Hate Speech Annotation in #china tweets

Jinghua Xu
Heidelberg University
jinghua.xu@stud.uni-heidelberg.de

Abstract

This work follows Xu and Weiss (2022) and collects #china tweets posted in 2022. We first automatically label hate speech using the same automatic annotation pipeline in Xu and Weiss (2022) to filter data, then sample 800 tweets identified as hateful in each year from 2020 to 2022, and conduct manual annotation on the 2400 samples by three trained annotators from different ethic and cultural groups. The annotations achieved inter-annotator agreement slightly above 0.4, resulting in a nearly balanced benchmark dataset of hate speech sourced from Twitter associated with the topic #china. Our strongest baseline model achieves an F-score as high as 0.827 trained and evaluated on our dataset.

1 Introduction

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 (Schmidt and Wiegand, 2017).

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 in 2020, 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. Xu and Weiss (2022) collected all 2M #china tweets[1] posted in the two years following the Covid pandemic, and designed an automatic annotation pipeline to identify hate speech in order to analyze hate speech level across 2020 and 2021. In order to improve the reliability of the automatic annotation pipeline in Xu and Weiss (2022), this work extends the work to 2023 and sample 800 tweets for each year identified as hateful in the automatic annotation pipeline to conduct manual annotation on. We sample 2400 tweets and manually annotate by three annotators from different ethic and cultural groups. Our annotation achieved inter-annotator agreement at slightly above 0.4. We obtain a nearly balanced Twitter hate speech benchmark dataset with the topic #china. Our strongest baseline model achieves an F-score as high as 0.827 trained and evaluated on our dataset.

In the following sections of this report, we first briefly discuss related works. We then describe out data collection, filtering and annotation pipelines, and present annotation results. Finally, we discuss results of baseline models trained and evaluated on our dataset, and conclude our finds.

The code and data of this paper are released at github.com/JINHXu/hate-speech-annotation.

[1] English tweets, with replies, quotes and retweets excluded.

2 Related Work

Most hate speech datasets focus on hate speech targeting a specific group or topic. For instance, Warner and Hirschberg (2012) labeled anti-Semitic hate speech 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. (2019) 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 raw data for manual annotation. In order to create a hate speech dataset, most research starts from filtering data by searching by keywords in order to gather texts more likely to be hateful and conduct manual annotation on the selected texts.

While a benchmark dataset of anti-Asian hate speech has been created in He et al. (2021) following the Covid pandemic, few work has been done speficially on hate against China-realted aspected in general. Shen et al. (2022) studies anti-China sentiment on social media between 2016 and 2021 using data from Reddit and 4chan. So far there has not been many known research towards creating benchmark dataset for hate speech specifically under the topic of #china.

3 Data Collection & Sampling

In addition to the 2,172,333 tweets posted with #china in 2020 and 2021 collected in Xu and Weiss (2022), this work collects 1,123,871 English #china tweets posted in 2022.[2] In order to obtain a sample of sufficient amount of hate speech, this work employs three LLM-based classifiers[3] to first automatically label all 3M tweets, and aggregate label predicted by three models: COVID-HATE BERT model (He et al., 2021), the HateXplain BERT model (Mathew et al., 2020), and the Twitter RoBERTa Hate model (Barbieri et al., 2020).

We sample 800 tweets for each year in 2020, 2021, and 2022 from data labelled as hateful following the above steps, leading to a dataset of 2396 tweets (4 tweets skipped by annotator 2). Figure 1 and table 1 shows number of tokens in the sampled tweets.

	Avg.	Min.	Max.
Num. tokens	43.9	3	96

Table 1: Mean, minimum, and maximum number of tokens in 2396 sampled tweets.

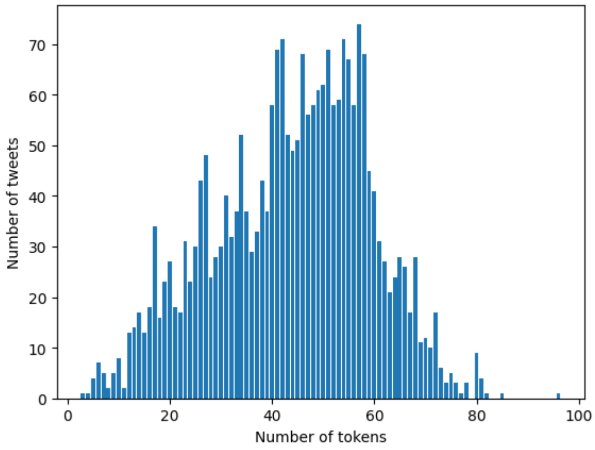

Figure 1: Token count distribution in 2396 sampled tweets.

4 Data Annotation

Three trained annotators from different ethic groups and cultural backgrounds participated in the annotation process. Each annotator were asked to label all 2400 tweet sampled in the previous step. Each annotator were provided with annotation guidelines[4] with definition of hate speech and examples of both cases. We used the annotation tool label-studio[5] to ease the annotation process. The final label is determined by majority vote of annotations from three annotators.

Table 2 shows annotation results by each annotator and the final label by aggregation. It can be seen from the table that annotator 0 labels hate speech most aggressively amongst the three annotators, whereas annotator 2 label more data as non-hateful. Meanwhile, annotator 1 shows a more neutral attitude. By aggregating label, we obtain a nearly balanced dataset.

[2] Retweets, replies and quotes are excluded, to be consistent with Xu and Weiss (2022).

[3] LLM: Large Language Model

[4] https://github.com/JINHXu/hate-speech-annotation/blob/main/annotation_guidelines.md

[5] https://labelstud.io/

Annotator	**NEG**	**POS**
Annotator 0	997	1399
Annotator 1	1158	1238
Annotator 2	1418	978
Aggregated Label	1217	1179

Table 2: Annotation results by each annotator and final label by aggregation.

Figure 2 shows pairwise cohen's kappa score between each two annotators. It can be seen from the figure that annotator 1 and annotator 2 has the highest agreement, whereas annotator 0 in general has low agreement with the other two annotators.

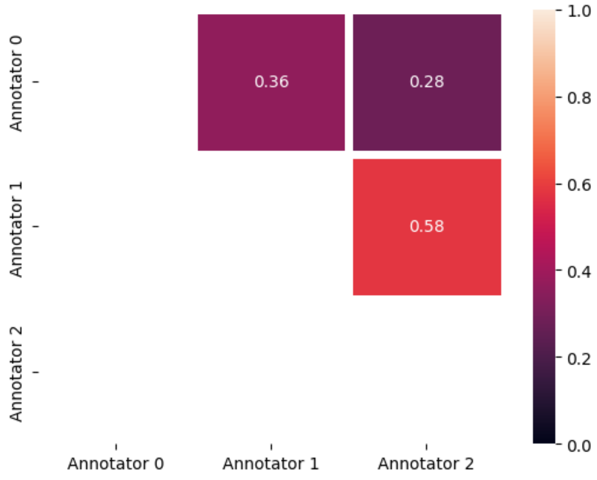

Figure 2: Pairwise inter-annotator agreement.

Table 3 show inter-annotator agreement scores amongst three annotators in different measurements. Our IAA in four meansurements are nearly around 0.40, which is better compared to several existing hate speech benchmark datasets including Del Vigna et al. (2017) and Ousidhoum et al. (2019).

Measurement	**Score**
Cohen's Kappa	0.4082
Fleiss' Kappa	0.4074
Krippendorff's alpha	0.4013
Scott's Pi	0.4012

Table 3: Inter-annotator agreement scores amongst three annotators in different measurements.

Table 4 shows example hateful and non-hateful tweets in our dataset.

Tweet	**Label**
This bullshit has gone on long enough ! That Chinese bitch lives in luxury ! Fuck the extradition, prisoner swap now; close the door on Chinese 5g we don't want it ! **Fuck Off #China !!!** https://t.co/Okg2UCqnP2	**1**
So Sad ! China Evicts Africans from their Homes Claiming they are Importing Coronavirus into China @MaziNnamdiKanu #China where did Corona virus originated? #China show the world proof of your claim. I say #Notoracist #Free the blacks https://t.co/1XNolmzvdb	**0**

Table 4: Example tweets labelled as hateful and non-hateful in our data.

5 Baselines

Following completion of data annotation, we additionally train a few baseline models using the labelled dataset. We split train, validation, and test data by ratio 64: 16: 20. For data preprocessing, we following the practice of twitter data cleaning in Barbieri et al. (2020) by replacing user names and urls in tweets with placeholders. We use the training split of our dataset to fine-tune three pre-trained large language models as our baselines: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Twitter-RoBERTa (Barbieri et al., 2020). We evaluate each of the fine-tuned baseline models on the test split of our data.

- **BERT** stands for Bidirectional Encoder Representations from Transformers pre-trained on data from English language. It is a stack of transformer encoder layers with 12 attention heads, i.e., fully connected neural networks augmented with a self attention mechanism.

- **RoBERTa** is a reimplementation of BERT with some modifications to the key hyperparameters and minor embedding tweaks. It uses a byte-level BPE as a tokenizer (similar to GPT-2) and a different pretraining scheme.

- **Twitter-RoBERTa** is a RoBERTa model trained on approx. 58 M tweets.

Table 5 show the baseline performance on our dataset. It can be seen from the table that the models in general perform well on our dataset, with Twitter-RoBERTa having the best performance achieving highest precision, recall and f-score amongst the three models, while BERT and RoBERTa both show nearly the same performance on our dataset.

Model	F1-Score	Precision	Recall
BERT	0.798	0.748	0.856
RoBERTa	0.798	0.778	0.818
Twitter-RoBERTa	**0.827**	0.796	**0.860**

Table 5: Baseline model performance on our dataset.

6 Conclusion & Future Work

In this annotation project, we collected all over 1M tweets (English) posted with #china in the year of 2022 in addition to data for 2020 and 2021 collected in Xu and Weiss (2022). We filter data using three robust large language models for automatic labelling hate speech, and aggregate labels predicted by three models. We sample 800 tweets from tweets identified as hateful in the automatic filtering pipeline for each year, resulting in a dataset of approx. 2400 #china tweets spanning three years following the outbreak of the Covid pandemic. We conduct manual annotation on sampled data and achieved inter-annoator agreement at slightly above 0.4 across three trained annotators from different ethnic and cultural groups. Through the above process, we obtained a nearly balanced benchmark dataset for hate speech on Twitter associated with the hashtag #china for future analysis. In addition to produsing a benchmark dataset, we train three robust baseline models using the training split of the dataset and test on the validation split. The strongest baseline Twitter-RoBERTa achieved an F-score as high as 0.827.

For future work, we intend to use the benchmark dataset to further fine-tune stronger models to create a reliable automatic pipeline for labelling all 3M #china tweets posted in 2020, 2021, and 2022, and conduct further socio-political analysis based on automatically labelled data.

Acknowledgements

I thank Janosch Gehring and Natalia Minakova for supporting the annotation project by each contributing 2400 annotations.

References

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, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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.

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.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

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*.

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.

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*.

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*.

John T Nockleby. 2000. Why Internet Voting. *Loy. LAL Rev.*, 34:1023.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International*

Workshop on Natural Language Processing for Social Media, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Xinyue Shen, Xinlei He, Michael Backes, Jeremy Blackburn, Savvas Zannettou, and Yang Zhang. 2022. On xing tian and the perseverance of anti-china sentiment online. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 944–955.

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.

Jinghua Xu and Zarah Weiss. 2022. How much hate with china? a preliminary analysis on china-related hateful tweets two years after the covid pandemic began.

A Annotator ID

Annotator 0: Jinghua Xu
Annotator 1: Janosch Gehring
Annotator 2: Natalia Minakova