu, “As the coronavirus spreads, so does online racism targeting Asians, new research shows,” *Washington Post,* Apr. 08, 2020. www.washingtonpost.com/technology/2020/04/08/coronavirus-spreads-so-does-online-racism-targeting-asians-new-research-shows/.

[2] “STOP AAPI HATE,” *Asian Pacific Policy and Planning Council*. http://www.asianpacificpolicyandplanningcouncil.org/stop-aapi-hate/.

[3] S. Tavernise and R. A. O. Jr, “Spit On, Yelled At, Attacked: Chinese-Americans Fear for Their Safety,” *The New York Times*, Mar. 23, 2020.

[4] “UN leads call to protect most vulnerable from mental health crisis during and after COVID-19,” *UN News*, May 13, 2020. https://news.un.org/en/story/2020/05/1063882.

[5] E. Chen, *echen102/COVID-19-TweetIDs*. GitHub, 2020. https://github.com/echen102/COVID-19-TweetIDs.

[6] *Hydrator*. Documenting the Now. GitHub, 2020. https://github.com/DocNow/hydrator.

[7] D. Mottl, *Mottl/GetOldTweets3*. 2020. https://github.com/Mottl/GetOldTweets3.

[8] C. Kantis, S. Kiernan, and J. Socrates Bardi, “Timeline of the Coronavirus,” *Think Global Health*. https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus.

[9] C. J. Hutto, *cjhutto/vaderSentiment*. GitHub, 2020. https://github.com/cjhutto/vaderSentiment.

[10] “Trump stands by China lab origin theory for virus,” *BBC News*, May 01, 2020. https://www.bbc.com/news/world-us-canada-52496098.

[11] NBC Asian America, “Didn't realize President Trump was spokesperson for all of Asian America.” *Twitter*, May 12, 2020. [Online]. Available: https://twitter.com/NBCAsianAmerica/status/1260336965158912001.

[12] Brown, K., “What does COVID-19 "social distancing" look like?...” *Twitter*, March 21, 2020. [Online]. Available: https://twitter.com/oregongovbrown/status/1241573048458686469?lang=en.

[13] A. R, A. Patil, S. Rayar, and V. K M, “Comparison of VADER and LSTM for Sentiment Analysis,” *International Journal of Recent Technology and Engineering*, vol. 7, no. 6S, Mar. 2019, [Online]. Available: https://www.ijrte.org/wp-content/uploads/papers/v7i6s/F03040376S19.pdf.

[9]

[14] V. D. Chaithra, “Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 4452–4459, Oct. 2019.

[15] S. Mullis and H. Glenn, “New Site Collects Reports Of Racism Against Asian Americans Amid Coronavirus Pandemic,” *NPR*, Mar. 27, 2020. https://www.npr.org/sections/coronavirus-live-updates/2020/03/27/822187627/new-site-collects-reports-of-anti-asian-american-sentiment-amid-coronavirus-pand.

[16] CDC, “Coronavirus Disease 2019 (COVID-19),” *Centers for Disease Control and Prevention*, Feb. 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/faq.html.