Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter

Lizhou Fan1,2 | Huizi Yu2,4 | Zhanyuan Yin3,4

1Program in Digital Humanities, University of California, Los Angeles, California, USA
2Department of Statistics, University of California, Los Angeles, California, USA
3Department of Mathematics, University of California, Los Angeles, California, USA
4Department of Economics, University of California, Los Angeles, California, USA

Abstract
As the COVID-19 pandemic has unfolded, Hate Speech on social media about China and Chinese people has encouraged social stigmatization. For the historical and humanistic purposes, this history-in-the-making needs to be archived and analyzed. Using the query “china+and+coronavirus” to scrape from the Twitter API, we have obtained 3,457,402 key tweets about China relating to COVID-19. In this archive, in which about 40% of the tweets are from the U.S., we identify 25,467 Hate Speech occurrences and analyze them according to lexicon-based emotions and demographics using machine learning and network methods. The results indicate that there are substantial associations between the amount of Hate Speech and demonstrations of sentiments, and state demographics factors. Sentiments of surprise and fear associated with poverty and unemployment rates are prominent. This digital archive and the related analyses are not simply historical, therefore. They play vital roles in raising public awareness and mitigating future crises. Consequently, we regard our research as a pilot study in methods of analysis that might be used by other researchers in various fields.

KEYWORDS
coronavirus, COVID-19, hate speech, pandemic, Twitter

1 | INTRODUCTION
As the COVID-19 pandemic has unfolded,1 many local authorities have locked down cities through social distancing orders to slow down the speed of virus transmission. Consequently, daily routines and social norms have changed and social media platforms, such as Twitter, have emerged as one of the primary places where people create and exchange information relating to the virus.

Although people are physically distant, the information they have shared through social media streams during this global health and information crisis is influential (Xie et al., 2020). Some of the posts and commentaries being shared online, however, reflect or even amplify contemporary social stigmatization, discrimination and outright Hate Speech, especially directed at China and the Chinese people. To understand their potential to influence public opinion and behavior, these controversial tweets (Twitter posts) need to be documented and analyzed in a timely manner. In early February 2020, therefore, we began capturing and analyzing tweets related to the coronavirus and China, with a specific focus on Hate Speech.

In this paper, we present an initial analysis of the resulting archive of tweets together with the analytical methods we have used. Our broader goal has been to examine the role that an archive could play in informing the
public, educating the society, and preventing or smoothing future crises (Chew & Eysenbach, 2010). Furthermore, we are methodologically interested in identifying effective processes, from archiving to analyzing, that can be used in real time during a rapidly evolving crisis of this magnitude.

2 | STUDY DATA AND METHODS

2.1 | Data collection

2.1.1 | Building a database of tweets

We have been documenting tweets related to China since the start of the COVID-19 pandemic. “China” and “Coronavirus” are words that have been associated in the public mind ever since the Chinese authorities informed the World Health Organization (WHO)'s China office of the occurrence of pneumonia cases with unknown cause in Wuhan, China on December 31, 2019 (Ravelo & Jerving, 2020). During the critical period between when the coronavirus emerged to the ensuing outbreaks, we have queried the Twitter API using "china+and+coronavirus" and have collected 3,457,402 English-language tweets (key tweets).

2.1.2 | External data collection

We have also been obtaining coronavirus cases information in the U.S. from the dataset provided by John Hopkins University. To further detect whether there might be any systematic trend in Hate Speech within groups of states that share certain characteristics such as demographic composition, we have extracted supplementary information from the U.S. Census Bureau, the U.S. Bureau of Labor Statistics and the Federal Election Commission (Eileen & Leamon, 2017).

2.2 | Visualizing trends as a method for preliminary analysis

In this project, we utilize data visualization as an exploratory research method and have been able to effectively identify patterns and correlation among variables (e.g., new cases, number of hate tweets) by using bar-chart, gradient maps and word cloud.

2.2.1 | Hashtag trends

Word count tables and Word Clouds are straightforward representations of discourse. We have generated the word cloud in Figure 1 as a demonstration of the overall trend of tweets according to their hashtags, creating a summary of user input topics.

Among the top frequently used hashtags in the archive, several main categories can be discerned, including location, person, organization, and abstract concept. One clear trend is the occurrence of locations, as shown in Table 1 and Figure A1 in Appendix III. “wuhan”, with more than 24,000 occurrences, is the top hashtag used in this archive. Other locations used as hashtags include Italy, USA, Hong Kong, and Hubei (of which Wuhan is the capital). This preliminary finding confirms our intuition that analyzing the trend and speech discourse together with demographics could be productive.

In terms of contextual information, two of the frequently used hashtags, “chinesevirus” and “wuhanvirus”, are associated with discriminatory comments. Both hashtags are violations of the WHO's convention for naming new human infectious diseases (WHO, 2015). This preliminary finding suggests that Twitter users have often used discriminatory speech regarding specific groups, thus leading us to conduct further investigation using lexicon-based information extraction methods.

2.2.2 | The trend of tweets

Figure 2 provides an aggregated histogram and line plot that analyze the relationship between the number of new cases versus the number of key tweets per day. We can observe an overall increasing trend for both the number of new cases and the number of key tweets over time. Interestingly, we can also observe a substantial surge in the number of key tweets from March 16 to March 19. We believe this occurrence might be caused by the burgeoning of confirmed cases in the U.S. and increasing media attention.

2.3 | Data pre-processing

After removing stop words and ignoring non-textual strings, we use lexicon-based methods for the primary data wrangling of both Hate Speech detection and aspect-based emotion scoring. For Hate Speech detection, we use the Hatebase dictionary, which includes thousands of discriminative words. For emotion extraction, we use an emotion lexicon created using crowdsourcing with Mechanical Turk (Mohammad & Turney, 2010).

For each tweet, we tabulate the number of words related to each element in the emotion dictionary and then generate 10 variables that include sentiments (negative, positive) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and dividing them by the
tweet’s word count to normalize. For tweets with words defined as hatewords, we label them as “Hate Speech”.

To identify location information, we use the *carmen* Python package based on geocoding tools (Dredze, Paul, Bergsma, & Tran, 2013). We extracted location information for 1,401,374 tweets from the archive and found that 40.53% of them are from the U.S.

Finally, we average the normalized sentiment and emotion variables by date and state to create the aggregated datasets. Each dataset also contains the total number of key tweets and the percentage of Hate Speech on the aggregated levels. For the state-level dataset, we further introduce characteristic variables *PoliticalParty*, *AsianPercentage*, *UnemploymentRate*, and *PovertyRate*.

### 2.4 Statistical methods

#### 2.4.1 Decision tree classifier

Decision tree is a widely used machine learning technique that uses a tree-like model for making decisions and analyzing possible consequences. The target of building a decision tree is to minimize the entropy of each leaf so that it contains either only Hate Speech or Non-Hate Speech. Using Gini Impurity, the decision tree model could also return the importance of features, from which we can identify the best Hate Speech indicator.

#### 2.4.2 Network methods

Another way to analyze and visualize Hate Speech behavior across each state is to use network methods. We can consider states to be nodes in a network and use methods such as modularity and centrality to evaluate the network. Network modularity is a measurement to show network homophily because similar nodes tend to connect with each other. If we assign groups to the states using some criterion (e.g., poverty), we can calculate the modularity of the network to show whether under such a criterion, states that are similar have homogeneity in Hate Speech behavior. Network centrality is a measurement to describe node property. Different centrality
measurements can describe different Hate Speech behavior of each state.

3 | RESULTS

3.1 | Analyzing hate speech and emotions

To better understand the relationship between emotions and Hate Speech, we aggregate the negative sentiments and observe a similar trend in the negative sentiment score and the number of Hate Speech over time, as shown in Figure 3.

To further analyze the relationship between emotions and Hate Speech, we implement a binary decision tree classifier to create a model that can infer whether a tweet is Hate Speech (negative class stands for Hate Speech). The decision tree model is built based on the eight emotion features in 3,461,929 positive samples and 25,467 negative samples (Train Set).

Based on Table 2 and Table 3, we believe that with a 100% specificity, our decision tree classifier could successfully detect all Hate Speech defined by the lexicon-based approach. With similar precision, our model's prediction of a tweet belonging to Non-Hate Speech is definite. Our negative predictive value is 87.87%, thus we can conclude that our model is proficient in detecting Hate Speech.

Figure A2 in Appendix III provides an overview of the generated tree, from which we can also obtain the contributions of features in detecting Hate Speech.

In Table 4, based on the decision tree, we extract the feature importance of each emotion and conclude that surprise and fear are the two important Hate Speech indicators. As shown in Figure A2 in the Appendix IV, although the root node uses joy as a splitting criterion, it
makes little contribution to finding Hate Speech.\textsuperscript{13} The actual nodes that contribute the most to detecting Hate Speech are shown in Figure 4; thus, tweets satisfying all

---

**FIGURE 3** Number of New Cases per day in the U.S. (Blue Bars) and Sentiment and Hate Speech Trends

**TABLE 2** Confusion Matrix of the Test Set

| Predicted | Actual |
|-----------|--------|
|           | Positive | Negative |
| Positive  | 691877   | 0         |
| Negative  | 610      | 4418      |

**TABLE 3** Confusion Matrix Measurements

| Measurements | Value |
|--------------|-------|
| Accuracy     | 99.91%|
| Sensitivity  | 99.91%|
| Specificity  | 100.00%|
| Precision    | 100.00%|
| NPV          | 87.87%|

*Note: NPV is Negative Predicted Value.*

**TABLE 4** Importance of Each Emotional Feature in Picking Hate Speech

| Emotion    | Feature Importance |
|------------|--------------------|
| Surprise   | 28.28%             |
| Fear       | 22.78%             |
| Anticipation| 13.43%            |
| Anger      | 10.99%             |
| Disgust    | 7.80%              |
| Trust      | 7.77%              |
| Sadness    | 6.83%              |
| Joy        | 2.11%              |
three splitting criteria are likely to be Hate Speech. Moreover, tweets satisfying attributes demonstrated in Table 5 are likely to be Hate Speech.

3.2 Analyzing hate speech and demographics

Using the aggregated state-level dataset, we compare the number of Hate Speech occurrences against the number of confirmed cases of COVID-19 by state. The gradient maps show a high correlation between the two, suggesting that shared characteristics across states might influence the number of Hate Speech occurrences and confirm cases simultaneously. To further investigate and visualize homogeneity between states and factors associated with Hate Speech, we build a network model and conduct modularity and centrality analyses.

According to the steps and conditions\textsuperscript{14} specified in Figure 6a, we visualize the resulting network in Figure 6b.

3.2.1 Network modularity analysis

The modularity of our network, $Q$, indicates whether our grouping criteria, for example state poverty rate, are informative in revealing each state’s Hate Speech behavior:
\[ Q = \sum_{i=1}^{r} (e_i - a_i^2) \]  

(1)

where \( e_r \) and \( a_i \) are the proportion and the expected proportion of similar states pairs within each group. If \( Q > 0 \), the similarity of states within groups is comparatively larger than that of between groups, which indicates that our method of splitting the groups is informative in revealing each state's different Hate Speech behavior.

We analyze the effects of four grouping criteria on Hate Speech behavior: political party, poverty rate, Asian percentage, and unemployment rate. For political party, we group the states by Republican states and Democratic states according to the 2016 Presidential Election Results. For the other factors, we group all the states into quadrisections. According to the division, we calculate the modularity of each criterion (Table 6). We conclude that the poverty rate and the unemployment rate of each state are informative factors in studying the Hate Speech behavior of each state.

### 3.2.2 Network centrality analysis

Centrality of a network is a measurement to describe the importance of each node. In our case, we are interested...
in the degree centrality and page rank centrality of each node. Degree centrality vector, \( \sigma_c \), can be used to measure the number of states that have similar Hate Speech behavior with another state:

\[
\sigma_c = \frac{A1}{|V|}
\]

(2)

where \( A \) means the adjacency matrix; \( 1 \) means the all-one vector; \( |V| \) means the number of nodes. In our case, higher degree centrality means the state’s Hate Speech behavior is common among all states. Otherwise, the state may have a unique Hate Speech behavior (either too high or too low).

Compared with degree centrality, page rank centrality, \( \sigma_p \), gives high centrality to nodes that connected with high centrality notes, that is page rank centrality groups “potentially” similar states in a more extreme way. Therefore, for those states that have low page rank centrality, their Hate Speech behaviors have higher levels of distinctions. For the centrality of each state:

\[
\sigma_p,i = a \sum_{j=1}^{n} A_{ij} \sigma_p,j + \beta
\]

(3)

where \( A \) refers to the adjacency matrix, \( k_j \) refers to the degree of node \( j \), and \( a, \beta \) are constants.

From Figure 7a,b we can discern the states that perform “uniquely” in Hate Speech. We can observe that California, New York, Florida, Texas, Pennsylvania, Illinois (from degree centrality), and Massachusetts, Rhode Island, North Dakota (from page rank centrality) have low centrality compared with other states. Combined with Figure 5, we observe that those states either have extremely high confirmed cases (such as New York, Massachusetts, Illinois, California) or have extremely low confirmed cases (such as North Dakota). Therefore, further study might usefully compare the Hate Speech behaviors of different states.

\section*{5 | DISCUSSIONS AND FUTURE WORK}

\subsection*{5.1 | Related work and our focus}

We realize that our dataset could be a good complement to the “COVID-19-TweetIDs” dataset, which has a more general scope (Chen, Lerman, and Ferrara, 2020). We also notice that scholars from various fields are analyzing COVID-19-related Twitter datasets by using topic modeling methods and characterizing individual user behaviors (Chen, Yang, et al., 2020). Compared to these investigations, our scope is more comprehensive, encompassing archival, analytical, and methodological processes, and including utilizing machine learning and network methods for visualization and analysis.

\subsection*{5.2 | Limitation and discussion of data scope}

With the Twitter API’s limitations on continuity of searching, we have several data gaps. Since we either normalize the variables or analyze correlations, we expect this discontinuity to have minimal influence on our analysis. While the number of key tweets quickly drops then slowly increases after March 19th, this drop may be due to delayed adoption of the term “COVID-19” in place of “coronavirus”, thus potentially diverting much of the information stream.

\subsection*{5.3 | Limitation and improvement in information extraction}

Restricted by the scope of the dictionaries, lexicon-based methods likely underestimate the sentiment/emotion score by failing to identify malformed words and colloquial expressions (e.g., gr8, RIP) (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Its also not sufficiently sophisticated to capture overall context and complex syntax scenarios (e.g., ironic language use). To combat these limitations and extract more precise indicators for Hate Speech, we plan to create a manually labelled dataset and employ other supervised learning models such as a support vector machine.
ACKNOWLEDGMENTS
We thank Prof. Anne Gilliland for supervising our research and providing generous help. We also thank Prof. Todd Presner, Prof. Ashvin Gandhi, and Dr. Heather Zinn-Brooks for their inspiring suggestions.

ENDNOTES
1 By April 26, 2020, 2,810,325 confirmed cases and 193,825 deaths had been reported as shown by WHO COVID-19 dashboard at https://covid19.who.int/
2 January 31, 2020 to April 7, 2020, approximately 40 days
3 https://systems.jhu.edu/research/public-health/ncov/
4 https://www.census.gov/programs-surveys/acs; https://www.census.gov/programs-surveys/saipe.html
5 https://www.bls.gov/lau/data
6 Twitter is a US-based social media platform, so the apparent bias is not surprising. Because most of the tweets in this archive are in English and most of the users are located in the U.S., we associate the number of tweets with the number of new cases in the U.S.
7 https://hatebase.org/
8 Examples of Hate Speech are presented in Appendix I Table A1
9 Some unincorporated territories of the U.S. such as Guam and USVI are dropped due to insufficient sample size.
10 Descriptive results for the state level dataset are presented in Appendix II Table A2
11 For a set of items with J classes, where \( p_i \) is the fraction of items labeled with class \( i \) in the set, the Gini Impurity of the class is \( I_G(p) = 1 - \sum_{i=1}^{J} p_i^2 \). The decision tree algorithm chooses one feature that could minimize Gini Impurity and uses it as the decision criterion.
12 Fear, Anger, and Disgust could be directed at haters as well as Hate Speech, and Sad indicates negative emotions which are related but not limited to hate. For the convinence of the first and overall result, we aggregate them together, while distinguish them in later analyses.
13 The Joy node only separates out a large amount of Non-Hate Speech (the left branch), and a significant amount of Non-Hate Speech remain unidentified.
14 We normalize the number of Hate Speech tweets by dividing them into the population of each state/district.
15 Here, \( e_i = \frac{1}{n}\sum_{l=1}^{n} \sum_{j=1}^{n} A_{ij} \delta_{g_i, r} \delta_{g_j, r} \) and \( a_i = \frac{1}{m}\sum_{r=1}^{m} k_i \delta_{g_i, r} \), where \( m \) stands for the number of edges, \( r \) stands for the number of groups, \( A \) stands for the adjacency matrix, \( \delta \) stands for the Kronecker delta, and \( k \) stands for the degree of node \( i \), and \( g \) stands for group \( i \).
16 Similar states pair stands for two states that share the same or similar feature. For example, both states are Democratic States; both states have high poverty rate.
17 https://developer.twitter.com/en/docs/tweets/search/overview/s

REFERENCES
Chen, E., Lerman, K., and Ferrara, E. (2020). Covid-19: The first public coronavirus twitter dataset. arXiv preprint arXiv:2003.07372.
Chen, L., Yang, T., Luo, J., Lyu, H., and Wang, Y. (2020). In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19. arXiv preprint arXiv:2004.10225.
Chew, C., and Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one, 5(11), e14118.
Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. AAAI Workshop - Technical Report. 20–24.
Eileen, J., Leamon, J. B. (2017). Federal elections 2016 election results for the U.S. president, the U.S. senate and the U.S. house of representatives. Technical Report, Federal Election Commission.
Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon (pp. 26–34).
Ravelo, J. L. and Jerving, S. (2020). Covid-19 — A timeline of the coronavirus outbreak. Retrieved from https://www.devex.com/news/covid-19-a-timeline-of-the-coronavirus-outbreak-96396
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267–307.
Xie, B., He, D., Mercer, T., et al. (2020). Global health crises are also information crises: A call to action. Journal of the Association for Information Science and Technology, 1–5.
World Health Organization. (2015). WHO issues best practices for naming new human infectious diseases. WHO. Retrieved from https://www.who.int/mediacentre/news/notes/2015/naming-new-diseases/en/.

How to cite this article: Fan L, Yu H, Yin Z. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. Proc Assoc Inf Sci Technol. 2020;57:e313. https://doi.org/10.1002/pra2.313
APPENDIX I.

TABLE A1  Examples of Hate Speech on Twitter

| Tweet ID  | Date       | Tweet                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-----------|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 12248490476099**** | 2/1/2020  | After SARS, Bird Flu, Swine flu and all that, how could something like coronavirus be allowed to spread for so long that the first person to get it is elusive? China do not f**k around, one c**t starts coughing the whole province is in masks in minutes.                                                                                          |
| 123718116467216**** | 3/10/2020 | What do we call this China Coronavirus then what it is and you Idiot’s call us Racist I put it where it should belong at the feet of the China Government we did not get this from another Country it came from China now we are all in it’s path so keep on being ignorant about this                                                                                             |
| 123795213637523**** | 3/12/2020 | If they diagnosed all these people with the coronavirus in China 2 months ago then why TF DID THEY EVEN LET THEM LEAVE CHINA, WHAT IDIOT LET THESE PEOPLE GO AROUND THE WORLD AND SPREAD IT                                                                                                                                                                                                                     |
| 123815218206889**** | 3/12/2020 | Well, yes, the novel coronavirus COVID 19 did, in fact, originate from Wuhan, China probably from filthy Commie Ch*nk’s open air meat markets the Wuhan Virus is a foreign virus. Call it the ChineseVirus, perhaps. And, out of spite, Americans should boycott Chinese stuff.                                                                                   |
| 123815202117115**** | 3/12/2020 | Down £5 k on my investments because some mad c**t eat a bat in China and started the coronavirus. Mad ting                                                                                                                                                                                                                                                                                                                                                       |
| 124076900418213**** | 3/19/2020 | Anyone who thinks China the pestilence nation is a leader in this is an idiot China covered this up for at least 2 months and let it fester. If anything they declared biological war on the planet. FoxNews                                                                                                                                                                                                                     |
| 124255771854143**** | 3/24/2020 | Anyone who believes Russia's numbers is an idiot. And I do not believe China's numbers either. Bullshit COVID19 coronavirus                                                                                                                                                                                                                                                                                                                                                       |

Note: Hate Speech occurrences are marked red with parts masked for the sake of the scholarly audience; the last four digits of Tweet IDs are masked for privacy reasons.

APPENDIX II

TABLE A2  Descriptive Statistics of State Level Data

|                  | Mean     | 95% CI Upper |
|------------------|----------|--------------|
| State            |          |              |
| Total Tweets (Key word containing) | 13466.80 | 7995.17      |
| Hate Speech Percentage | 0.00594 | 0.00520      |
| Emotion          |          |              |
| Anger Score      | 0.01905  | 0.01882      |
| Anticipation Score | 0.02252 | 0.02234      |
| Disgust Score    | 0.01635  | 0.01611      |
| Fear Score       | 0.02708  | 0.02670      |
| Joy Score        | 0.00882  | 0.00872      |
| Sad Score        | 0.02432  | 0.02393      |
| Surprise Score   | 0.01710  | 0.01691      |
| Trust Score      | 0.02763  | 0.02739      |
| Positive Score   | 0.03213  | 0.03193      |
| Negative Score   | 0.05219  | 0.05158      |
| Demographic      |          |              |
| Asian Percentage | 0.04195  | 0.02661      |
| Unemployment Rate | 0.03562 | 0.03318      |
| Poverty Rate     | 0.12864  | 0.12085      |
| Political Party  |          |              |
| Democratic       | 0.38     |              |
| Republican       | 0.62     |              |
FIGURE A1  Top five hashtags per day and their numbers of occurrences

APPENDIX IV: DECISION TREE CLASSIFIER FOR Distinguishing Hate Speech