Analysis of COVID-19 Offensive Tweets and Their Targets

Song Liao
Clemson University
Clemson, SC, USA

Ebuka Okpala
Clemson University
Clemson, SC, USA

Long Cheng
Clemson University
Clemson, SC, USA

Mingqi Li
Clemson University
Clemson, SC, USA

Nishant Vishwamitra
University of Texas at San Antonio
San Antonio, TX, USA

Hongxin Hu
University at Buffalo, The State University of New York
Buffalo, NY, USA

Feng Luo
Clemson University
Clemson, SC, USA

Matthew Costello
Clemson University
Clemson, SC, USA

ABSTRACT

During the global COVID-19 pandemic, people utilized social media platforms, especially Twitter, to spread and express opinions about the pandemic. Such discussions also drove the rise in COVID-related offensive speech. In this work, focusing on Twitter, we present a comprehensive analysis of COVID-related offensive tweets and their targets. We collected a COVID-19 dataset with over 747 million tweets for 30 months and fine-tuned a BERT classifier to detect offensive tweets. Our offensive tweets analysis shows that the ebb and flow of COVID-related offensive tweets potentially reflect events in the physical world. We then studied the targets of these offensive tweets. There was a large number of offensive tweets with abusive words, which could negatively affect the targeted groups or individuals. We also conducted a user network analysis, and found that offensive users interact more with other offensive users and that the pandemic had a lasting impact on some offensive users. Our study offers novel insights into the persistence and evolution of COVID-related offensive tweets during the pandemic.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Networks** → **Social media networks**.

KEYWORDS

COVID-19, Twitter, Offensive tweets

ACM Reference Format:
Song Liao, Ebuka Okpala, Long Cheng, Mingqi Li, Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Matthew Costello. 2023. Analysis of COVID-19 Offensive Tweets and Their Targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599773

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599773

1 INTRODUCTION

The COVID-19 pandemic caused a rise in the number of users of social media platforms like Twitter for information acquisition, communication, and entertainment. Although social media platforms serve as sources of information about the pandemic, they also enable the rapid spread of misinformation (*e.g.*, false claims, rumors, and conspiracy theories) and toxic content (*e.g.*, cyber-harassment) related to COVID-19. Misinformation spurred an increase in the attribution of blame for the virus's origin to China; in turn, this led to a rise in anti-Asian rhetoric and hate [47]. Physical incidents directed towards Asian communities have likewise increased, particularly after former-President Donald Trump publicly called the virus the "Chinese virus" on Twitter. Following these derogatory statements, the number of hateful incidents targeting Asian communities rose [11, 21].

Due to the offline implications of online offensive speech related to COVID-19, it is important to characterize offensive online contents and the users responsible for disseminating them. Existing works on cyber-hostility related to the COVID-19 pandemic have studied the evolution of Sinophobic content [47], but only focused on brief periods of the pandemic. Moreover, existing works mainly studied hate related to COVID-19 specifically towards Asians without considering other targets during the COVID-19 pandemic. Caleb *et al.* [63], for instance, studied the evolution of anti-Asian hate on Twitter. In this work, we seek to conduct a long-period analysis to understand the changes and persistence of COVID-related offensive tweets beyond the peak of the pandemic.

We aim to answer the following research questions. RQ1: How did COVID-related offensive speech evolve on Twitter throughout the pandemic? RQ2: Which groups or individuals were discussed and targeted by offensive rhetoric when discussing COVID-19? What are the characteristics of offensive tweets towards different targets? RQ3: For offensive users, how did they change in terms of posting offensive tweets during the pandemic, and what are the possible factors that influenced offensive users?

Our contributions. We make the following contributions and offer the following findings:

- We collected a large-scale dataset (named `COVID-OFFENSE`) with different types of keywords related to COVID-19. Our dataset contains over 747 million tweets collected between

January 1, 2020 and June 30, 2022. We labeled 1,679 offensive tweets with various types of COVID-19 related offensive language. We fine-tuned a BERT model using our annotated tweets and used the fine-tuned model for tweet classification. Our COVID-OFFENSE dataset is available at https://github.com/CUSecLab/2023-KDD-Covid-Twitter-Analysis.
- We conducted a comprehensive analysis of COVID-linked offensive tweets during the pandemic. Specifically, we performed fine-grained temporal, linguistic and user network analysis of offensive tweets, and the targets and authors of the identified offensive tweets. Our analysis reveals that 1) the real-world events were potentially connected to the rise of COVID-related offensive tweets; and 2) several targets received a large number of offensive tweets containing abusive words, which could negatively affect these targets during the pandemic.
- Using the tweets classified as offensive, we analyzed how offensive users became offensive over time and the influence of interactions between offensive users. We found that users posted more offensive tweets from the start of the pandemic, which remained high until the end of our analysis period. We also found that offensive users interacted more with other offensive users by retweeting and mentioning each other.

2 RELATED WORK

Due to the pandemic and lockdown, more COVID-19-related cyber-hostility was increasingly prevalent on social media. The pandemic spawned new COVID-related research starting from COVID-related dataset collection. Chen *et al.* [9] collected 123 million tweets from January to March, 2020. Banda *et al.* [4] collected 285 million tweets from January 2020 to June 2021. COVID-related datasets have also been collected in other languages [2, 20, 45]. Literature on COVID-19 dataset collection was followed by other COVID-19 related research such as misinformation detection [48, 49, 56], vaccination [13, 16, 42], and user emotion analysis [34, 62]

While hate and offensive speech have been studied extensively in the literature using different methods such as word embeddings [54, 55], deep neural networks [58, 59], and transformer based methods such as BERT [15, 37], these methods do not generalize well to new phenomenons like COVID-19 due to the new language variation it introduces [31]. Various works have studied COVID-related hateful behaviors. Schild *et al.* [47] analyzed the emergence of Sinophobic content during the outbreak of the COVID-19 pandemic on both Twitter and 4chan. Caleb *et al.* [63] developed COVID-HATE, a dataset of anti-Asian hate and counter-hate tweets. Uyheng *et al.* [51] developed a dynamic network framework to understand the spread of hate speech and hateful communities during the COVID-19 pandemic. Vishwamitra *et al.* [52] discovered hate-related keywords associated with COVID-19 in hateful tweets on Twitter using BERT attention. An *et al.* [3] analyzed users who posted anti-Asian messages when the pandemic began. Nghiem *et al.* [40] developed a multitask (level of aggression, target, and type of hate speech detection) training approach that incorporates agreement between data annotators for anti-Asian hate speech detection on Twitter. Researchers in [1, 18, 26, 33] also investigated hate speech detection during the COVID-19 pandemic.

Distinction from existing works. Our work distinguishes itself from existing works [3, 4, 9, 40, 47, 51, 63] in three ways. First, we extend our content analysis by analyzing Twitter data beyond all the above datasets and analyses. The extension allows us to ascertain whether offensive rhetoric continued or changed throughout the pandemic. Second, we study the targets of offensive speech beyond Asian targets during the pandemic and the characteristics of offensive tweets towards different targets. Finally, we analyze how offensive users changed in terms of posting offensive tweets and their interaction networks. Furthermore, we do not use a hard coded list of anti-Asian slurs to identify offensive users as opposed to An *et al.* [3] . Our offensive tweet classifier is trained on annotated COVID-19 related tweets and not on general hate or offensive dataset as done by Uyheng *et al.* [51] . Li *et al.* [31] showed that models trained on traditional hate or offensive datasets do not generalize well on COVID-19 related datasets.

3 DATASET AND CLASSIFIER

In this section, we describe COVID-OFFENSE, a large-scale dataset that contains potential COVID-linked offensive tweets. We collected COVID-related tweets for a period of 30 months and then annotated a subset of tweets for training a COVID-Twitter-BERT model to identify offensive tweets. This study has been approved by our institution's institutional review board (IRB).

3.1 Data Collection

We started our data collection in October 2020 and selected keywords using a snowball sampling technique [57]. First, we used the Twitter Streaming API to collect real-time tweets for one week with initial seed keywords related to COVID-19 and cyber-hostility, such as "COVID19", "coronavirus", and "Chinavirus" [9]. By manually reviewing the gathered tweet hashtags, we selected COVID-related hashtags and hashtags that have the potential to be associated with an offensive tweet based on their co-occurrence frequency with existing COVID-19 keywords and added them as new keywords for our data collection. In total, we used 120 keywords for data collection. The first category of our keywords is about COVID-19 which includes general keywords such as "coronavirus", "COVID", "virus", "pandemic", "lockdown", and "stayathome". The second category is about China, such as "Chinavirus", "Wuhanvirus" or "kungflu". Other categories include "mask", "boomer", "vaccine", "covidiot" (someone who ignores public health or safety guidelines), "Qanon", "Trump", "Gates", "Fauci", and "WHO" (World Health Organization). Table 6 in Appendix A lists the complete categories and keywords.

Twitter's Streaming API can only be used to collect real-time tweets, and Twitter's official Search API has a rate limitation. Therefore, we used an open-source tool snscrape[1] to collect the tweet ids of the COVID-related tweets during the period from January 1, 2020, to October 15, 2020, and then used the Twitter official API to retrieve the tweets content. After that, we used the Twitter Streaming API to collect real-time COVID-related tweets from October 16, 2020, to June 30, 2022. Our COVID-OFFENSE dataset spans 30 months from January 1, 2020 to June 30, 2022. We obtained 747

[1] https://github.com/JustAnotherArchivist/snscrape

Data	Duration	Data Collection Tool	# of Tweets	# of Users
Historical Tweets	Jan 1, 2020 - Oct 15, 2020	snscrape tool + Twitter Official API	239,017,320	24,988,390
Real-time Tweets	Oct 16, 2020 - Jun 30, 2022	Twitter Streaming API	508,305,375	34,800,883
Total	**Jan 1, 2020 - Jun 30, 2022**	-	**747,322,695**	**46,803,691**

Table 1: Statistics of our COVID-OFFENSE dataset.

million tweets published by 46 million users after removing all non-English tweets and retweets. Table 1 illustrates the statistics of our dataset.

3.2 Annotation

Annotation Strategy Due to the lack of a unified definition of offensive speech, determining what is offensive content is non-trivial. While it is easy to identify racial and sexist slurs [54], a text can be offensive without having any of these slurs. Based on different definitions of offensive speech in the literature [14, 27, 46, 53] and industry [17, 50], we present a specific definition related to COVID-19 used in our data labeling. We define *offensive content* as:

language used to attack a person or a group based on their social categories, such as race, sex, sexual orientation, gender, national origin, religion, disability, occupational status, or political belief. More specifically, text that promotes/incites violence, contains dehumanizing comparisons, tries to segregate/exclude, harass with/without racial epithet, expresses inferiority and contains profanity/offensive language are all considered offensive.

Context and the target of a tweet play a critical role in our labeling. Tweets with keywords such as fu*k, b*tch do not necessarily make those tweets offensive. Before labeling such tweets as offensive, we ensured that the tweet explicitly targets a person/group. Tweets that combine a location or person's name with a name variant used in referencing COVID-19, such as "virus", are labeled offensive. For example, tweets containing "China virus" or "Trump virus" are labeled offensive because they are hateful and encourage racist and xenophobic behavior [17, 30]. Tweets that do not fall in the above definition of offensive content, even if they contain offensive keywords, are labeled as non-offensive (None) given the context. For example, the tweet: *"I've only seen my family twice this year , f*ck the second lockdown"* is not considered offensive because it does not target a person or a group.

Tweets Sampling Due to the size of our dataset, we sampled a subset for labeling. To identify potential offensive tweets, we randomly sampled 36,000 tweets from our dataset and used the `Perspective` tool [2] to identify possible offensive tweets for labeling. `Perspective` scores texts based on how toxic they are and gives a score between 0 and 1. 1,235 tweets having a toxic score greater than 0.9 were retained. Tweets with a toxic score of less than 0.9 could also be offensive, so we randomly retained 450 tweets with toxic scores lower than 0.9 to mitigate the potential bias of the Perspective tool. We removed the hyperlinks but kept any user handles in tweets allowing us to determine explicitly when an offensive keyword is being used to refer to a person or group.

Labeling Process Three internal annotators labeled 1,679 tweets after removing duplicates. All the annotators were trained and given a guideline containing the definition and examples of offensive tweets. They were also made aware of the effects of extensive exposure to offensive content. In the first round of annotation, each annotator labeled 200 tweets. After labeling, all annotators reviewed the labeled results and agreed on the tweets that were labeled differently. The review process provided new insights, and the annotation guideline (*i.e.*, the definition that included more categories like political beliefs) was updated appropriately to reflect the new insight. Using the updated annotation guideline, we relabeled the 200 tweets. The second round consisted of each annotator labeling 400 tweets and reviewing the labeled results as done in the first round. Inconsistent labeling was minimal as annotators became more familiar with the definition of offensive content. In the third round, each annotator labeled 1079 tweets, and we repeated the review process of the previous steps. Majority voting was used if two annotators have the same label which is different from the third annotator. For inter-rater agreement score, we obtained a Fleiss' Kappa value of 0.53, which is within the moderate level of agreement [29], showing the difficulty in annotating offensive content. Of the 1,679 tweets, 554 were labeled as offensive and 1,125 as non-offensive.

3.3 Classifier Training

Pre-trained Model	Precision	Recall	F1
`BERT-Base`	0.868	0.849	0.858
`BERT-Large`	0.88	0.847	0.863
`BERTweet`	0.86	0.836	0.848
`COVID-Twitter-BERT`	**0.901**	**0.878**	**0.889**

Table 2: Performance of the fine-tuned BERT models.

We used the BERT (Bidirectional Encoder Representations from Transformers) [15] model for classification by taking the output representation of the classification token and feeding it to a feed-forward network with the softmax function. BERT has been used to improve various Natural Language Processing (NLP) tasks and it achieved excellent performance with only a small number of data for fine-tuning. We trained and compared four different BERT variants- `BERT-Base` (with 12 layers, 12 attention heads, and 768 hidden vectors), `BERT-Large` (with 24 layers, 16 attention heads, and 1024 hidden vectors), and BERT based models `BERTweet` [41] (trained with 850m English tweets based on RoBERTa [32]) and `COVID-Twitter-BERT` [38] (trained with 160m tweets based on BERT-large). `COVID-Twitter-BERT` achieved the best performance with 0.901, 0.878, and 0.889 average precision, recall, and F1 score, respectively. We used the fine-tuned `COVID-Twitter-BERT` model in the rest of our analysis for classifying tweets into two classes - offensive and non-offensive. We manually validated 100 classified offensive tweets, and our model correctly classified 88 tweets.

4 METHODOLOGY

This section describes the data analysis techniques used in our work. Figure 1 shows the overview of our analysis. First, we collected our dataset of COVID-related tweets with Twitter API using

[2] https://www.perspectiveapi.com

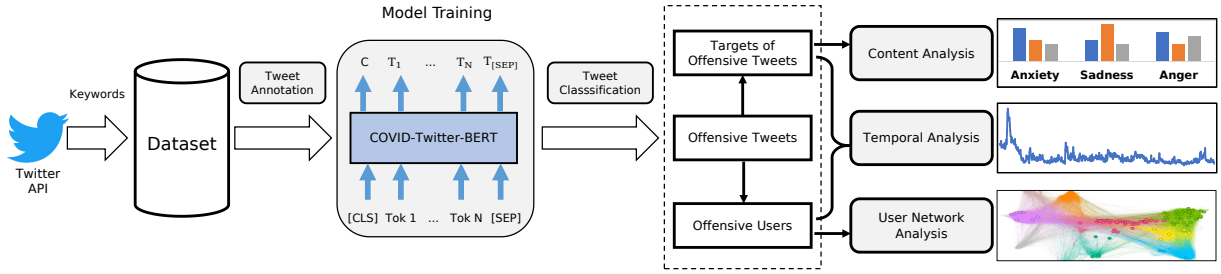


Figure 1: Overview of our analysis of COVID-related offensive tweets and targets.

specific keywords related to COVID (§ 3.1). Next, we randomly sampled and annotated a small number of tweets (§ 3.2) and used the annotated data to fine-tune a *COVID-Twitter-BERT* model (§ 3.3). The fine-tuned model was used to classify tweets as offensive or not offensive. After detecting offensive tweets, we extracted offensive tweets' targets and analyzed each target. We analyzed how users became offensive and the interaction networks of the offensive users. Various state-of-the-art analysis techniques, including temporal analysis (§ 4.1), content analysis (§ 4.2) and user network analysis (§ 4.3), were used to analyze the identified offensive tweets, offensive tweet targets, and offensive tweet authors.

4.1 Temporal Analysis

Our dataset spans a long observational period of the pandemic. We first conducted a temporal analysis of the offensive tweets to understand how offensive tweets evolved during the pandemic and how different pandemic events might have contributed to the rise of offensive speech.

The temporal analysis uses data mining techniques on objects or events chronologically ordered, following a time sequence. One crucial technique in the temporal analysis is change point detection, which can be implemented based on supervised methods, such as decision tree, Bayesian Networks, SVM [44, 60, 61], and unsupervised techniques, such as CUSUM [24], ChangeFinder [28], or Pruned Exact Linear Time (PELT) [25]. In our work, we used the PELT algorithm to obtain the change points in the weekly offensive tweets. We then analyzed the potential corresponding real-world events around the change point dates of the tweets, and checked how different pandemic events might have contributed to the rise of offensive speech and their potential correlations.

4.2 Content Analysis

Topic modeling. As many events unfolded during the pandemic, it is expected that discussions about the events must have taken place on Twitter. We used topic modeling [10] to identify popular topics, especially topics about individuals or groups as targets. Given a large corpus without any prior annotation, a topic model discovers the main themes in the corpora. It discovers groups of words close to each other in documents, and these words represent the topics. Each topic is a probability distribution over words in the vocabulary. We selected the BTM (biterm topic model) [10] technique in our analysis as it performs well on all tasks such as classification accuracy, topic coherence, and efficiency [43]. We trained weekly topic models and observed how weekly topics evolved. For each week, we selected the top topics with the most representative word in each topic. We also computed the frequency of words in all weekly topics to find the content that people cared about and discussed, which shows the overview of topics people cared about during the whole period.

Word2vec model. We explored how Twitter users discussed COVID-19 and the context of the discussion toward different targets, by utilizing word2vec modeling [36]. Specifically, we used the `skip-gram` model [35] with negative sampling, a shallow neural network to predict the context of words. We trained separate `word2vec` models for each week in our dataset (130 weeks in total). The word2vec models allow us to understand the extent of racist rhetoric and how this behavior persisted or declined throughout the pandemic.

Linguistic analysis. We used the Linguistic Inquiry and Word Count (LIWC) [8] tool to explore the sentence pattern and to obtain the various categories of words for offensive tweets toward different targets. LIWC is a text analysis program that calculates the percentage of words in a given text that fall into different linguistics and it is widely used to study emotions or causal words.

4.3 User Network Analysis

We analyzed the interactions between users to explore the possible factors that influenced offensive users to publish more offensive tweets. There are several interactions between users on Twitter, and we focused on users' mentions and retweet activity since they can easily be extracted from tweet objects directly [12]. After obtaining the interactions between all users in our dataset, we built the user mentions and retweet network for all users to have an overview of the users' network. For the user mentions activity, we extracted all the users with the "@" symbol in a tweet. For the retweet activity, although we have removed retweets when cleaning our dataset, we used the original tweets that contain retweets for this analysis.

5 RESULTS

Number of sampled tweets	128,353,749
Number of offensive tweets	3,645,890 (2.8%)
Number of users	20,622,186
Number of users who published offensive tweets	1,656,810 (8.0%)

Table 3: Statistics of sampled tweets in our analysis.

This section presents the findings of our analysis of COVID-related offensive tweets. Due to the size of our dataset and the time needed to classify all tweets, we randomly sampled 1 million tweets from each week in our dataset for our analysis. The first three weeks did not have enough tweets because they were at the start of the pandemic, so we used all tweets in the three weeks. As a result, we sampled 128,353,749 tweets (from the 130 weeks in

our dataset) published by 20,622,186 different users. Using our fine-tuned `COVID-Twitter-BERT` model, 3,645,890 (2.8%) of the sampled tweets posted by 1,656,810 unique users were classified as offensive[3]. Table 3 shows the statistics of the sampled tweets used in our analysis.

5.1 Analyzing Offensive Tweets

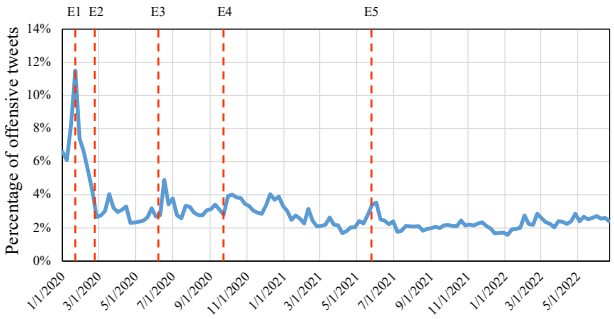


Figure 2: Percentage of weekly offensive tweets.

Index	Change Point	Real-world Event
E1	January 23, 2020	China imposed a lockdown in Wuhan
E2	February 23, 2020	Italy imposed lockdown to several cities
E3	June 1, 2020	Trump photo op at St. John's Church
E4	October 2, 2020	Trump was diagnosed with COVID-19
E5	June 1, 2021	Fauci's emails from April 2020 released

Table 4: Top 5 change points of weekly offensive tweets percentage and the possible corresponding real-world events.

5.1.1 Temporal Analysis of Offensive Tweets . We first conducted a temporal analysis of the offensive tweets to understand how offensive tweets evolved during the pandemic (RQ1). Figure 2 shows the percentage of weekly offensive tweets in our sampled tweets. The top 5 change points are shown in Figure 2, and the possible corresponding real-world events and dates are shown in Table 4. The highest percentage of offensive tweets appeared on January 23, 2020, when China imposed a lockdown in the city of Wuhan during the early stages of the COVID-19 outbreak. Another explanation could be our annotation strategy of labeling tweets as offensive if they associated a location or national origin with "virus" as in "China virus" or "Wuhan virus". The percentage of offensive tweets dropped until the end of February (E2). When coronavirus cases were confirmed in Europe and Italy went into lockdown, this caused an increase in the discussion about the virus. Tweets condemning Trump for posing for a picture with the bible on June 1, 2020 (E3) and tweets discussing his COVID-19 diagnosis on October 2, 2020 (E4) also resulted in a rise of offensive tweets. Only one change point is observed throughout 2021. This change point (E5) appeared in June 2021 when Fauci's emails from 2020 leaked. The percentage of offensive tweets remained steady until the end of our dataset. It is noteworthy that the percentage of offensive tweets decreased by 1.5% after 2021 (from 3.8% in 2020 to 2.3% in 2021 and 2022), and Welch's t-test showed a significant difference ($p < .0001$).

[3]It took 27 days to classify all sampled tweets on a cloud server with the NVIDIA Tesla P100 GPU.

5.1.2 Topic Modeling for Offensive Tweets. To explore the commonly discussed topics during the pandemic, we used topic modeling to identify popular topics. We trained weekly topic models and selected topics from the list of representative tokens based on the frequency of the tokens in all weeks. After merging similar tokens, such as China/Chinese to "Chinese" and Trump/realDonaldTrump to "Trump", the most frequent topics were "Chinese", "virus", "mask", "Trump" and "vaccine" . It is interesting that the topic "virus" appeared in most of weeks. After manual examination of some tweets, we found that while most of the users were using "COVID" or "coronavirus", some users continued to use "China virus", "Wuhan virus" or "Trump virus" in 2022, which led to such tweets being classified as offensive. Our later analysis using Word2vec modeling [36] demonstrates the tight association of these words based on their high similarities, as shown in Figure 10 in the Appendix B. The results of our topic modeling answer RQ2 in part, and show that the group (China and Boomer) and highly frequent individuals (Trump and Fauci) were highly discussed during COVID-19.

5.2 Analyzing the Targets of Offensive Tweets

Having identified groups or individuals discussed during COVID-19 in Section 5.1.2, this section analyzes the targets of offensive tweets and the characteristics of offensive tweets towards targets, answering the rest of RQ2.

5.2.1 Identifying the Targets of Offensive Tweets. To understand the rate/amount of offensive tweets targeting different groups or individuals, their periods of occurrence during the pandemic, and the extent of offensive rhetoric toward targets, we first identified targets of offensive tweets based on our topic modeling results. We identified the topics that corresponded to either an individual or group and were highly associated with COVID-19. Consequently, we found seven frequent targets, including "Chinese", "Trump", "Fauci", "Boomer", "Gates", "Cuomo" and "Covidiot".

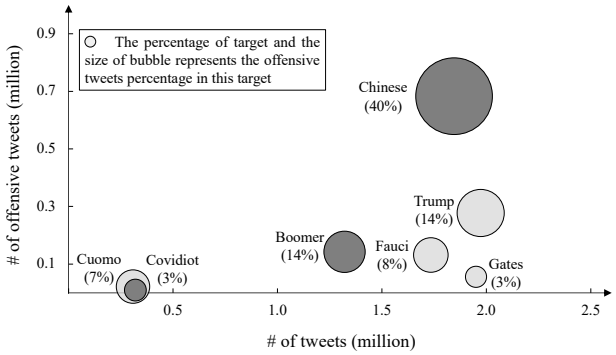


Figure 3: Targets of offensive tweets and the corresponding percentages of offensive tweets by targets.

Figure 3 shows the number of tweets, the number of offensive tweets and the percentage of offensive tweets regarding each target. Group targets are coded with dark gray, and individual targets are coded with light gray. The percentage beside each target bubble represents the percentage of offensive tweets regarding the target, and the size of the bubble is relative to the percentage of offensive tweets in the target it represents. The target category "Chinese" has the most significant number of offensive tweets and the largest

percentage of offensive tweets, with a percentage of 40%. Trump has the highest number of tweets. Although "Cuomo" has the fewest number of tweets, it has a high percentage of offensive tweets at 7%, possibly stemming from the perception that the former New York governor mishandled the response to the pandemic in New York. "Boomer" had fewer tweets but more offensive tweets than "Fauci" and "Gates", indicating that people had more negativity towards this group. **The results show that several targets received large number of offensive tweets during the pamdemic.**

5.2.2 Temporal Analysis of Different Targets. Next, we performed temporal analysis to model the characteristics of different targets of offensive tweets. Figure 4 shows the proportion of offensive tweets corresponding to each target over time. We observed that "Chinese" and "Trump" had the highest percentage of offensive tweets and a similar trend in 2020. Manually examining the tweets reveals that users referred to the coronavirus as the "Chinese virus" or "Trump virus". Trump supporters used the former and the latter was used to attack Trump and his supporters. For example, in March 2020, after Trump called the coronavirus the "Chinese virus", the percentage of offensive tweets about Chinese and Trump both increased. A user wrote *"ChinaVirus? No, the TrumpVirus is"*. A similar occurrence was observed in October 2020, when Trump was diagnosed with the COVID-19 virus. The percentage of offensive tweets towards Trump started reducing from February 2021 while the percentage of offensive tweets towards Chinese remained high throughout our study. Interestingly, the percentage of offensive tweets towards Fauci increased in 2021, surpassing Trump. The percentage of offensive tweets towards other targets remained below 5% except for Boomer, which was high only in the first few months of 2020.

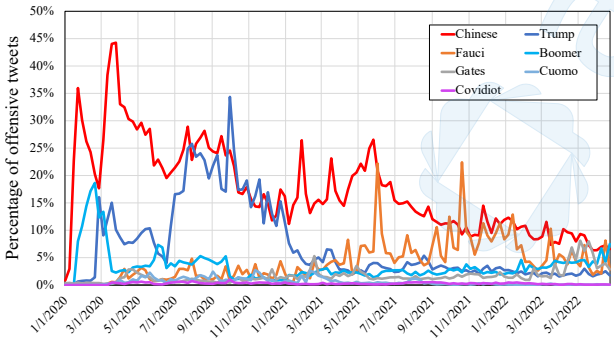


Figure 4: Proportion of offensive tweets on seven targets related to COVID-19.

To explore whether offensive tweets towards different targets surged during certain short periods or were persistent over a long observational period, we calculated the percentage of each target's weekly offensive tweet number normalized by the target's overall offensive tweet number, and the results are shown in Figure 5. We observed that most discussions about Chinese and Boomer were within the first three months of 2020 during the initial stages of the COVID-19 outbreak. Tweets about Boomer mocked older people for spreading the virus when the pandemic started. For example, one user tweeted *"I'm absolutely pissed about this. One fu*king boomer infected my brother"*. "Covidiot", frequently used to describe

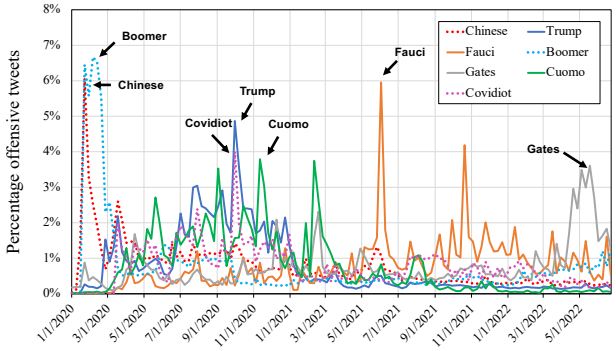


Figure 5: Percentages of daily tweets about seven individuals or groups (solid lines represent individuals and dotted lines represent groups).

individuals who did not abide by COVID-19 safety instructions such as wearing a mask and social distancing, had a similar trend with Trump in October 2020. "Covidiot" was also frequently used to denigrate Trump and his supporters. For example, a user published *"@realDonaldTrump absolutely fu*king stupid, covidiot"*. Cuomo's highest surge happened in November 2020 and March 2021. Trump and Fauci's surges are likely due to real-world events, as discussed in the previous sections. Gates was frequently targeted in 2022 after proposing a plan to prevent the next pandemic. **Such results indicate that for most of the targets, the offensive tweets toward them were concentrated in short periods.** Such an outburst of offensive tweets towards these groups or individuals could negatively affect these targets.

5.2.3 Linguistic Analysis of Different Targets. Next, we explored what words users used to discuss different targets by conducting a linguistic analysis. We utilized Word2vec modeling [36] to find the most similar tokens of each target and Linguistic Inquiry and Word Count (LIWC) [8] tool to get the word usage of different targets.

Target	Abusive Words	Example
Chinese	virus, boycott, racist, started, chickity	China start world war by china virus. We can't do any thing against china but we can boycott made in china goods.
Trump	loser, racist, asshole, jackass, sucker	The China virus in China the Trumpvirus in America!
Boomer	fu*king, shit, ass, b*tch, fu*k	Coronavirus said fu*k them boomers.
Fauci	fraud, ass, asshole, fu*k, fire	Fauci is a fu*king fraud. He needs to be fired.
Covidiot	asshole, moron, twat, loser, idiot	You're a coronavirus mass murderer. Covidiot to the hague with you.
Cuomo	killer, sexual, murderer, sexually, killed	Democrat governors like Cuomo are the true China covid killers.
Gates	hell, satan, nig*a, fu*k, cum	This is fu*king bill gates speaking.

Table 5: Frequently used abusive words and example offensive tweets regarding each target.

We trained separate Word2vec models for each week to understand the extent of offensive rhetoric throughout the pandemic. A representative week is provided in Table 7 in the Appendix B, which contains the most similar words to the targets in the week of March 18, 2020. We found that for some targets, the most relevant

words had racial slurs or abusive words, such as "trumpvirus", "ass" and "b*tch". We calculated the frequency of the similar words in all weeks for each target. Table 5 lists the top 5 most frequently used abusive words towards targets and example tweets. For the target "Chinese", people frequently used "Chinese virus" and called for the boycott of products from China. For "Trump", some tweets regularly called him an asshole. We note that some abusive words are more hateful than offensive. Abusive words were also often used in discussions of other targets such as "ass" and "fu*k". "Gates", "Boomer", and "Cuomo" were associated with "hell", "fu*king", and "killer", which frequently appeared in 51, 29, and 24 weeks targeting these targets, respectively.

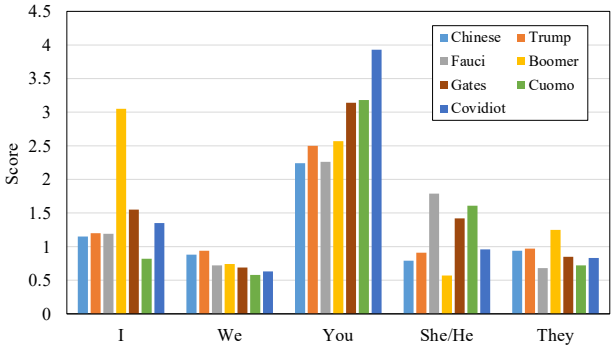


Figure 6: Person pronouns scores of LIWC for different targets of COVID-19 related offensive tweets.

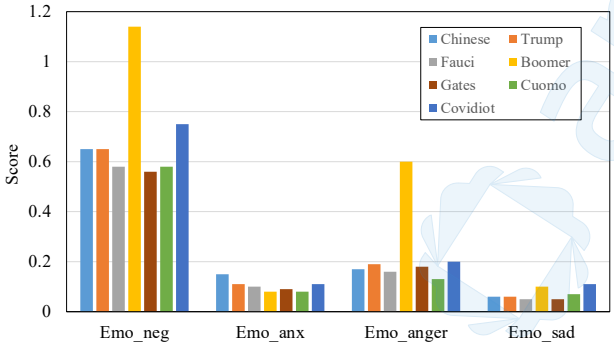


Figure 7: Negative emotion scores of LIWC for different targets of COVID-related offensive tweets.

We also used the LIWC tool to calculate the various categories of words for different targets. We found a significant difference when referring to the personal pronouns and emotion scores of different targets, which helps us better understand the characteristics of offensive tweets targeting them. Figure 6 shows the person pronouns scores, and Figure 7 shows the negative emotion scores for each target. By combing them with the commonly used abusive words regarding each target, we observed interesting patterns for several targets. The target "Boomer" had the highest score of the first person singular pronoun ("I") and the third person pronoun ("They"), as shown in Figure 6. Considering that abusive words, such as "fu*k", were frequently used towards "Boomer" (as shown in Table 5), it is not surprising that this group had the highest score for anger emotion and negative emotion, as observed in Figure 7. For example, a user tweeted *"Fu*k boomers and I hope they all get taken out with the virus"*. For other targets such as "Covidiot", tweets used more of a second pronoun ("you"), indicating they were replying to other users and labeling them as a group to attack them, as the example tweet in Table 5 shows. Different with 'Fauci", "Cuomo" and "Gates", which all contain tweets using "he" to criticize them, more tweets discussing "Trump" used "we", such as *"We need to call it TrumpVirus because its from bats and trump is bat shit crazy"*. Such pronoun shows the possible political group of users. **Our results show that for several targets, they received offensive tweets with extremely abusive words, which led to these tweets being negative and possibly hateful.**

Our analyses about targets of offensive tweets show that for several targets, such as Trump, Boomer and Fauci, there was a large number of offensive tweets with abusive words, some of which are hateful surged in short periods during the pandemic. Such high offensive rhetoric towards groups or individuals could negatively affect them, especially vulnerable groups like Boomer, who were more at risk at the start of the pandemic, and Trump, who was later diagnosed with COVID.

5.3 Analyzing Offensive Tweet Users

In this section, to answer RQ3, we characterized users who posted offensive COVID-related tweets (whom we call offensive users). We analyzed how these users became offensive over time and how they interacted with others in their networks during the pandemic.

As described in the Dataset section, over 1.6 million users published at least one offensive tweet, and not all of them were offensive users. To find the active offensive users, we first removed users that posted less than 100 tweets since they likely did not produce enough prior tweets for our analysis. For the remaining users, we calculated the percentage of offensive tweets for each user and then used the k-means algorithm to classify each user based on their offensive tweet percentage. The advantage of k-means is that it is an unsupervised learning method and we can set k = 2 to classify users as offensive or not in our dataset without setting up a threshold for the percentage of offensive tweet [62]. As a result, over 3,000 users were classified as offensive users. For each offensive user, we collected the user's monthly tweets from January 2019 to June 2022, and for each month, we collected ten tweets, resulting in 450 tweets for each user. In total, we collected 684,433 past tweets from 2,445 offensive users using the `snscrape` tool. The collected tweets were classified as offensive or non-offensive using our fine-tuned `COVID-Twitter-BERT` model.

5.3.1 Changes of Offensive Users. First, we calculated the monthly percentage of offensive tweets posted by offensive users. For comparison, we also randomly sampled the same number of normal (non-offensive) users, collected their past tweets and classified their tweets using our fine-tuned model. Figure 11 in Appendix C shows the results and it is not surprising that offensive users had a higher percentage of offensive tweets than normal users in all months. The percentage of offensive tweets by normal users only increased in the first three months after the COVID-19 outbreak and then decreased to pre-pandemic levels. On the contrary, the percentage of offensive tweets by offensive users increased a lot and then fluctuated, showing that offensive users were more influenced by real-world events.

The percentage of offensive tweets by offensive users also remained elevated until 2022 and was significantly higher than the number before the pandemic. **These results indicate that the pandemic had a lasting impact on these offensive users.** We also used Welch's independent sample t-tests to compare the percentage of offensive tweets before and after the pandemic. For offensive users, the result shows a significant difference ($p < 0.0001$), and for normal users, there is no significant difference.

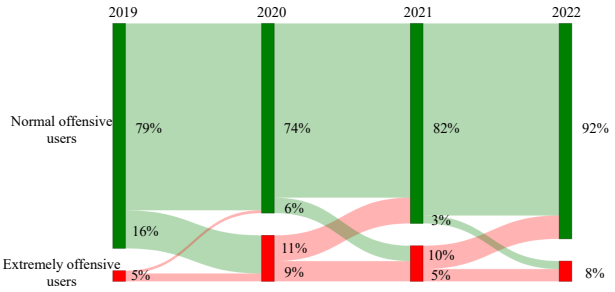


Figure 8: Change of offensive users and non-offensive users over four years.

Next, we summarized the percentage of yearly offensive tweets from offensive users. We further divided the 2,445 offensive users as extremely offensive or normally offensive based on their past tweets by using the k-means (k = 2) clustering algorithm. Note that the new division for offensive users is based on their tweets that we newly collected instead of COVID-related tweets from our dataset. Figure 8 shows the change in offensive users in 4 years. As expected, in 2020, a considerable number of normally offensive users became extremely offensive users (16%), and in 2021 most of them became normally offensive again. The high percentage of extremely offensive users in 2020 (20%) and 2021 (15%) indicates that users became more offensive and cynical during the pandemic. For example, only 3% and 9% of a user's tweets were offensive in 2019 and 2022, but the percentages jumped to 44% and 66% in 2020 and 2021, respectively. Additionally, 1% of users remained extremely offensive in all four years. The user who posted the most offensive tweets in 2020 had 75%, 87.5%, 78%, and 44% of his/her tweets rated as offensive tweets in the four years respectively, showing that the extremely offensive users also became more offensive during the pandemic. Compared to before the pandemic, the number of extremely offensive users increased after the COVID-19 outbreak, corresponding to the results shown in Figure 11. We thus answer RQ3 in part, that offensive users became more offensive during the pandemic and remained offensive thereafter.

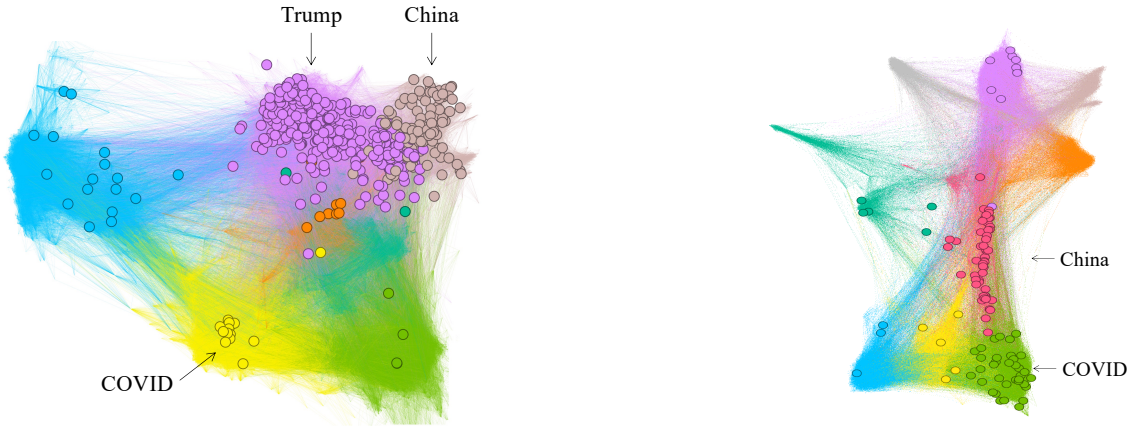
5.3.2 User Network Analysis. We further analyzed the user network based on users' interactions. The user mention network and the user retweet network of all the users in our dataset are shown in Figure 9, generated by the tool "Gephi" [5]. In the figures, each node represents a user in our dataset. There is an edge between two nodes if a user has an interaction with another user. For simplicity, we removed users (nodes) that interacted with less than 10 users and all the interactions (edges) that were less than 20, leaving 77,416 nodes and 1,727,712 edges. This filtering removes noise from small communities or weak connections among users and enables better network visualization. After filtering, we used the Louvain community detection method [7] to cluster users, which is a commonly used approach to extract communities from large networks. At last, we employed the ForceAtlas2 algorithm [23], a continuous graph layout algorithm, to take into weights to organize the nodes and make the connected nodes close. The color shows the group that users are in and the distance shows the interaction frequency between different groups. The more interactions they have with each other, the closer they are in the figure. We highlighted the offensive users and the circles represent the offensive users.

It is interesting that most offensive users were clustered into the same groups in both mention and retweet networks (3 groups in the mention network and 2 groups in the retweet network). **Such grouping shows that offensive users had more interactions with other offensive users compared with normal users** (72 vs. 17). We also found a few offensive users were not clustered into the same groups in Figure 9 because they didn't interact with other offensive users frequently. There might exist other reasons that led to these users becoming offensive, such as real-world factors, and this is not explored in our study. After checking the tweet content of offensive users, we labeled the topics they discussed in Figure 9. In the user mention network, a large group of offensive users mentioned Trump's Twitter account and expressed their opinion, which led to them being grouped; another group of offensive users discussed "China". There is also another small group discussing "COVID". For the user retweet network, we also found two groups discussing "China" and "COVID", separately. The retweet network is not as dense as the mention network because users can mention several people in one tweet, creating more interactions. We thus answer the rest of RQ3, showing that offensive users interact more when discussing the same topic.

6 DISCUSSION

6.1 Implications

Our series of analyses provide insights into the sources and targets of COVID-related offensive language on Twitter. Our temporal analysis of offensive tweets found a potential connection between the rise and fall of offensive COVID-19 related tweets and events in the physical world. Indeed, taking the target "China/Chinese" as an example, linking China to the origin and consequences of the pandemic by powerful individuals, especially then-President Donald Trump, contributed to the coarsening of Twitter dialogue regarding China and individuals of Chinese descent. The increase in Sinophobia on Twitter is troubling for myriad reasons. For one, exposure to hateful online material – especially when repeated - can have deleterious effects on online users, leading to mood swings, depression, anger, fear, and mistrust [39]. Additionally, online forums can quickly transform into echo chambers, resulting in the reinforcement and normalization of ideas, including fringe and hateful ideas. In this case, the notion that the Chinese government engaged in a nefarious plot to unleash a deadly pandemic rapidly gained traction on numerous social media platforms, including Twitter, and was accepted by some users as truth rather than conspiracy. The subsequent rise in anti-Asian hate crimes was unfortunately predictable, as extant research demonstrates that frequent exposure to online hate speech has the capacity to foster radicalization, and even offline violence [19].



(a) User mention network **(b) User retweet network**

Figure 9: Network between users in our dataset. Circles mean the offensive users. and different colors mean users in different groups because they have more interactions. The results show that offensive users had more interactions with other offensive users (close to each other) and they discussed similar topics.

In sum, this study has important implications regarding online speech. First, public figures, especially politicians, should be prudent with their speech on social media, recognizing their outsized ability to shape public discourse. Further, this work accentuates the need to grapple with the difficulty of regulating online speech. With the exceptions of speech that can incite imminent lawless action, pose true threats, or serve as fighting words, hate speech is allowable in the U.S. [6]. But, while the government is largely restrained in the fight against online hate, the private sector has remedies at its disposal. Namely, social media sites can choose to ban hateful speech on their platforms, something they are increasingly likely to do as financial incentives dictate such action. Online users therefore play a vital role in online speech regulation, as advertisers flock to social media sites with large user bases. We hope that by shining a light on the spread and dangers of Sinophobia on Twitter, this paper encourages more online users to actively mitigate hate by not only choosing not to engage in it, but also patronizing social media sites that are proactive in regulating harmful online content.

In addition, the analysis we have conducted in this study provides insights into how offensive language evolves during global events and how such language can target individuals and communities, especially minority/marginalized communities. Minority communities are often scapegoated during such events; this research enables policymakers to design better policies that ensure that every citizen's human right and dignity is preserved.

6.2 Limitation

Our work has several limitations. First, our dataset consists of two different data collection methods across different periods. Twitter may have deleted some tweets or users when we collected past tweets, likely making our dataset incomplete. This is because we started our data collection in October 2020. After using the regression discontinuity design [22] to compare the weekly tweet number before and after we changed our data collection method (snscrape tool vs. Twitter Streaming API), we found a significant difference in terms of tweet count. We also sampled 30,000 past offensive and non-offensive tweets and found that offensive tweets were more likely to be removed. This indicates that more offensive tweets had likely been removed when we collected past tweets. Even so, to the best of our knowledge, our COVID-OFFENSE dataset is one of the largest COVID-related datasets. Second, we used a relatively small number of labeled tweets to train a BERT model used in the classification of our dataset. We plan to label more offensive data and train a deep learning model for better classification performance. Yet, the accuracy of our model is still around 90%, which benefits from the performance of the BERT model pre-trained on COVID-related tweets. Third, in the user analysis, we sampled only a part of users' past tweets and we only considered the mention and retweet data while Twitter provides other interactions such as likes, follows, and replies. We plan to conduct a more comprehensive analysis of users' social networks in our future work.

7 CONCLUSION

In this paper, we collected a large-scale Twitter dataset over a 30-month period during the pandemic and performed a comprehensive analysis of offensive tweets and their targets as the pandemic progressed. We found that the ebb and flow of offensive tweets reflect events in the physical world and the percentage of offensive tweets decreased after the pandemic. Indeed, our results demonstrate a rise in offensive tweets with abusive words targeting different individuals or groups in a short period, as the pandemic worsened and prominent politicians used disparaging language. We observed how offensive tweet targets changed and how events in the real world potentially caused such offenses to increase in a short period. Our study sheds light on the attributes of online users who author offensive tweets related to the pandemic and reveals that tweeting offensive material is partly a function of interacting with others.

ACKNOWLEDGMENT

This work is in part supported by National Science Foundation (NSF) under the Grant No. 2031002, 2114920, 2228616, 2120369, 2114982, 2129164, 2228617 and 2245983.

REFERENCES

[1] Shivang Agarwal and C Ravindranath Chowdary. 2021. Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. *Expert Systems with Applications* 185 (2021), 115632.

[2] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315* (2020).

[3] Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-Asian hateful users on Twitter during COVID-19. *arXiv preprint arXiv:2109.07296* (2021).

[4] Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia* 2, 3 (2021), 315–324.

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 3. 361–362.

[6] Lauren E Beausoleil. 2019. Free, hateful, and posted: rethinking first amendment protection of hate speech in a social media world. *BCL Rev.* 60 (2019), 2101.

[7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[8] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* (2022).

[9] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6, 2 (May 2020), e19273. https://doi.org/10.2196/19273

[10] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 2928–2941.

[11] San Francisco Chronicle. 2020. *Coronavirus: Asian American groups compile hate crime reports as trump persists in 'Chinese virus' attacks.* https://www.sfchronicle.com/bayarea/article/Coronavirus-Asian-American-groups-compile-hate-15144295.php

[12] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the echo chamber: the Italian vaccination debate on Twitter. In *Proceedings of the International AAAI conference on web and social media*, Vol. 14. 130–140.

[13] Giuseppe Crupi, Yelena Mejova, Michele Tizzani, Daniela Paolotti, and André Panisson. 2022. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 102–113.

[14] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. arXiv:1703.04009 [cs.CL]

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[16] Marco Di Giovanni, Francesco Pierri, Christopher Torres-Lugo, and Marco Brambilla. 2022. VaccinEU: COVID-19 vaccine conversations on Twitter in French, German and Italian. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1236–1244.

[17] Facebook. 2023. Facebook Community Standards: Hate Speech. https://www.facebook.com/communitystandards/hate_speech

[18] Komal Florio, Valerio Basile, and Viviana Patti. 2021. Hate speech and topic shift in the covid-19 public discourse on social media in Italy. In *8th Italian Conference on Computational Linguistics, CLiC-it 2021*, Vol. 3033. CEUR-WS, 1–7.

[19] Abraham H Foxman and Christopher Wolf. 2013. *Viral hate: Containing its spread on the Internet.* Macmillan.

[20] Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Naist covid: Multilingual covid-19 twitter and weibo dataset. *arXiv preprint arXiv:2004.08145* (2020).

[21] Angela R Gover, Shannon B Harper, and Lynn Langton. 2020. Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice* 45, 4 (2020), 647–667.

[22] Guido W Imbens and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics* 142, 2 (2008), 615–635.

[23] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014), e98679.

[24] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. 2001. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining*. IEEE, 289–296.

[25] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* 107, 500 (2012), 1590–1598.

[26] Bumsoo Kim, Eric Cooks, and Seong-Kyu Kim. 2021. Exploring incivility and moral foundations toward Asians in English-speaking tweets in hate crime-reporting cities during the COVID-19 pandemic. *Internet Research* (2021).

[27] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*. 1–11.

[28] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque, and Juan Carlos Nuno. 2008. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences* 105, 13 (2008), 4972–4975.

[29] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[30] Carmen Lee. 2021. HateIsAVirus: Talking about COVID-19 'hate'. *Viral discourse* (2021), 61–68.

[31] Mingqi Li, Song Liao, Ebuka Okpala, Max Tong, Matthew Costello, Long Cheng, Hongxin Hu, and Feng Luo. 2021. COVID-HateBERT: a Pre-trained Language Model for COVID-19 related Hate Speech Detection. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 233–238.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[33] Runjing Lu and Yanying Sheng. 2020. From fear to hate: How the COVID-19 pandemic sparks racial animus in the United States. *arXiv preprint arXiv:2007.01448* (2020).

[34] Lydia Manikonda, Mee Young Um, and Rui Fan. 2022. Shift of User Attitudes about Anti-Asian Hate on Reddit Before and During COVID-19. In *14th ACM Web Science Conference 2022*. 364–369.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[37] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*. Springer, 928–940.

[38] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv:2005.07503 [cs.CL]

[39] Matti Näsi, Pekka Räsänen, James Hawdon, Emma Holkeri, and Atte Oksanen. 2015. Exposure to online hate material and social trust among Finnish youth. *Information Technology & People* (2015).

[40] Huy Nghiem and Fred Morstatter. 2021. " Stop Asian Hate!": Refining Detection of Anti-Asian Hate Speech During the COVID-19 Pandemic. *arXiv preprint arXiv:2112.02265* (2021).

[41] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 9–14.

[42] Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 782–793.

[43] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[44] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)* 6, 2 (2010), 1–27.

[45] Ibrahim Sabuncu and Zeynep Yurex. 2020. Corona Virus (COVID-19) Turkish Tweets Dataset. https://doi.org/10.21227/0wf0-0792

[46] Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10, 1 (1 Dec. 2020). https://doi.org/10.1186/s13673-019-0205-6

[47] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2020. "Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. arXiv:2004.04046 [cs.SI]

[48] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A Duo-generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. In *Proceedings of the ACM Web Conference 2022*. 3623–3631.

[49] Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of COVID-19 misinformation on social media. In *Proceedings of the ACM Web Conference 2022*. 3706–3717.

[50] Twitter. 2023. Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[51] Joshua Uyheng and Kathleen M Carley. 2021. Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science* 6, 1 (2021), 1–21.
[52] Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. 2020. On analyzing covid-19-related hate speech using bert attention. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 669–676.
[53] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899* (2017).
[54] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
[55] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access* 6 (2018), 13825–13835.
[56] Maxwell Weinzierl and Sanda Harabagiu. 2022. Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse. In *Proceedings of the ACM Web Conference 2022*. 3196–3205.
[57] Kai-Cheng Yang, Pik-Mai Hui, and Filippo Menczer. 2019. Bot electioneering volume: Visualizing social bot activity during elections. In *Companion Proceedings of The 2019 World Wide Web Conference*. 214–217.
[58] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web* 10, 5 (2019), 925–945.
[59] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*. Springer, 745–760.
[60] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing*. 312–321.
[61] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*. 247–256.
[62] Assem Zhunis, Gabriel Lima, Hyeonho Song, Jiyoung Han, and Meeyoung Cha. 2022. Emotion Bubbles: Emotional Composition of Online Discourse Before and After the COVID-19 Outbreak. In *Proceedings of the ACM Web Conference 2022*. 2603–2613.
[63] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. arXiv:2005.12423 [cs.SI]

Appendix A CATEGORIES AND KEYWORDS IN OUR DATASET

Category	Keywords
COVID	coronavirus, covid19, pandemic, virus, outbreak, plandemic, china, stayhome, covid-19, covid, corona, wuhan, covid_19, lockdown, coronavirusoutbreak, stayathome, socialdistancing, pandemic, coronaoutbreak, stayhomestaysafe, staysafestayhome, covid__19, covid-19, learntherisk, howwegothere, nonewnormal, gopsuperspreaders, herdimmunity, stopthecovidchaos, lockdown2
China	wuhancoronavirus, wuhanvirus, chinesevirus, chinavirus, coronaviruschina, ccpvirus, chinacoronavirus, chinaliedpeopledied, wuflu, kungflu
Mask	mask, antimask, maskless, maskfree, unmask, nomask, nomasks, unmaskarizona, unmaskamerica, nomasknancy, nomaskmandate, antimaskers, nomaskmandates, maskoff, maskoffamerican, maskdontwork, maskoffamerica, unmaskthetruth, unmasked
Boomer	boomerremover, bommer, babyboomers, babyboomer, boomers, boomersooner, okboomer
Vaccine	vaccine, vaccines, coronavirusvaccine, russianvaccine, covidvaccine, covid19vaccine, vaccineinjury, fluvaccine, rnavaccines, vaccineswork, pfizer, pfizervaccine
Covidiot	covidiot, covidiots, covididiots, covidiotinchief
Qanon	qanon, qanons, qanondon
Fauci	fauci, faucithefraud, drfauci, tonyfauci, drfaucitimecover, faucihero, faucifraud, firefauci, criminalfauci, followthefauci
Cuomo	killercuomo
Trump	trumpvirus, trumpkills, trumppandemic, trumpvirusdeathtoll193k, trumpvirusdeathtoll186k, trumpviruscatastrophe, trumpkillsamericans, trumplied200kdied, trumpliedpeopledied, trumphascovid, trumpcovid, trumpcovid19, trumpcovidhoax, covidcaughttrump, trumpcrimefamily, trumpisbroke, trumpvirusdeathtoll210k, trumpispathetic, trumpcrimefamilyforprison, trumpvirusdeathtoll225k
Gates	billgates, gates, gatesofhell, exposebillgates
WHO	#who

Table 6: All categories and keywords in our dataset.

Appendix B WORD2VEC MODELING RESULTS

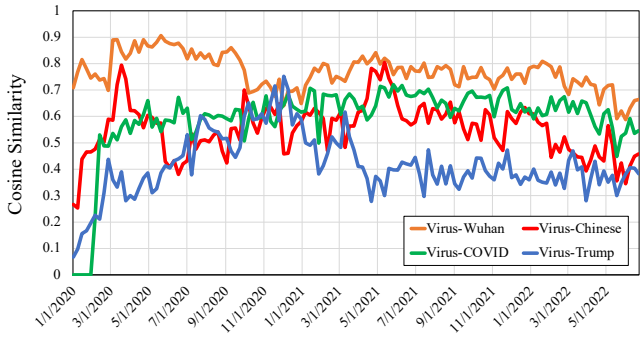


Figure 10: Cosine similarities between "Wuhan", "Chinese", "Trump", "COVID" and "Virus" using Word2vec modeling, showing that the "Wuhan", "Chinese" and "Trump" were used with "virus" together as frequently as "COVID" and "virus".

Chinese	Trump	Fauci	Cuomo	Boomer
virus	president	dr	governor	**remover**
china	donald	anthony	dilley	boomers
wuhan	**trumpvirus**	faucis	trump-asorous	doomer
communist	goodbyegop	**ass**	andrew	ok
government	trumps	barr	blasio	girl
originated	theyknew	listen	gov	simp
racist	potus	shiva	hates	xd
deadlyvirus	chrissy	hero	**b*tch**	lankford
people	trump-pandemic	trump	nonessen-tial	millennial
came	coronavirus	**asskisser**	eh	lol

Table 7: Most similar words to targets "Chinese", "Trump", "Fauci", "Cuomo" and "Boomer" for the week of March 18th, 2020 (Offensive words are highlighted).

Appendix C USER ANALYSIS

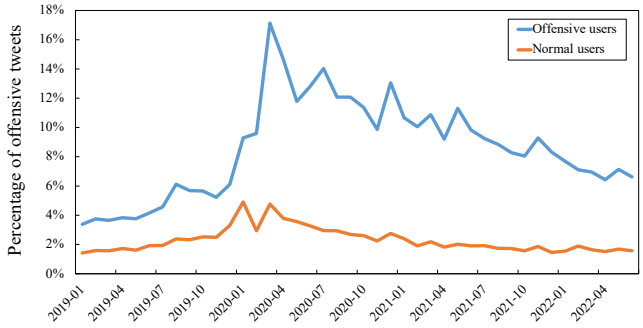


Figure 11: Monthly offensive tweets percentage for offensive users and normal (non-offensive) users.