ISSN:2790-5179 DOI: 10.56028/iajhss.1.2.1

Empirical Study of Hate Speech in Social Media During COVID-19 Crisis in the United States

Chenghan Wen[1,a], Deema Alnuhait[1], ShanghuLiu[1,b],Wenda Zhou[2,c],Jun Zhang[3,d]

[1]Columbia University, [2]Cornell University, [3]Tianjin University of Finance and Economics

[a]cw3330,daa2182, [b]sl4840@columbia.edu,[c]wz325@cornell.edu,[d]hellenjunzhang@gmail.com

Abstract

As the COVID-19 outbreak, hate speech on social media towards Chinese and other Asian groups has encouraged "Sinophobia". To capture the situation of hate speech on Twitter, we hydrated the tweets discussing COVID-19, written in English language and have location within the USA. These tweets hydrated from Twitter API in span of 5 months (153 days). We have obtained 543,943 tweets in which we identify 40,579 Hate Speech occurrences. We categorized and analyzed them according to Pysentimiento model which is based on BERT models and, Latent Dirichlet Allocation Model (LDA). The results indicate that there are substantial associations between the increased amount of hate speech and the increased rate of deaths due to COVID-19, increased rate of new COVID-19 cases, and negative tests rate.

Keywords

social science, Covid, hate speech, data mining

1. Introduction

The COVID-19 pandemic has dramatically changed the world. It resulted to extreme fear, frustration, and stress for many people as the world deals with this invisible enemy. Many lockdowns and social distancing orders were enforced to minimize the spread of the pandemic. This abrupt changes affected people's lives both economically and psychologically. Having the fact that the first COVID-19 cases were discovered in Wuhan city, Hubei, China, resulted in an increasing number of Sinophobia (nyt, 2021). Fan et al. (2021) mentioned that the COVID-19 pandemic has unfolded hate speech on social media about China and Chinese people.Besides, Mehta (2021) noted that Twitter has noticed 900% increase in hate speech towards Chinese people.

In this empirical research, our goal is to spot lights on the main contributions of COVID-19 effecting the increased rate of hate speech towards Chinese on Twitter within the United States. so the main objective of the research is to address the research questions, "what are the most important factors that affect hate speech related Tweets?" and "what are the main topics of hate speech Tweets?" Through the literature research, we hypothesized that the increased number of new COVID-19 deaths and new positive cases are the key factors that affect hate speech tweets. In addition, they also hypothesized that the main topics of hate speeches on Twitter are focused on preventive and pandemic issues. In this case, the independent variables are COVID-19 new positive and negative cases, deaths, tests, hospitalization and Trump's tweets while the dependent variable is the rate of hate speech-related content on Twitter in the United States.

2. Related work

2.1. The effect of COVID-19 pandemic on negative sentiments

A subgroup of the growing body of literature in studying the factors of COVID-19 contributing to the negative sentiment of people. According to (Dai et al., 2021) one of the probable contributors is the intolerance of uncertainty. The COVID-19 pandemic is an extremely stressful thing even in modern day history, it has brought great uncertainties which promoted negative reactions from many people, as they continuously seek normalcy. Similarly, (Herbert et al., 2021) reported that people constantly expressed their fear of losing their lives due to the pandemic. Since no one has a concrete answer to most of their questions, it results to negative sentiments. To make it worse, some of these negative sentiments are targeted towards a ethnicity. This worsened xenophobia across the globe. In social media a study by Tariq Soomro et al. (2020) analyzed over 18 million COVID-19 related tweets to study the effects of COVID-19 number of cases on the public sentiment. According to their study, there exists a positive correlation between the negative sentiment and the strikes of COVID-19 cases.

2.2. Increase in hate speech due to COVID-19

There has been an significant increase in hate speeches due to the ongoing COVID-19 pandemic. In a study, (Gray and Hansen, 2021) concluded that the pandemic has increased the criminal rate against Chinese people in London. Similarly, the Watch (2020) claimed that pandemic made Asians as "human punching bag" as people get attacked and harassed on streets and on social media platforms. At the same time, on Twitter, there is an alarming increase in hate tweets targeting at Asians, which turned into physical attacks later. Fan et al. (2021) reported that they collected 3,457,402 key tweets about China and to COVID-19 in which about 40% of the them are from the U.S and 25,467 tweets contains hate speech content.

2.3. Trump sentiment effects on people's attitude towards Chinese

Former President Trump was criticized for calling the COVID-19 virus as "Chinese Virus." According to Reja (2021), this remark from Trump helped the rise of anti-Asian twitter content. This is not the first time Trump had a war of words with China (Hu, 2018). Trump made use of the pandemic to invoke the envy of Chinese, which unfortunately misleadingly convinced a group of people that the virus is caused and spread by Chinese. Trump's constant tirades against China and the Chinese people fueled the already worsening xenophobia in America. A study by Dhanani and Franz (2020) analyzed an online survey collected from 1141 US residents showed that the participants being more supportive on Donald Trump are more likely to have more bias toward Asian people. Therefore, the country saw an alarming increase of attacks against Asians. The hashtag StopAsianHate is in response to the unprovoked attack on Asians in the United States.

3. Dataset

First, we built the hate speech tweet dataset by scoring each COVID-19 related tweet with a hate score in span of 5 months (153) days. Then we perform factor analysis on the factors of hate speech, set the threshold, and filter out the factors that have a stable influence on the time series distribution. Finally, we integrated machine learning models to find the importance of each factor, i.e., we get which factors can influence hate speech most.

3.1. Data Collection

We built our overall dataset with three sub-datasets. The first dataset is COVID-19 Twitter chatter dataset(Banda et al., 2021), whose data gathering started from March 11th yielding over 3.3 million tweets a day. According to our selected period of time, we collected 83,624,190 unique tweets related to COVID-19 chatter. The second dataset is COVID-19 statistic dataset from The COVID-19 Tracking Project(Group, 2021), whose data includes daily increase in new case, death, negative test, hospitalization, total test in 5 months. The third one is Trump's Twitter posts dataset(Brendan, 2021). Twitter has permanently suspended Trump's account (January 8th, 2021), but this site checks Twitter every 60 seconds and records every Trump tweet into a database from 2016, which consists of 5454 trump's tweets in span of 5 months (153) days (from August 1, 2020 to December 31, 2020).

3.2. Data Processing

In general, data processing includes data selection and data cleaning. The specific techniques for each dataset we use are outlined as follows.

3.2.1. COVID-19 Twitter Chatter Dataset

In our tweet dataset consisting of 83,624,190 unique tweets, we filtered tweets whose content were not written in English and whose locations were not the United States to obtain 543,943 tweets. We input the content of each tweet into a hate speech detection model(Pérez et al., 2021). Pysentimiento is a multilingual Python toolkit for sentiment analysis and other social NLP tasks, and we use BERTpre-trained model(Nguyen et al., 2020) to initiate the parameters. After nearly 24 hours hate speech detection for all tweets, we obtained our dataset consisting of 40,579 hate speech tweets.

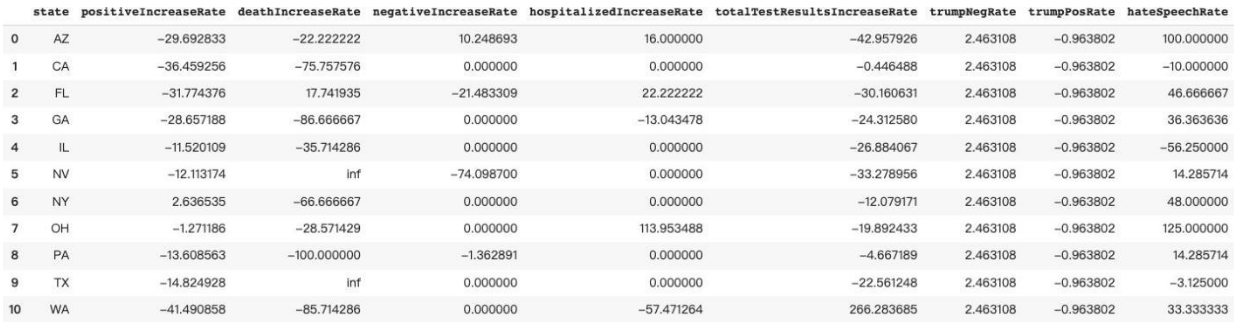
For 40,579 hate speech tweets, we divide them into 56 states and sort by date. We discard those states whose total number of hate speech is lower than 1000 in 5 months. 11 states and 34295 tweets are left, and they are AZ, CA, FL, GA, IL, NV, NY, OH, PA, TX and WA. 4 days are deleted from 153 days, because most of 11 states have 0 hate speech on those days.

3.2.2. COVID-19 Statistic Dataset

The factors we selected in COVID-19 statistic dataset are the new cases of COVID-19, death cases due to COVID-19, tests of COVID-19, hospitalizations of COVID-19, negative tests of COVID-19, which show us the relationship with hate speech, according to previous related works. The definition can be found in Table 1.

3.2.3. Trump's Twitter Dataset

The factor we selected in Trump's Twitter dataset is the sentiment scores of tweets. We filter out 5454 Trump's tweets in 5 months and analyse every tweet with sentiment analysis model(Pérez et al., 2021), and calculated sentiment scores in positive, negative and neutral. We use pre-trained models(Nakov et al., 2019)(Nguyen et al., 2020) in Pysentimiento for sentiment analysis.



	state	positiveIncreaseRate	deathIncreaseRate	negativeIncreaseRate	hospitalizedIncreaseRate	totalTestResultsIncreaseRate	trumpNegRate	trumpPosRate	hateSpeechRate
0	AZ	−29.692833	−22.222222	10.248693	16.000000	−42.957926	2.463108	−0.963802	100.000000
1	CA	−36.459256	−75.757576	0.000000	0.000000	−0.446488	2.463108	−0.963802	−10.000000
2	FL	−31.774376	17.741935	−21.483309	22.222222	−30.160631	2.463108	−0.963802	46.666667
3	GA	−28.657188	−86.666667	0.000000	−13.043478	−24.312580	2.463108	−0.963802	36.363636
4	IL	−11.520109	−35.714286	0.000000	0.000000	−26.884067	2.463108	−0.963802	−56.250000
5	NV	−12.113174	inf	−74.098700	0.000000	−33.278956	2.463108	−0.963802	14.285714
6	NY	2.636535	−66.666667	0.000000	0.000000	−12.079171	2.463108	−0.963802	48.000000
7	OH	−1.271186	−28.571429	0.000000	113.953488	−19.892433	2.463108	−0.963802	125.000000
8	PA	−13.608563	−100.000000	−1.362891	0.000000	−4.667189	2.463108	−0.963802	14.285714
9	TX	−14.824928	inf	0.000000	0.000000	−22.561248	2.463108	−0.963802	−3.125000
10	WA	−41.490858	−85.714286	0.000000	−57.471264	266.283685	2.463108	−0.963802	33.333333

Figure 1: Data structure

3.2.4. Data Structure

Y - Hate Speech Label We choose the data of 11 states from our hate speech dataset. We calculate the increase rate of every day's hate speech tweets, based on the previous day's metric.

X - Hate Speech Factors We choose the 11 states data from COVID-19 Statistic dataset. We normalize each factor and then calculate the increase rate of each factor, based on the previous day's value. The normalization methodology aims to normalize all the data sharing the same region at different time frame, by subtracting their means and then applying the standard deviation.

All the data are made into timely and state-related tables. i.e. the data of 2020/8/3 is calculated from 2020/8/2, shown in Figure 1. Thus, we have 148 such tables.

3.3. Data Visualization

First, we can visualize all hate speech factors from 8/1/2020 to 12/31/2020 in Figure 7. Our study wants to explore the effect of factors on hate speech over a 5-months period, rather than at particular time points. Therefore, we plot the trend of factors X over 153 days, but we are unable to do so in a way that clearly identifies the value for each day. Through these graphs we can observe the trend, where the histogram shows the daily increase number and the line reveals the oscillation of daily increase rate .

For the hate factor, we similarly visualize the trend of hate speech over 5 months, as shown in Figure 3.

ISSN:2790-5179 DOI: 10.56028/iajhss.1.2.1

X	New Case Death Negative Test Hospitalization Total Test Trump's tweet	Daily increase in confirmed plus probable cases of COVID-19 based on the previous day's value. Daily increase in fatalities with confirmed OR probable COVID-19 case diagnosis. Daily increase in number of unique people with a completed PCR test that returns negative. Daily increase in number of individuals who have ever been hospitalized with COVID-19. Daily increase in test including antigen tests and viral (PCR) tests. Daily average sentiment score of Trump's tweets, obtained from sentiment model
Y	Hate speech	Daily increase in hate speech tweets.

Table 1: Definition of Hate Speech factors (X) and label (Y)

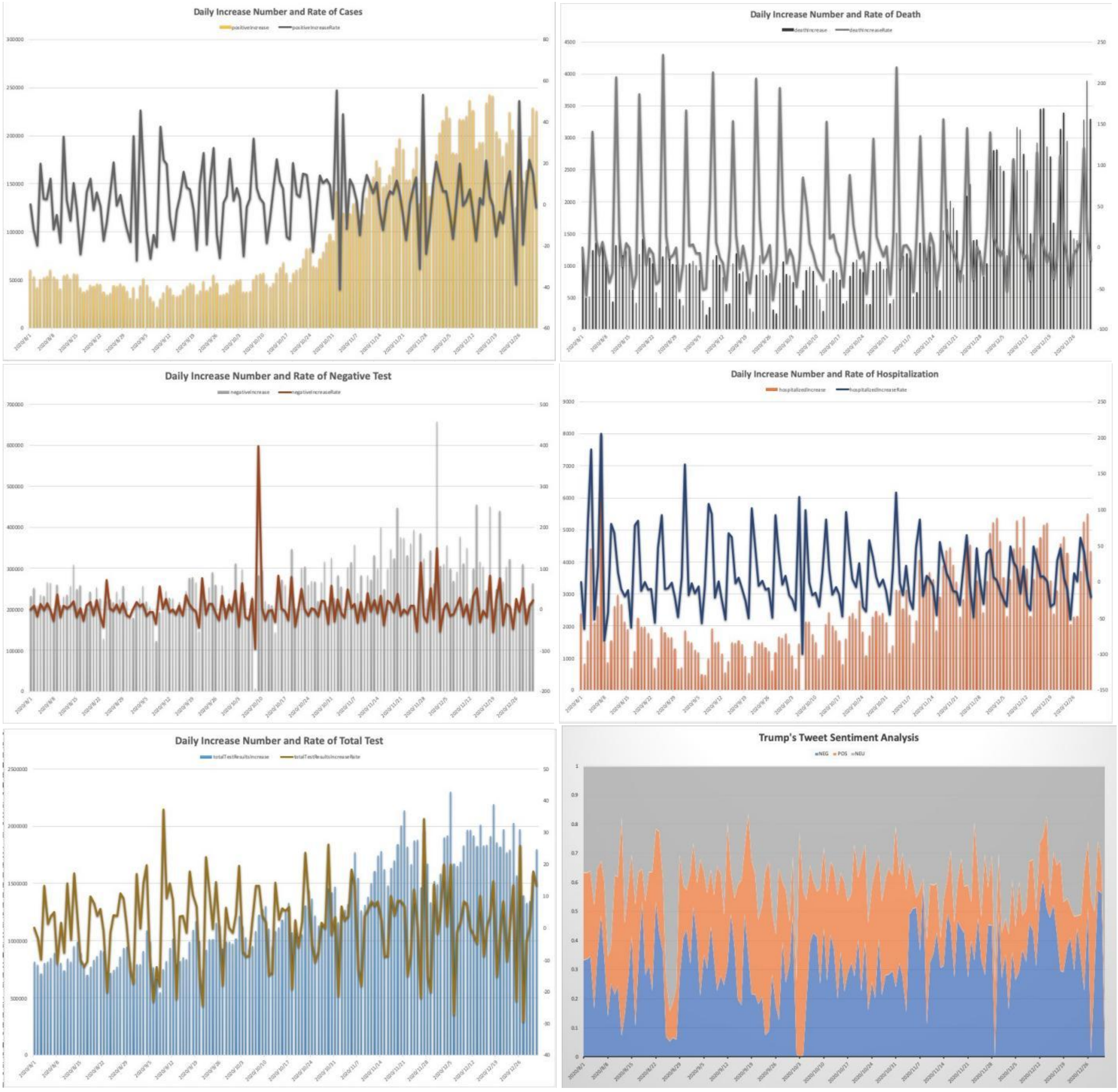


Figure 2: Hate speech factors trend from 08/01/2020 to 12/31/2020. For COVID-19 data, the histogram shows the daily increase amount (values on left side of graph) and the line reveals the daily increase rate (values on right side of graph). For trump's tweet sentiment data, the daily average scores of all tweet sentiment scores are categorized into positive, negative and neutral, whose sum of score is 1.

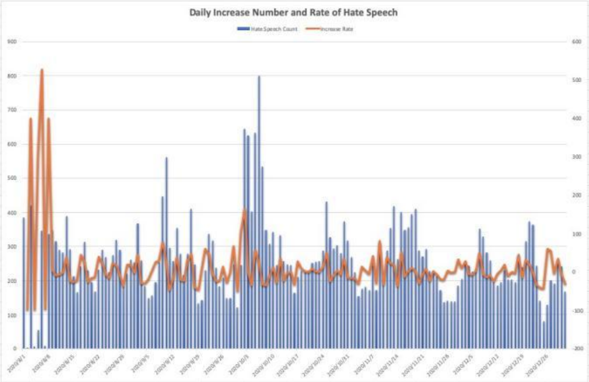


Figure 3: Hate speech factors trend from 08/01/2020 to 12/31/2020. The histogram shows the daily increase number (values on left side of graph) and the line reveals the daily increase rate (values on right side of graph).

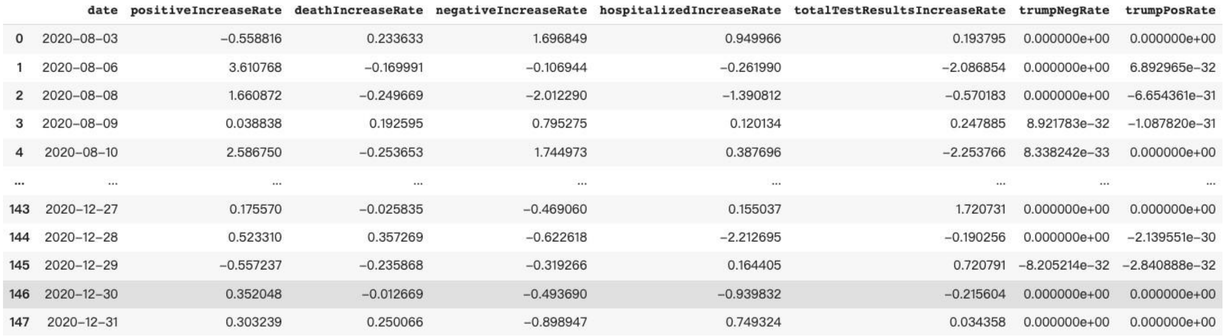
4. Methodology

We will introduce the stable factors selection, topic models and factor importance models.

$$n(v) = \frac{4N\pi}{V} * \frac{m}{2 * \pi * k * T}^{\frac{3}{2}} * v^2 * e^{\frac{-mv^2}{2*k*T}} dv$$

4.1. Stability Factor Selection

From Figure 1, every day's data consists of 11 states' data. On cross-sectional data, by using the linear regression model, the hate speech increase rate Y and the increase rate of factors X are linearly regressed to obtain the coefficient β of each factor on that day. In this way, a β sequence can be obtained for each factor, as shown in Figure 4.



	date	positiveIncreaseRate	deathIncreaseRate	negativeIncreaseRate	hospitalizedIncreaseRate	totalTestResultsIncreaseRate	trumpNegRate	trumpPosRate
0	2020–08–03	–0.558816	0.233633	1.696849	0.949966	0.193795	0.000000e+00	0.000000e+00
1	2020–08–06	3.610768	–0.169991	–0.106944	–0.261990	–2.086854	0.000000e+00	6.892965e–32
2	2020–08–08	1.660872	–0.249669	–2.012290	–1.390812	–0.570183	0.000000e+00	–6.654361e–31
3	2020–08–09	0.038838	0.192595	0.795275	0.120134	0.247885	8.921783e–32	–1.087820e–31
4	2020–08–10	2.586750	–0.253653	1.744973	0.387696	–2.253766	8.338242e–33	0.000000e+00
...	...	...	...	...	...	...	...	...
143	2020–12–27	0.175570	–0.025835	–0.469060	0.155037	1.720731	0.000000e+00	0.000000e+00
144	2020–12–28	0.523310	0.357269	–0.622618	–2.212695	–0.190256	0.000000e+00	–2.139551e–30
145	2020–12–29	–0.557237	–0.235868	–0.319266	0.164405	0.720791	–8.205214e–32	–2.840888e–32
146	2020–12–30	0.352048	–0.012669	–0.493690	–0.939832	–0.215604	0.000000e+00	0.000000e+00
147	2020–12–31	0.303239	0.250066	–0.898947	0.749324	0.034358	0.000000e+00	0.000000e+00

Figure 4: β sequences for each factor.

The regression coefficient β represents the influence of the factor of that day on the hate speech. When the absolute value of β is large, the factor of that day is considered to have a significant influence on Y. We hope to find the factors that have a stable impact on Y for a long time in the past. For this reason and based on the Sharpe ratio(Sharpe, 1994), we construct a stability formula for β.

$$stability = \frac{E(abs(\beta_t))}{std(\beta_t)},$$

$$t = 08/01/2020, 08/02/2020, \ldots, 12/31/2020$$

We calculated the stable scores for each factors, as shown in Table 2. A larger value indicates that the factor has a more stable influence on Y over a long period of time, so we selected the factor with a largervalue. By comparing the graphs and setting the threshold value (0.5), we select the factors with stable influence on the time series and discard hospitalization factor.

Factor	Stability score
positiveIncreaseRate	0.522074
deathIncreaseRate	0.626562
negativeIncreaseRate	0.518176
hospitalizedIncreaseRate	0.400534
totalTestResultsIncreaseRate	0.566007

Table 2: Stability scores of COVID-19 factors

Parameters	Description
max depth	The maximum depth of the tree, deeper trees tend to be overfitting and shallow trees tend to be underfitting.
learning rate	The weight reduction factor of each weak learner also known as step, which is in the range [0, 1].
n estimators	The maximum number of weak learners, always considered with learning rate.
early stopping rounds	Stop training early and prevent overfitting.

Table 3: Parameters description of random forest and Xgboost.

4.2. Model

We will talk about topic model for finding the most popular topics in hate speech and factor importance model for scoring the factors.

4.2.1. Topic Model

Topic Models are a type of statistical language models used for uncovering hidden structure in a collection of texts, it solves task of dimensional reduction, unsupervised learning, and tagging. Latent Dirichlet Allocation Model(Blei et al., 2003) is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities. We build a LDA model with 6 topics where each topic is a combination of keywords, and each keyword contributes a certain weight to the topic.

4.2.2. Factor Importance Model

From the above process, we have screened out the factors with relatively stable effects. There may be a linear or non-linear relationship between these factors and hate speech. In order to measure the magnitude of the correlation, we need to score the importance of the factors. For this purpose, we adopted decision tree. It is because decision tree method not only has a good fitting effect on the linear and non-linear relationship, but also can evaluate the classification ability of the factor.

In the model selection, we adopt random forest(Liaw and Wiener, 2002) and Xgboost(Chen and Guestrin, 2016) models, both of which are ensemble methods of decision trees and have better fitting effects.

The random forest is a parallel algorithm model with low variance but low accuracy. Xgboost is a serial algorithm model with high accuracy but may lead to overfit and thus high variance. Therefore, we combine and compare the results of both scores on factor importance to obtain more credible analysis conclusions.

The main parameters of Xgboost and random forest during the training process are shown in Table 3.

The algorithm is as follows:

1. Divide the dataset into a training set and a validation set. The division ratio is 0.75 and 0.25, respectively.
2. Find optimal parameters on the training set sample by Grid Search method. The validation set is used to control whether to stop training early during training.
3. For the optimal parameters obtained in step 2, construct XgBoost and random forest on the full dataset and fit the data.
4. Obtain the importance of the factors by trained classifiers.

5. Result and Discussion

We will also talk about the results of topic model and factor importance model.

5.1. Topic Model Result

Inspired by (Shi and Chen, 2020), we need to name the topic from the keywords. For example, for topic 0, we have 10 keywords, we find that most of keywords are related to preventive measure issue, such as wear, mask, stay, home, and we name topic 0 as preventive issue topic. The same process applies to rest of them.

Id	Topic	High probability vocabulary
0	Preventive Issue	mask,country,want,home,die stay,state,government,wear,stop
1	Emotion Issue	fuck,lies,white,going,die american,virus,america,shut,ojos
2	Election Issue	vote,america,biden,save,blue tono,piel,harris,harris2020,end
3	Vaccine Issue	vaccine,fuck,bitch,away,stop country,real,did,killing,mask
4	China Issue	biden,china,vote,president,virus america,does,money,pandemic,right
5	Pandemic Risk	risk,symptom,dead,country,america pandemic,world,president,died,virus

Table 4: Result of LDA and six topics of hate speech

For every hate speech tweet, we calculate the six topic scores, and choose the topic with max topic score. We can see the distribution of six hate speech topics in Figure 5. Thus, preventive issue, pandemic risk issue and china issue are the main topics of hate speech tweets.

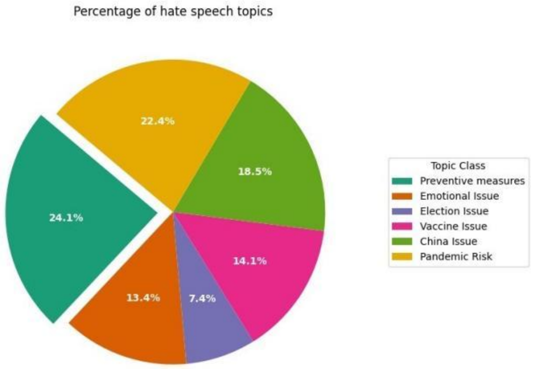


Figure 5: Distribution of six hate speech topics

5.2. Factor Importance Model Result

By grid search, we choose the parameters of model with the lowest mean error in the validation set, and and the optimal parameters corresponding to Xgboost and random forest are listed on Table 5

Parameters	Xgboost	Random forest
max depth	3	3
learning rate	0.3	-
n estimators	20	10

Table 5: Optimal parameters of Xgboost and random forest

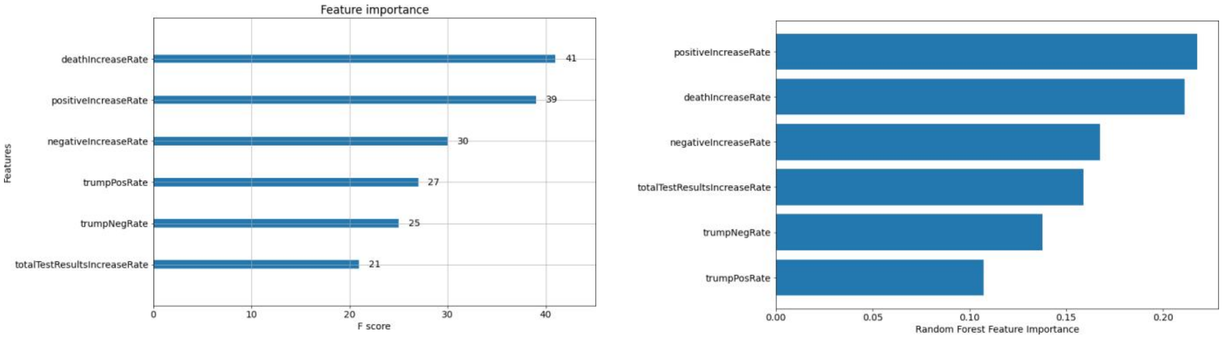


Figure 6: Feature importance score of Xgboost (left) and Random forest (right).

The results of the random forest and Xgboost are shown in Figure 6. The results of the Xgboost model show that the increase rate of death, new cases (positive) and negative test are ranked high, indicating that these 3 factors have the greatest influence on the the increase rate of hate speech on Twitter. The results of the random forest model show that new cases (positive), death and negative test are ranked in different order, but they are still the top three factors.

Therefore, we can conclude that new cases of COVID-19, death of COVID-19 and negative test of COVID-19 have the most influence on hate speech on Twitter. This finding also suggests that the constant updating of daily new cases during the COVID-19 epidemic is a major trigger for hate speech, while the death toll continues to cause panic and the negative sample factor may be an inconvenience for people who have to go to pharmacy for testing.

6. Conclusion

This research summarize the situation of hate speech towards Chinese people on Twitter, in span of 5 months (153) days. Considering only the tweets written in English and within the USA. A total of 40,579 tweets out of 543,943 tweets obtained contains hate speech content. Subsequently, they analyzed these tweets using Pysentimiento model which is based on BERT models and, Latent Dirichlet Allocation Model (LDA). The results indicate that there are substantial associations between the amount of hate speech and the increased rate of deaths due to COVID-19, increase rate of new COVID-19 cases, and negative tests rate. Another finding is that the main topics of hate speech tweets are mostly related to the prevention of new cases, the risk of the pandemic as well as the hate against Chinese.

7. Future Work

For future work, we will investigate and analyze the Twitter accounts creating and sharing hate speech content, i.e. their age groups, educational and occupational background ..etc. Then find the main characteristics that they share. Another direction is to study the demographic affects of the states with the highest number of hate speech content. Additionally, to extend the work to include more popular platform (e.g., Reddit) and more common languages in the US (e.g., Spanish). Therefore, we could potentially use current open source machine learning models available online that classifies the type of users based on the hatefulness of the speech as well as their time spend on posting hateful speech, and look into some specific group of people include but not limited teenagers, people who were diagnosed of COVID-19 and politicians. This could be useful when we would like to know more about the majority type of people who are active users that spreads the hate speech and how irrational do they react when there is an increase of new COVID-19 cases.

In addition, in terms of the accounts themselves, we would dive deeper into the robot accounts that would lead to a skewed data sample.

References

[1] 2021. attacks on asian-americans in new york stoke fear anxiety and anger 2021. https://www.nytimes. com/2021/02/26/nyregion/asian-hate-crimes-attacks-ny.html. Online; accessed 15 De- cember 2021.

[2] 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):315–324.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022.

[4] Brendan. 2021. Trump twitter archive v2.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, pages 785–794, New York, NY, USA. ACM.

[6] Weine Dai, Guangteng Meng, Ya Zheng, Qi Li, Bibing Dai, and Xun Liu. 2021. The impact of intolerance of uncertainty on negative emotions in covid-19: Mediation by pandemic-focused time and moderation by perceived efficacy. International Journal of Environmental Research and Public Health, 18(8):4189.

[7] Lindsay Y. Dhanani and Berkeley Franz. 2020. Unexpected public health consequences of the covid-19 pandemic: a national survey examining anti-asian attitudes in the usa. International Journal of Public Health, 65(6):747–754.

[8] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. 2021. Stigmatization in social media: Documenting and analyzing hate speech for covid -19 on twitter.

[9] Chelsea Gray and Kirstine Hansen. 2021. Did covid-19 lead to an increase in hate crimes toward chinese people in london? Journal of Contemporary Criminal Justice, 37(4):569–588.

[10] The Atlantic Monthly Group. 2021. The covid tracking project.

[11] Cornelia Herbert, Alia El Bolock, and Slim Abdennadher. 2021. How do you feel during the covid-19 pandemic? a survey using psychological and linguistic self-report measures, and machine learning to investigate mental health, subjective experience, personality, and behaviour during the covid-19 pandemic among university students. BMC Psychology, 9(1).

[12] W. Hu. 2018. Trump’s china policy and its implications for the “cold peace” across the taiwan strait. China Review, 18:61–88.

[13] Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomforest. R News, 2(3):18–22. Ivan Mehta. 2021. [link].

[14] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2019. Semeval-2016 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.01973.

[15] 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, pages 9–14.

[16] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

[17] Mishal Reja. 2021. Trump’s ’chinese virus’ tweet helped lead to rise in racist anti-asian twitter content: Study. William F. Sharpe. 1994. The sharpe ratio. The Journal of Portfolio Management, 21(1):49–58.

[18] Wen Shi and Changfeng Chen. 2020. Issue highlighting and relevance construction: Twitter social robot’s construction of covid-19 epidemic discussion. Modern Communication: Journal of Communication University of China, 10.

[19] Zainab Tariq Soomro, Sardar Haider Waseem Ilyas, and Ussama Yaqub. 2020. Sentiment, count and cases: Analysis of twitter discussions during covid-19 pandemic. In 2020 7th International Conference on Behavioural and Social Computing (BESC), pages 1–4.

[20] Human Rights Watch. 2020. defending human rights worldwide 2020. https://www.hrw.org/world-report/2020. Online; accessed 15 December 2021.

A. Appendix

A.1 Robot Account

Increase of robot account users at hate speech tweets spike

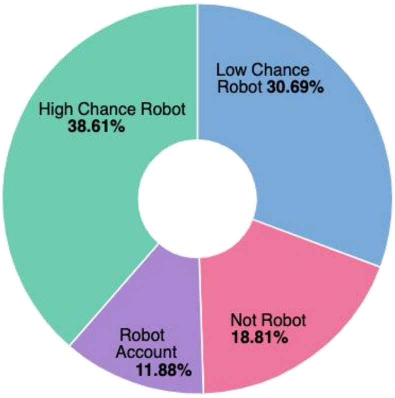


Figure 7: user robot score distribution at 08/08/2020

Integrating one of the open source API-Rapid, that assigns the detection scores which evaluates the likelihood of the account being fake, we perform the analysis of whether intentional robot account posts to cause public confederate is one of the factors that leads to a significant increase in the numbers of hate speech at given period. Rapid API used a machine learning model that detects robot tweeter users based on the amount of posts, the amount of followers that are fake and the the frequency of the posts etc. The higher the robot score of the user id is, the more likely that this account an robot account. After making a comparing the distribution as well as the average score of the users between July and Aug which has a significant increase We noticed the higher weight of the fake account users in the population, it could lead us to a study of how it would effect analysis of the correlation between new cases and the amount of hate speech posts in this research.

Given the reasoning that Trump's posts on Twitter constitutes one of the factors that would simulates the increase of the hate speech posts, we categorize the accounts based on the scores that the model assigned and dive into the portion of potential fake accounts among all users. It shows a accompany of more fake account at the point of spike when Donald Trump makes multiple hate speech post, which leads to a future study of the political propaganda in correlated to the tweeters as a confederate.

B. Frequent words appeared in hate speech tweet

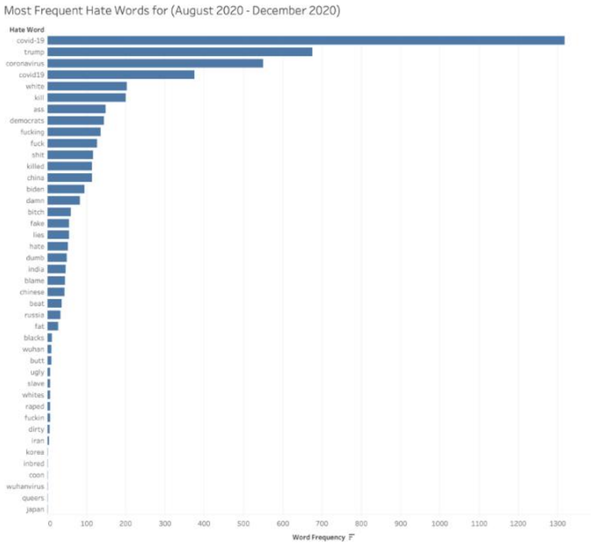


Figure 8: Frequent words appeared in hate speech tweets.