COVID-19 and Racism: An Ensemble Machine Learning Approach to Understanding the Effect of a Global Pandemic on Twitter Users' Attitudes

Bokang Jia, Domnica Dzitac, Samridha Shrestha, Komiljon Turdaliev, and Nurgazy Seidaliev
Computational Social Science, Spring 2020, NYUAD

Advised by: Prof. Talal Rahwan and Prof. Bedoor AlShebli

ABSTRACT

It is thought that the COVID-19 outbreak has significantly fuelled racism and discrimination, especially towards Asian people[10]. In order to test this hypothesis, in this paper, we build upon existing work in order to classify racist tweets before and after COVID-19 was declared a global pandemic. To overcome the difficult linguistic and unbalanced nature of the classification task, we combine an ensemble of machine learning techniques such as a Linear Support Vector Classifiers, Logistic Regression models, and Deep Neural Networks. We fill the gap in existing literature by (1) using a combined Machine Learning approach to understand the effect of COVID-19 on Twitter users' attitudes and by (2) improving on the performance of automatic racism detectors. Here we show that there has not been a sharp increase in racism towards Asian people on Twitter and that users that posted racist Tweets before the pandemic are prone to post an approximately equal amount during the outbreak. Previous research on racism and other virus outbreaks suggests that racism towards communities associated with the region of the origin of the virus is not exclusively attributed to the outbreak but rather it is a continued symptom of deep-rooted biases towards minorities[9]. Our research supports these previous findings. We conclude that the COVID-19 outbreak is an additional outlet to discriminate against Asian people, instead of it being the main cause.

This report is submitted to NYUAD's Computational Social Science Course in fulfillment of the class requirements.

جامعة نيويورك ابوظبي
NYU ABU DHABI

, ,

KEYWORDS

COVID-19, Coronavirus, Machine Learning, Natural Language Processing, Automatic Hate-Speech Detection, Racism

Reference Format:
Bokang Jia, Domnica Dzitac, Samridha Shrestha, Komiljon Turdaliev, and Nurgazy Seidaliev. . COVID-19 and Racism: An Ensemble Machine Learning Approach to Understanding the Effect of a Global Pandemic on Twitter Users' Attitudes. In . 8 pages.

1 INTRODUCTION

In December of 2019, a new disease of the coronavirus family, COVID-19, was detected in Wuhan, China. The World Health Organization (WHO) has declared the novel coronavirus a global pandemic due to an exponential increase in COVID-19 infections in the past months reaching, as of May 1st 2020, over 3.3 million cases and resulting in approximately 234,140 deaths worldwide[11]. Besides its effects on global health, the COVID-19 outbreak has significantly impacted the global economy, travel, political dynamics, and the public's actions as a whole[2]. Moreover, it is thought that the COVID-19 outbreak has significantly fuelled racism and discrimination, especially towards Asian people[10]. Over the past months, extremely influential politicians made public declarations that further associate the virus to China[10]. In addition, the number of racist attacks, including hate crimes, has increased severely since the pandemic was declared a global threat to our health[4]. Nonetheless, it is not the first time in history when an infectious disease outbreak is associated with communities that later have to suffer social, economic or political consequences. Other diseases such as SARS, the Middle East Respiratory Syndrome or Ebola had a negative impact on the communities associated with the cause of the outbreak[7][9]. Previous research that investigated the effect of Ebola on discrimination against African residents of Hong Kong suggests that social stigmatization of Africans was a continued symptom of deep-rooted biases towards minorities in Hong Kong not solely attributable to the 2014

Ebola outbreak. Rather, the outbreak was an additional excuse to discriminate against the minority African community in Hong Kong[9]. Further, other work that studied the effect of SARS on racism in Toronto, Canada suggests that the virus outbreak in the city affected citizens' attitude towards people who may be associated with the country of the origin of SARS[1]. As a further matter, recent research shows that these types of stigmas affect negatively the research and academic community as well[7]. However, previous work on the connection between infectious disease outbreaks and racism did not look quantitatively at the impact of social media on racism towards these targeted communities that are associated negatively with the region of the origin of the virus. In this paper, we are filling this literature gap by studying the connection between racism and social media. In order to do that, we analyse the dynamics in people's attitude towards Asian people on Twitter before and after the 29th of January 2020, the date when COVID-19 was officially declared a threat to our global health.

2 RELATED WORK

Previous work has looked at multiple training methods for datasets in order to classify hate speech or other similar linguistic tasks and the best-suggested method of training is the Linear Support Vector Machine Learning Model (SVC)[13]. Thus, in our study we use a linear SVC classifier to identify racist tweets that target Asian people between 1st of January and 31st of March 2020. The classifier is an updated version inspired by the open-source system provided by Davidson et al.[16] with adapted features that are meant to identify racism towards Asian people instead of just hate speech, as proposed in the original system. To further reduce overfitting, we use two other methods of training, namely a logistic regression and a Long Short Term Memory (LSTM) recurrent neural network (RNN)[12]. Moreover, other related work concludes that pre-trained embedding weights for Twitter datasets improve the performance of the classification tasks and thus, we included this in our classifiers[6]. Previous research also shows that in order to extract a deeper level of features, a Deep Neural Network Model is required[17]. However, all these methods of training models and improvements added to existing systems are still highly dependent on the linguistic nature of the classification task. According to previous literature, hate speech, particularly racism, is challenging to classify because of the high amount of offensive language and swear words on online platforms [14]. The key difference between the two is based on linguistic distinction [8]. Previous related work also suggests that hate speech towards certain groups comes from a set of stereotypical words, used in either a positive or negative way due to how previous research treated the hate-speech identification as a matter of word sense disambiguation [8]. The training data sets take the above definition of hate speech into consideration.

3 DATA

In order to investigate the dynamics in users' attitudes on Twitter towards Asian people before and after the COVID-19 outbreak, we use two different datasets of Tweets for training, validating and testing our models and two datasets of Tweets for our actual study. The two datasets used for training are collected by (1) Davidson et al[16] and (2) Waseem and Hovy [15]. Both datasets are manually annotated and contain only tweets written in English. The dataset provided by Davidson et al. has 24K tweets that contain hate speech keywords which were collected by using a crowd-sourced hate speech lexicon. These tweets are labelled into three categories: "hate speech", "offensive language" and "neither". Only 5% of the tweets in this dataset were labelled as "hate speech" and 76% were labelled as "offensive language". The remaining were labelled as "neither"[16]. On the other hand, the dataset provided by Waseem and Hovy contains 16k tweets that contain racist and sexist instances of hate speech. These tweets are labelled into three categories: "racism", "sexism" and "neither". The annotators labelled 20% of the Tweets as "sexism" and 12% as "racism". The remaining were labelled as "neither"[15]. For the sake of our study, we only considered the tweets labelled as "racism" and "neither". Moreover, in order to conduct our study, we used the first available COVID-19 dataset of tweets collected by Chen et al[3]. This dataset contains tweet ids of users that mention the pandemic since January 22, 2020. We downloaded the tweet ids that were collected by the above team based on keywords listed in their open-source documentation. The tweets we downloaded were shared by users between January 22 and March 31, 2020. In accordance with Twitter terms, we hydrate the tweet ids in order to convert them to actual tweets by using the Twarc tool. All the tweets from all three datasets had to undergo a preprocessing procedure. This includes standardizing counts of URLs and mentions, removing punctuation and the excess of whitespaces, as well as tokenizing the tweet by lowercasing and stemming the words. Besides these standard preprocessing tasks, we cleaned the COVID-19 dataset to account only for tweets written in English and that are geo-located in the United States of America. As a further matter, we used this data to train and run our models and we identified all tweets that were labelled as "racist" towards Asian people. Then, we used the tweet ids of all these racist tweets in order to trace down these users' posts before their first mention of the pandemic. In accordance with Twitter terms, we only had access to the last 32,000 tweets for each

	precision	recall	f1-score	support
0	0.46	0.60	0.52	164
1	0.96	0.91	0.94	1905
2	0.82	0.95	0.88	410
accuracy			0.89	2479
macro avg	0.75	0.82	0.78	2479
weighted avg	0.91	0.89	0.90	2479

Figure 1: SVC Model Metrics

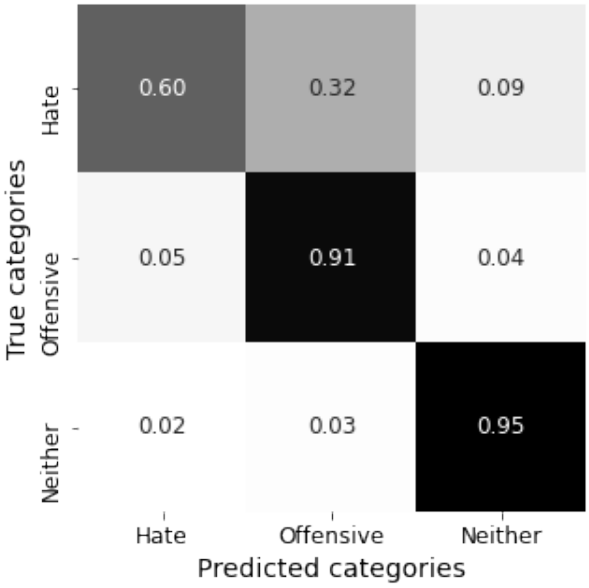

Figure 2: Confusion Matrix SVC Model

	precision	recall	f1-score	support
0	0.50	0.70	0.58	164
1	0.98	0.91	0.94	1905
2	0.84	0.96	0.90	410
accuracy			0.91	2479
macro avg	0.77	0.86	0.81	2479
weighted avg	0.92	0.91	0.91	2479

Figure 3: Ensemble Model Metrics

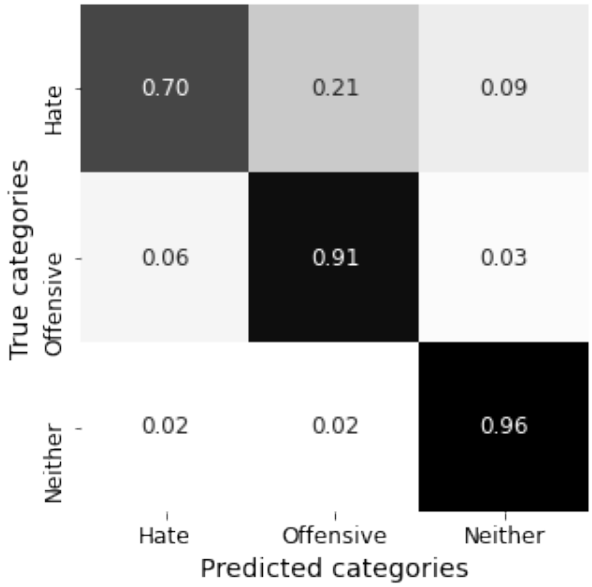

Figure 4: Confusion Matrix Ensemble Model

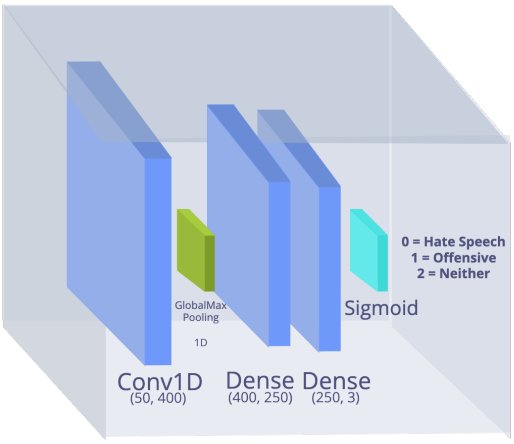

Figure 5: Neural Network model

user, however this was enough to form a new dataset of tweets before the COVID-19 outbreak. We used the latter dataset to see whether these users had a negative attitude on Twitter towards Asians even before the pandemic. This dataset has been preprocessed in a similar manner as the COVID-19 dataset.

4 METHODOLOGY

Analysing and quantifying the dynamics in the attitudes of Twitter users towards Asian people before and after the COVID-19 pandemic is not an easy task due to the subtle linguistic distinction between hate speech, specifically racism, and offensive language. Since existing methods primarily focus upon hate speech, we have contributed to the research literature by developing our own improved classification method consisting of an Ensemble Network of multiple detectors. We trained this network on datasets collected by both

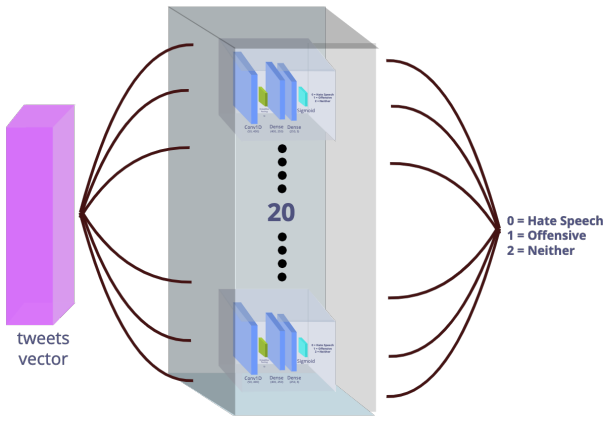

Figure 6: Neural Network ensemble model

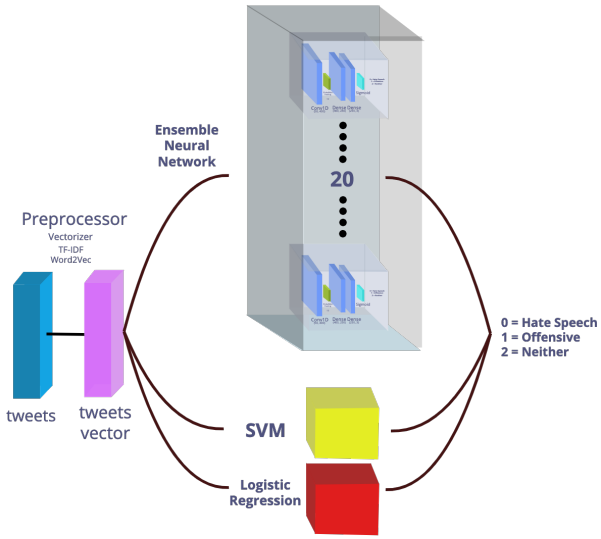

Figure 7: Combined ensemble model

Davidson et al, and also Waseem and Hovy, as mentioned above. This enabled us to build a strategy that was superior to existing methods. As a starting point, we improved upon the hate speech classifier created by Davidson et al. Rather than simply using an SVM model, our methodology combines this strategy with a Logistic Regression model. We extracted features by running the pre-processing pipeline as proposed by Davidson et al. which utilized Term Frequency-Inverse Document Frequency (TF-IDF), Penn POS tagging and Porter Stemming on unigrams, bigrams, and trigrams. This helps highlight important features within the tweets. These two models were trained using the hate speech dataset of tweets provided by Davidson et al. The output of these networks were put through an ensemble voting mechanism.

To boost the accuracy further and to introduce regularization, we also ran these features through a ensemble model of 20 Deep Neural Network (DNN) models. The tweets in this ensemble model framework are first preprocessed by generating convolutional neural network word embeddings from the raw twitter data. This embedding generation uses a pre-trained WordtoVec NLP model that was pretrained on 400 million tweets [17]. These DNN models were trained using the dataset collected by Waseem and Hovy. The output of this DNN collection was also fed through the voting mechanism. In order to reduce overfitting, we use hard voting method which results in the most accurate vote for the three systems. In addition to the DNN, we also applied a LSTM model, but unfortunately the results indicated that a LSTM would not be optimal for tweets, which are short and contain dense meanings. The original system proposed by Davidson et al. had a precision and recall of 46% and 60%, respectively See Fig 1,2. Our novel ensemble method produced a precision of 50% with a recall of 70%. See Fig 3,4

We limit our study to the Twitter users in the US to prevent any country bias. Majority of the tweets lack explicit country information, but there is a plethora of literature on the detection of tweet country based on its several features. In this paper, we used a tweet country classification model developed by Zubiaga et al [18]. The model used eight tweet-inherent features for classification such as tweet content, the user's self-reported location and the user's real name and was trained on two datasets, collected a year apart from each other. From the tests we conducted, the model had on average more than 85% accuracy on detecting the tweets from the US, which we deemed satisfactory.

Finally, the classification system was also passed through a RegEx keyword filter, for example "(R|r)acis.*" and "(C|c)hin.*", in order to adapt the system with features that are meant to identify racism towards particular groups of people instead of just hate speech, as proposed in the original system.

5 RESULTS

We obtain three sets of results. The first set of results correspond to general rate of hateful tweets by the US Twitter users. Specifically, we examine whether users become increasingly more hateful after the outbreak of COVID-19 and its global spread.

We ran our classifier on the country-filtered COVID-19 related dataset. We find that collective number of hateful tweets is generally steady during this period *See Fig 4.* However, there were two large spikes in the number of hate-speech tweets early February (1) and early March (2). This corresponds to 1. the rapid spread of the virus within China and first wave of significant active cases outside China and 2. the significant increase in the number of cases globally, specifically in the US. *See fig 11*

Note: There is a known gap in Chen et al's Covid-19 database around Feb 23 due to connectivity issues, this decrease is reflected in Figures 8, 11 and 12 [3]

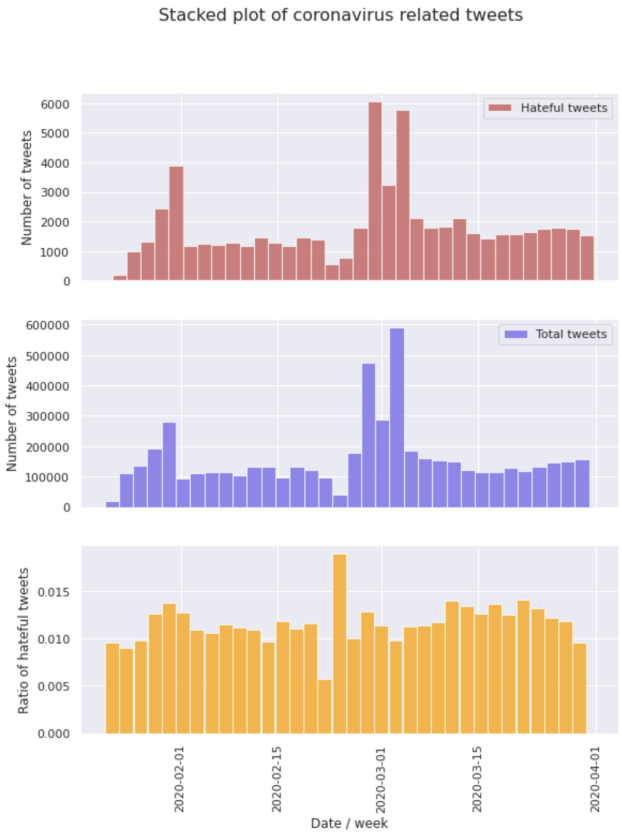

Figure 8: Classified hate speech tweets related to COVID-19

We note that the total number of tweets (before classification) varies over time. This could affect the conclusions regarding individuals. Therefore, we also examined if the number of hate-speech tweets corresponds to an increase in racist and hateful discussion on COVID-19. I.e, do people individually become more hateful? A graph of the ratio of hate speech to total tweets was plotted *See Fig 8.* We observe that the overall level of hate-speech remained steady (at around 1% of total tweets). This is a surprising finding as it indicates that the level of hate speech around COVID-19 has not increased significantly over the course of the pandemic.

In order to examine if the recent pandemic has caused people to become more racist, we would need to control for people's historic tweet behaviour going back to before the pandemic, and after. We selected approximately 60 thousand users who posted hateful tweets from the COVID-19 dataset during the month of January, whom we label as the "haters". We selected the month of January as it allowed us to have a fairly equal representation of the 4 months before and after the initial reaction to COVID-19. The two separated periods correspond to September 2019 - January 2020 and January 2020 - May 2020, respectively.

Of this, we scraped 20 thousand Twitter timelines through the Twitter API. Because of the limits posed by Twitter, we were only able to scrape the latest 1000 tweets of any given user. For many users this only covers the recent few weeks, however, for the some it covers the entire pre-pandmic and post-pandemic period. We aggregated the "hater" timeline tweets together, which produced the recent-heavy graph of hate tweets near May *See Fig 9.*

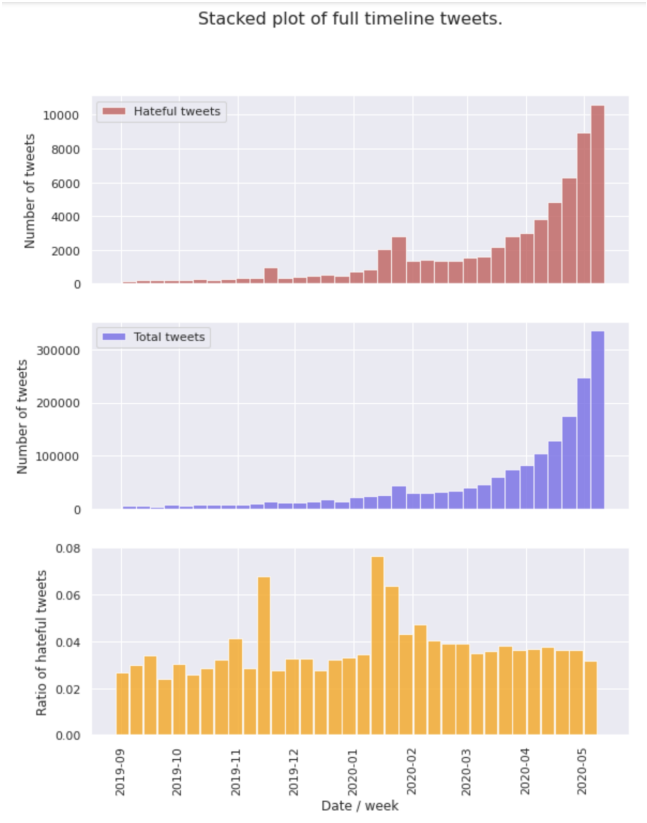

Figure 9: Classification of hate speech on racist users timelines

When we normalize this by the total number of timeline tweets, we obtain the ratio of real hate tweets of these users *See Fig 9*. It is important to note that while there are small spikes in the hateful tweets in February and March (corresponding to the rising cases in China and subsequently in the US), the baseline level of hate-speech does not change significantly in the pre-pandemic and post-pandemic periods. These "hater" users were already posting high level of hate speech before the pandemic and the COVID-19 appears to be just another context in which they continue to express their hate speech. This figure also shows that these "haters" are also the highest source of racist hate-speech with approximately 4% of their tweets being hateful as compared to the average of 1% *See Fig 8*

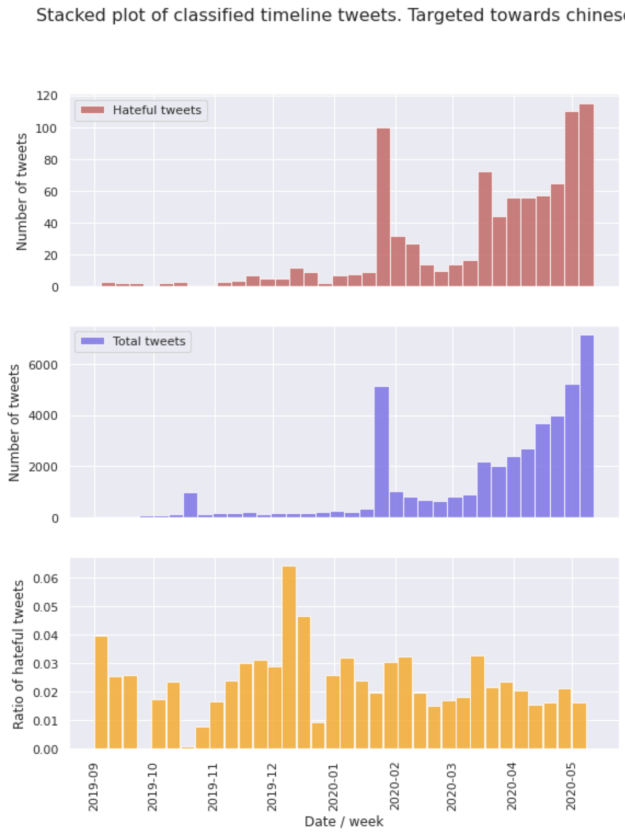

Figure 10: Classified and normalized hate speech timelines towards Chinese people.

In the third set of results *(See Fig 10)*, we examine whether there was an increase in the rate of racism on Twitter specifically targeted at the Chinese. Similar to the findings on general racism, we find that while the absolute volume of hate-speech increased towards Chinese people, the actual proportion of tweets towards them remained fairly constant - and in fact, slightly decreased. This goes to demonstrate that the pandemic did not cause people to become more racist towards the Chinese.

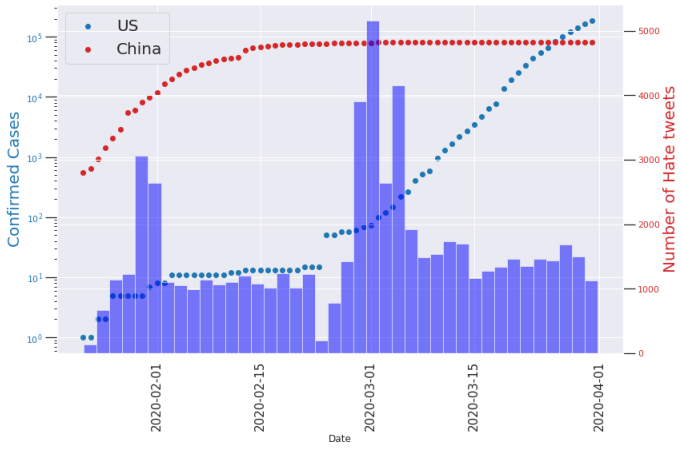

Figure 11: Hate Speech vs COVID-19 cases in the US and China

In *figure 11* we also overlay the cumulative number of real COVID-19 cases as obtained from Johns Hopkins University [5]. We do this for the primary countries in question, China and the US. In the figure, we observe that a increase in the number of COVID-19 cases can be seen directly preceding a large increase in the number of hate tweets. This occurred for China late January, and again for the US early March. This is likely correlated with panic within the community about an imminent spread of COVID-19 - leading to a temporary increase in racist, isolationist attitudes - which do not remain.

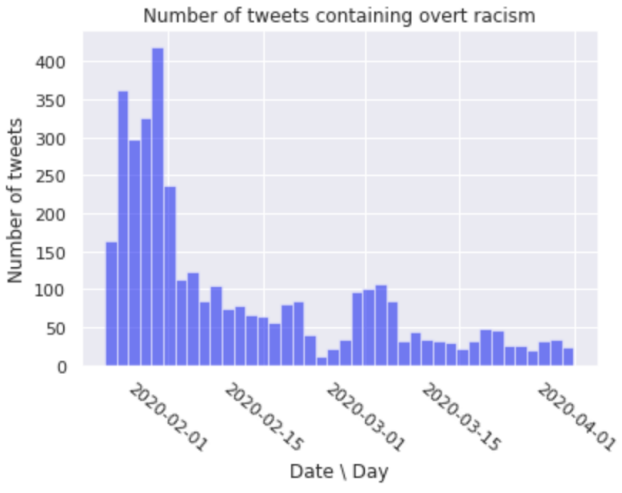

Figure 12: Chart of overt racism towards Chinese

Finally, in *figure 12* we analyzed whether overt racism i.e usage of clear racial slurs such as "chink" changed over time. What we found, which corroborates the findings in the previous graphs, is that usage of these obvious terms spiked during the peak spread of COVID-19 during late January and Early March in the US and China, but quickly faded in popularity.

6 DISCUSSION

Our results complement the findings in the existing literature regarding the relationship between racism and various virus outbreaks, such as SARS and Ebola, which showed that racism towards a particular group (a community from where the virus originated) is not directly caused by the virus itself, but rather is a continued wave of existing bias towards them.

Our primary contribution has been the usage of an ensemble of ML models instead of relying on one specific model. This enables us to increase the accuracy when identifying racist hate-speech by increasing regularization and reducing over-fitting. It also enables us to capture more intricate features of tweets that would otherwise be missed by a single model. This can be seen in our superior accuracy. Nevertheless, this comes at a cost of lower recall levels. However, for this analysis involving huge datasets (60M+ tweets), a higher precision is more desirable.

While we strove to make our analysis and results robust by following rigorous data processing procedures and making use of a wide array of ML models, our work is not exempt from some limitations. First limitation stems from that fact that we trained our models on out-of-distribution data as opposed to an in-domain data. Davidson et al.[16] was originally designed to detect racism towards African Americans (not Chinese as in our context) while the Waseem and Hovy dataset [15] was designed for classification of Islamophobic tweets along with tweets targeting African Americans. While we combined them in order to reduce their idiosyncratic limitations, training on in-domain data would have yielded more accurate results in detecting tweets specifically racist towards Chinese. In our future work we could overcome these limitations by manually annotating our own dataset (e.g. using Amazon Mechanical Turk) to generate training data for detecting tweets racist towards Chinese. Second limitation of our research is that we did not filter out potential tweets by bots. Since we found that a proportion of hateful tweets was fairly constant regardless of the total tweets posted by the users (bots tend to post more often than humans), we believe that this limitation would not alter the general findings of our research.

7 AUTHORS' CONTRIBUTIONS

All authors' contributions can be seen in Appendix 1. These are described in terms of scientific standards and are based on the guidelines given by PLOS Journals.

8 ACKNOWLEDGMENTS

We would like to express our special thanks to our professors, Talal Rahwan and Bedoor AlShebli, for the feedback, guidance and supervision provided throughout the development of our research project. Moreover, we would like to thank our peers from the Computational Social Science class at New York University Abu Dhabi from the Spring 2020 academic semester, for their feedback during the brainstorming of research ideas. As a further matter, we would like to acknowledge Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber for the open-source system and dataset that enabled us to conduct our research. On the same note, we would like to thank Zeerak Waseem and Dirk Hovy for the open-source dataset that we used in training our models. Last but not least, we would like to express our special thanks to Emily Chen, Kristina Lerman and Emilio Ferrara for providing the first public dataset of tweets related to COVID-19 that was of great use in our research.

REFERENCES

[1] S. Harris Ali and Roger Keil. 2006. Global Cities and the Spread of Infectious Disease: The Case of Severe Acute Respiratory Syndrome (SARS) in Toronto, Canada. In *Urban Studies*, Vol. 43. 491–509. https://doi.org/10.1080/00420980500452458

[2] Andrew Atkeson. 2020. What Will Be the Economic Impact of COVID-19 in the US? Rough Estimates of Disease Scenarios *(Working Paper Series)*. https://doi.org/10.3386/w26867

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. COVID-19: The First Public Coronavirus Twitter Dataset. arXiv:cs.SI/2003.07372 https://arxiv.org/abs/2003.07372

[4] Delan Devakumar, Geordan Shannon, Sunil S Bhopal, and Ibrahim Abubakar. 2020. Racism and discrimination in COVID-19 responses. *The Lancet* 395, 10231. http://www.sciencedirect.com/science/article/pii/S0140673620307923

[5] The Humanitarian Data Exchange. 2020. Novel Coronavirus (COVID-19) Cases Data. The Humanitarian Data Exchange. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

[6] Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. In *Proceedings of the Workshop on Noisy User-generated Text*. Beijing, China, 146–153. https://www.aclweb.org/anthology/W15-4322

[7] Shravan Hanasoge, Noriaki Horiuchi, Congcong Huang, Hepeng Jia, Na Young Kim, Mio Murao, Minah Seo, Rebecca Tan, and Jens Wilkinson. 2020. Visibility challenges for Asian scientists. In *Nature Reviews Physics*. https://doi.org/10.1038/s42254-020-0162-z

[8] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *AAAI Publications, Twenty-Seventh AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6419

[9] Judy Yuen man Siu. 2015. Influence of social experiences in shaping perceptions of the Ebola virus among African residents of Hong Kong during the 2014 outbreak: a qualitative study. In *International Journal for Equity in Health*. https://doi.org/10.1186/s12939-015-0223-6

[10] Nature. 2020. Stop the coronavirus stigma now.

[11] World Health Organization. 2020. Coronavirus disease 2019 (COVID-19): situation report, 52. World Health Organization.

[12] Georgios Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. In *Appl Intell 48*. https://doi.org/10.1007/s10489-018-1242-y

[13] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *CoRR*. http://arxiv.org/abs/1809.07572

[14] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in English on Twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*. Association for Computing Machinery, 415–425.

[15] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *SRW@HLT-NAACL*.

[16] Davidson Thomas , D. Warmsley , M. Macy and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *AAAI Publications, Eleventh International AAAI Conference on Web and Social Media*. https://arxiv.org/abs/1703.04009

[17] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving Hate Speech Detection with Deep Learning Ensembles. In *LREC*.

[18] Arkaitz Zubiaga, Alex Voss, Rob Procter, Maria Liakata, Bo Wang, and Adam Tsakalidis. 2016. Towards Real-Time, Country-Level Location Classification of Worldwide Tweets (dataset). https://doi.org/10.6084/m9.figshare.3168529.v2