How much hate with #china? A preliminary research on China-related hateful tweets two years after the Covid pandemic began

Jinghua Xu
University of Tübingen
jinghua.xu@student.uni-tuebingen.de

Abstract

Following the outbreak of a global pandemic, online contents are filled with hate speech. Donald Trump's "Chinese Virus" tweet shifted the blame for the spread of the Covid-19 virus to China and the Chinese people, which triggered a new round of anti-China hate both online and offline. This research intends to examine China-related hate speech on Twitter during the two years following the burst of the pandemic (2020 and 2021). Through Twitter's API, in total 2,172,333 tweets hashtagged #china posted during the time were collected. By employing multiple state-of-the-art pretrained language models for hate speech detection, a wide range of hate of various types is ensured to be detected, yielding an automatically labeled anti-China hate speech dataset. The analysis conducted in this research reveals the number of #china tweets and predicted hateful #china tweets changing over the two years time span, and identifies 2.5% of #china tweets hateful in 2020 and 1.9% in 2021. Both ratios are found to be above the average rate of online hate speech on Twitter at 0.6% estimated in Gao et al. (2017).

1 Introduction

With the popularity of online social media platforms, the masses are given the opportunity to freely express opinions on the Internet. However, this also provides the freedom for people to spread online hate speech. Hate Speech is commonly defined as any communication that belittles a person or a group based on some characteristic such as race, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). As the spread of online hate speech continues to grow, the detection of hate speech on social media has gained increasing significance and visibility.

The goals to study hate speech detection are manifold. In addition to reducing the toxicity of the online environment by censorship, analyzing online hate on specific topics can help reveal public sentiment and opinions on certain events. Furthermore, it can help establish a linkage between social factors in social science studies. For instance, Kim and Kesari (2021) links misinformation regarding China and Covid hate speech using the case of anti-Asian hate speech during the Covid-19 pandemic based on observational data.

Following the outbreak of a global pandemic, the Internet is awash with hate speech. With various restrictions against the virus carried out in countries all over the world, widespread disruption was caused in people's normal lives, which led to rising levels of anxiety, stress, and anger. On March 16, 2020, then US President Donald Trump linked the Covid-19 virus to China and the Chinese people by referring to Covid-19 as "Chinese Virus" in a tweet. The tweet shifted the blame for the global pandemic and redirects the anger to China and the Chinese people. And it set off a new round of "Sinophobia" both on the Internet and in real life.

While previous studies such as He et al. (2021) have examined the Covid-related online hate towards the larger Asian community, this research is particularly interested in anti-China hate on Twitter triggered after the beginning of the pandemic. With Kim and Kesari (2021) establishing the association between Covid misinformation regarding China and online Covid hate speech, this research intends to examine online hate specifically with #china. Through Twitter's API, all tweets (in English) hashtagged #china posted during the years 2020 and 2021 were collected, resulting in 2,172,333 tweets for analysis. In order to ensure wide coverage of various types of hate, this research employs an aggressive approach to predict hatefulness by using multiple pretrained language models for hate speech detection. The state-of-the-art models have proved excellent performance in previous work. The tweets automatically labeled by the pretrained

language models lead to a silver dataset of anti-China hate speech.[1] Based on the predictions, the analysis reveals the percentage of hateful tweets in #china tweets in 2020 (2.5%) and 2021 (1.9%). Both ratios are found to be bigger than the average ratio of online hate speech on Twitter (0.6%) identified in Gao et al. (2017). Furthermore, the analysis divulges the number of #china tweets and hateful #china tweets posted per day changing over the two years time span and presents an overview of the frequently mentioned keywords in the hateful #china tweets in 2020 and 2021.

2 Related Work

Previous studies on online Covid hate speech such as Nghiem and Morstatter (2021), An et al. (2021) and He et al. (2021) mainly focus on hate speech targeting the Asian group. There has been no previous work specifically on anti-China hate speech. Hate speech datasets created over the years mostly have a type of concentration in terms of hate target or topic. For instance, Warner and Hirschberg (2012) labeled hate speech that is anti-Semitic from Yahoo!'s newsgroup post and American Jewish Congress's website; Kwok and Wang (2013) created a balanced dataset of non-hateful and hateful tweets targeting the African community; Burnap et al. (2014) collected hateful tweets related to the murder of Drummer Lee Rigby in 2013; Basile et al. (2019a) proposes a dataset that contains hate speech targeting women or immigrants. In order to create such datasets, the sources of hate speech data are many. These range from user comments on newspaper articles to online social media content from Facebook, Twitter, Reddit and other platforms.

The fact that the majority of hate speech datasets are restricted to a specific type of hate or topic is partially due to the sparsity of online hate speech and the method used to collect initial data for manual annotation. In order to create a hate speech dataset, a number of research start from filtering data through searching by keywords in order to gather texts more likely to be hateful and conduct manual annotation on the selected texts. Since the outbreak of the Covid-19 pandemic, along with Donald Trump's "Chinese Virus" tweet, various Covid misinformation regarding China, and daily trendings such as #fuckchina appearing more frequently on Twitter, a high level of anti-China hate can be sensed to have been developed. The keyword #china has become potentially a good indicator leading to a sufficient number of online hateful speech. This research is thus interested in exploring hate associated with the keyword "China" as a preliminary study in building an anti-China hate speech dataset.

Methods for hate speech detection include conventional rule-based approaches (e.g. keyword-based detection, sourcing metadata) and data-driven approaches (MacAvaney et al., 2019). The statistical methods include supervised and unsupervised learning approaches, with supervised methods more widely applied. Various supervised learning models have proved good performance on the task in previous work. These models include both classic machine learning models and neural network models. For classic models, Support Vector Machines (Noble, 2006), Naive Bayes (Webb et al., 2010), and Logistic Regression (Wright, 1995) have been most commonly used. A previous study Putri et al. (2020) compared some of the classic models for hate speech detection using Indonesian tweet data. Apart from the classic models, neural network models have also been widely used for the task. For instance, Badjatiya et al. (2017) investigated various deep learning models for hate speech detection using the benchmark dataset proposed in Waseem and Hovy (2016). Amongst neural network models, long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) models have been most widely applied and presented excellent performance. A number of systems proposed in previous work are based on or partially based on LSTM (Gao et al., 2017; Bisht et al., 2020; De la Pena Sarracén et al., 2018). In addition to the traditional models, pretrained language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have proved to advance the state of the art on NLP tasks including hate speech detection. These pretrained models have shown superior performance without overfitting, and are currently among the top performers for the task of hate speech detection.

Despite the advanced performance of various supervised learning methods, these approaches require a large amount of annotated data, which are costly to create and often restricted to specific types. Many unsupervised methods for hate speech detection have been developed and used over the years.

[1]The code and data of this paper are released at github.com/JINHXu/how-much-hate-with-china

For instance, Gao et al. (2017) proposed a weakly supervised two-path bootstrapping system, which was designed to capture both implicit and explicit hate speech with the minimum requirement for annotated data. The bootstrapping system contains two learning components: a slur term learner and an LSTM classifier. Due to the low reliability of the slur term learner, the overall system performance is modest. Later work such as "Snorkel" proposed in Ratner et al. (2020) is able to achieve better performance with weak supervision by statistically modeling the process of rule-based labeling and training high-accuracy machine learning models for various NLP tasks including text classification.

3 Data

The data have been collected through Twitter's API. This research collects all English tweets[2] posted with the hashtag #china in 2020 and 2021. In 2020, in total 1,236,335 tweets are collected, and 935,998 tweets in 2021. Table 1 presents an overview of the number of #china tweets in each quarter over the two years.

Year	Q1	Q2	Q3	Q4
2020	336,017	393,513	295,283	211,522
2021	217,831	232,949	243,974	241,244

Table 1: The number of tweets posted with the hashtag #china per quarter in 2020 and 2021.

In addition to tweet content and creation time, the dataset also includes other metadata of each tweet including author id, tweet id, like count, quote count, reply count, retweet count, and source.

4 Method

In order to ensure the reliability of the predictions, this research employs state-of-the-art pretrained language models for hate speech detection. Such models' predictions can be biased based on the data they were trained on. For instance, models trained on hate speech targeting the Asian community may be more sensitive to anti-Asian hate speech and less sensitive to other types of hate speech such as sexist hate language. The hate towards China may be of any type: nationality hate, racism, Covid hate, immigrants hate, etc. Thus, as a means to ensure wide coverage of various types of hate, three pretrained models trained on different types of hate speech data were used to identify hate. The selected models are trained or partially trained on English tweet data. Each tweet identified as hateful by any of the three models is considered hateful in this research.

4.1 COVID-HATE BERT model

The COVID-HATE BERT model is a BERT model trained on the anti-Asian hate speech dataset COVID-HATE (He et al., 2021). The dataset contains 3355 English tweets manually labeled in three categories: hate speech, counterspeech, and neutral. The BERT model trained on this dataset was able to achieve an average macro F1 score of 0.832 and a per-class F1 score on hate speech of 0.762 on the COVID-HATE test data. This research uses the pretrained language model directly to classify the collected tweets and post-process the labels into binary by merging the neutral and counterspeech categories. Given that the model is trained on the COVID-HATE data and proved high F-score, the model is expected to predict Covid-related hate speech with good accuracy.

4.2 HateXplain BERT model

The HateXplain BERT model is a BERT model primarily trained on the HateXplain dataset (Mathew et al., 2020).[3] The dataset consists of 20K posts (in English) from Gab and Twitter. Each data sample is annotated with one of the hate/offensive/normal labels, target communities, and rationales of the label by each annotator. The target group of hate speech in the dataset covers race, gender, religion, sexual orientation, and miscellaneous (e.g. indigenous, refugee/immigrant). The BERT model trained on this dataset was able to reach a macro F1 score of 0.68. Thus the model is anticipated to be able to identify hate speech of a wide range of types with reliable performance.

Before feeding data to the HateXplain BERT model for inference, each tweet was preprocessed by cleaning the URLs, emojis, user tokens following the steps of preprocessing tweets for hate speech detection suggested in Pérez et al. (2021).

4.3 Twitter RoBERTa Hate model

The Twitter RoBERTa Hate model is the best Twitter masked language model retrained in TweetEval (Barbieri et al., 2020), an evaluation framework

[2]Exclude retweets, quotes, and replies.

[3]Additional data from Gab, Twitter, and Human Rationales were included for training to boost model performance.

of Twitter-specific classification tasks, on the hate speech dataset proposed in Basile et al. (2019b). The model performance ranks among the top ones on the leaderboard in TweetEval. The data used for retraining the RoBERTa model is composed of non-hateful and hateful English tweets targeting immigrants and women. The original dataset proposed in SemEval-2019 Task 5 (Basile et al., 2019b) also contains Spanish tweets.

Following the steps suggested in Barbieri et al. (2020), each tweet was preprocessed by replacing URLs and user mentions with placeholders before feeding to the model to predict.

5 Analysis

5.1 Overview

Among the 1,236,335 #china tweets collected in 2020, 2.5% (31,500) are identified as hateful. While the number went down to 935,998 in 2021, with 1.9% (17,872) classified as hateful. Both ratios of hateful tweets in #china tweets in the years 2020 and 2021 are above the average percentage of online hateful language on Twitter estimated in Gao et al. (2017). Overall, the number of both #china tweets and hateful #china tweets declined in 2021 from 2020, with the hateful percentage also decreased.

5.2 Daily Number and Hateful Rate Analysis

Figure 1 shows the number of tweets and hateful tweets posted with the hashtag #china per day in the year 2020. Several visible summits of the number of daily #china tweets can be seen in January, April, May, and June. The peaks appear mostly in the first two quarters of 2020, which was the beginning time of the global pandemic. It is notable that in the last two quarters of 2020, the number of tweets hashtagged #china posted per day went down to fluctuating near a lower level at 2500, with no visible spikes. It is worth mentioning that the most outstanding spikes over the year were not triggered by the "Chinese Virus" tweet, instead, these happened in April and June. Both peak values surpassed 17500 (April) and 10000 (June) per day respectively, with no known events closely related to China that happened during the time.

Figure 2 provides a closer look into the number of hateful #china tweets posted each day in 2020. It is clear from the chart that the "Chinese Virus" tweet triggered a rise in the number of hateful tweets targeting China. However, the bigger

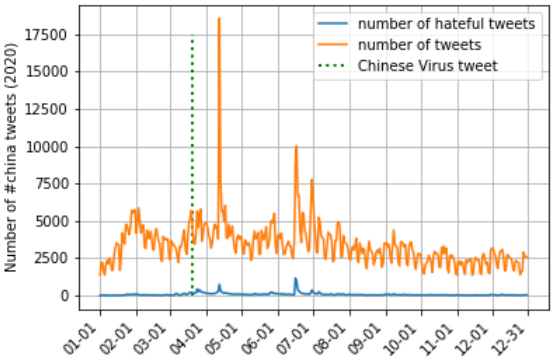

Figure 1: The number of #china tweets and hateful #china tweets per day in the year of 2020.

apex values were reached in April and June. These two peaks are synced with the two most notable peaks of the daily number of #china tweets. Overall, it can be seen from the figure that the major summits and spikes of the daily number of hateful #china tweets were in the first two quarters of 2020. In the last two quarters, the number, in general, maintains in a low level with no remarkable surges.

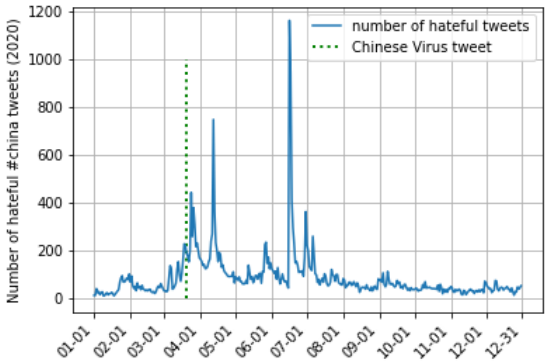

Figure 2: The number of hateful #china tweets posted per day in 2020.

Figure 3 shows the hateful tweet rate in #china tweets on each day over the year 2020. It can be seen from the chart that the hateful tweet rate in #china tweets is above the average rate on Twitter of 0.6% most time of the year. It is notable that the hateful rate had been climbing before the "Chinese Virus" tweet, only the tweet further increased the rate to a higher value at 8%, which was followed by another apex in April. The most outstanding spike appeared in June, the maximum peak value was reached at a percentage as high as over 12%. In general, the major peaks in hateful rates are consistent with these the number of hateful tweets

in 2020. Since the beginning of the pandemic in February, the hateful tweet rate of #china tweets stays at a higher level until the end of the year with several notable surges in the second quarter.

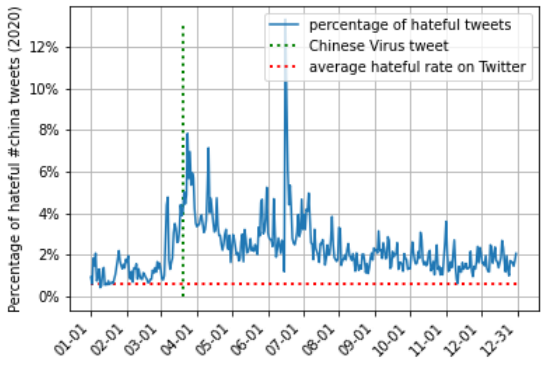

Figure 3: The hateful rate of #china tweets each day in 2020.

Figure 4 shows the number of tweets posted with the hashtag #china on each day in 2021, and the number of these identified as hateful per day. Apart from several apexes in April, May, October, and December, the number of #china tweets posted per day in 2021 generally fluctuates around 2500. The overall level of the number of tweets hashtagged #china posted per day is lower than that in 2020.

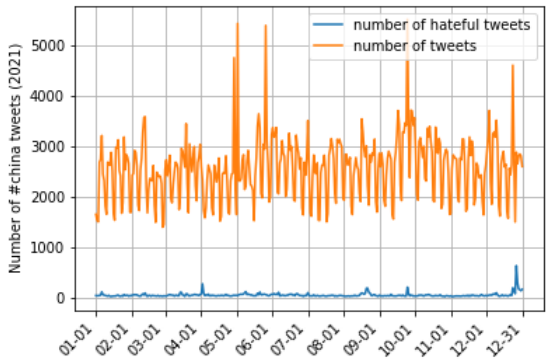

Figure 4: The number of #china tweets and hateful #china tweets per day in the year of 2021.

Figure 5 presents a better view over the number of #china tweets identified hateful each day in 2021. It can be seen from the chart that the number of hateful tweets posted each day rarely surpasses 100. For most time of the year, the number maintains at a low level except for the few peaks in April, August, September, and December, with the surge in December especially outstanding, which led to the maximum apex value of over 600 hateful #china tweets on one day in 2021.

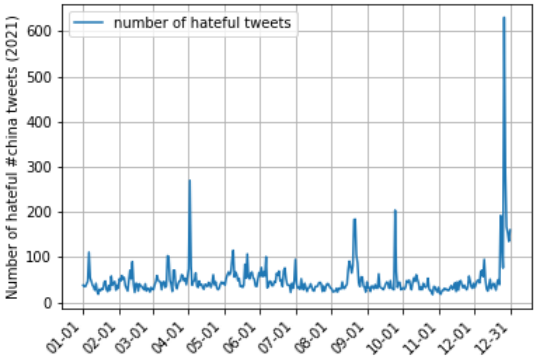

Figure 5: The number of hateful #china tweets posted per day in 2021.

Figure 6 shows the hateful rate in #china tweets posted in 2021 per day. It can be seen that the several peaks in the percentage change are in harmony with the peaks in the number of hateful #china tweets per day over the year. Apart from the smaller peak values below the level of 5% in April, May, June, and October, the percentage reached a more outstanding apex value of over 10% at the end of March, and around 8% in around mid-August. It is also notable that at the end of December 2021, there was an outstanding spike that pushes the daily rate to over 20%, which was the highest hateful rate over the entire year of 2021. Overall, throughout the entire year of 2021, the hateful rate generally maintains at a low level apart from the few apexes, however, the rate is still overall above the average hateful rate of 0.6% on Twitter.

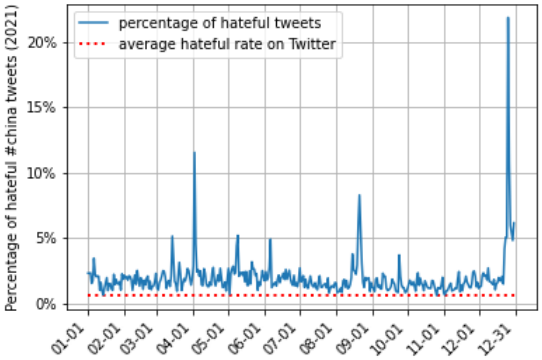

Figure 6: The hateful rate of #china tweets each day in 2021.

5.3 Hateful Keywords Analysis

Figure 7 and Figure 8 show the word clouds generated from the hateful tweets posted with the hashtag #china in 2020 and 2021 respectively. Both figures present a visual overview of the most important and frequently mentioned keywords in the hateful #china tweets each year. With no surprise, the keyword "China" is most frequently mentioned in both years as all tweets were collected according to the keyword.

Figure 7: The word cloud of hateful #china tweets posted in 2020.

It can be seen from Figure 7 that in 2020 some of the Covid-related terms such as "coronavirus", "virus", "chinesevirus", and "covid19" are frequently mentioned in those #china tweets detected as hateful. Additionally, the f-word shows a strong presence in these hateful tweets according to the word cloud. Furthermore, several countries including India, Pakistan, and USA/America have been frequently mentioned in the hateful #china tweets posted in 2020.

When it comes to the word cloud of hateful #china tweets posted in 2021, it can be seen from Figure 8 that the Covid-related terms are no longer frequently involved. Instead, the most relevant and frequently mentioned terms appear to be Africa-related issues and events such as the Tigray Genocide. Based on the key terms such as "Africa", "Ethiopia", "lending", and "dollars", it can be sensed that the hate associated with #china in 2021 may be related to Chinese loans to Africa. And it is worth mentioning that "US" is a commonly frequently mentioned keyword in the hateful #china tweets in both 2020 and 2021.

Figure 8: The word cloud of hateful #china tweets posted in 2021.

6 Limitations & Future Work

This research has several limitations. First of all, the analysis of this research is based on hateful tweets identified through an aggressive approach by using three pretrained language models. Although the state-of-the-art models have been carefully chosen and proved advanced performance on hate speech detection in previous work, this research did not provide evaluation metrics on the method through manually annotating a random data sample to directly illustrate the reliability of the analysis, which is based on the predictions of the proposed method. This research should thus serve as preliminary. Future work should consider conducting manual annotation on a random data sample to evaluate and improve the proposed system for anti-China hate speech detection. Furthermore, by manually labeling the tweets identified as hateful in this study, a large-scale anti-China hate speech dataset can be created on top of this work.

Secondly, the analysis of this study focuses on presenting the statistical results by mainly examining the numerical data (e.g. number of hateful #china tweets posted per day) and their distribution over time. I suggest for future political science study to further investigate the social factors and events behind the abrupt surges and outstanding peak values. For instance, the spike of both the

number of #china tweets and hateful #china tweets in December 2021 remains unexplained, since no significant event known to be related to China has happened during the time.

Thirdly, the scope of the analysis can be further extended in future work. The analysis conducted in this research is limited to tweets, with quotes, replies and retweets excluded. Future work can consider expanding the investigation scope by including these retweets and replies for analysis. Furthermore, future research should also consider analyzing the users, geo-locations, and hate type (e.g. racism) of the hateful #china tweets. Additionally, the analysis scope can be extended both chronologically and geographically, i.e. examining the hate speech in the previous years before the Covid pandemic in order to see if the hate with #china had been phenomenal before the outbreak, and comparing the hateful rate across different countries.

7 Conclusion

This paper presents a preliminary analysis on online hate associated with #china on Twitter through examining hateful speech in tweets posted with the hashtag #china over the two years after the outbreak of the Covid-19 pandemic. Over two million #china tweets posted in 2020 and 2021 were collected in this research. Through an aggressive approach by utilizing three state-of-the-art pretrained language models for hate speech detection, this study identified a wide range of hateful tweets of various hate types in these #china tweets, leading to a large-scale automatically labeled anti-China hate speech dataset. The analysis conducted in this study presents an overview on the number of #china tweets and hateful #china tweets, as well as the hateful rate per day distributed over the two year time span. It was found out that the number of #china tweets and hateful #china tweets went down in 2021 from 2020, so did the hateful tweet rate. Although in both years, the identified hateful rate in #china tweets is above the average rate of 0.6% on Twitter. Through keyword examination, the analysis presents the most frequently mentioned keywords in the identified hateful #china tweets. This analysis finds the most relevant key term mentions in the two years to be vastly different, with Covid-related terms most frequently mentioned in 2020 and Africa-related keywords in 2021.

References

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*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019a. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019b. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Akanksha Bisht, Annapurna Singh, HS Bhadauria, Jitendra Virmani, et al. 2020. Detection of hate speech and offensive language in Twitter data using LSTM model. In *Recent Trends in Image and Signal Processing in Computer Vision*, pages 243–264. Springer.

Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):1–14.

Gretel Liz De la Pena Sarracén, Reynaldo Gil Pons, Carlos Enrique Muniz Cuza, and Paolo Rosso. 2018. Hate speech detection using attention-based LSTM. *EVALITA evaluation of NLP and speech tools for Italian*, 12:235.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jae Yeon Kim and Aniket Kesari. 2021. Misinformation and hate speech: The case of Anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety*, 1(1).

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

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*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

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*.

William S Noble. 2006. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

John T Nockleby. 2000. Why internet voting. *Loy. LAL Rev.*, 34:1023.

TTA Putri, S Sriadhi, RD Sari, R Rahmadani, and HD Hutahaean. 2020. A comparison of classification algorithms for hate speech detection. In *Iop conference series: Materials science and engineering*, volume 830, page 032006. IOP Publishing.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

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*, pages 88–93.

Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. 2010. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714.

Raymond E Wright. 1995. Logistic regression.