COVID-19 Anti-Asian Racism Hate Speech Detection using Transfer Learning

Nihal D'Souza, Saffrin Granby, Oksana Necio

Department of Linguistics, University of British Columbia

nihal01@student.ubc.ca

sgranby@student.ubc.ca

necioo@student.ubc.ca

***Abstract*— Since the start of the COVID-19 pandemic, there has been a widespread increase in the amount of hate-speech being propagated online against the Asian community. This paper builds upon and explores the work of He et al. Their COVID-HATE dataset contains 206 million tweets focused around anti-Asian hate speech. Using tweet data from before the COVID-19 pandemic, as well as the COVID-HATE dataset from He et al, we performed transfer learning. We tested several different models, including BERT, RoBERTa, LSTM, and BERT-CNN. Some of these models hindered the performance of He et al's model, while others improved it. We also explored the geographical extent of the ani-Asain tweets from the COVID-HATE dataset.**

***Keywords*— Hate Speech, Counter Hate, Twitter, COVID-19, BERT, RoBERTa, LSTM**

I. Introduction

Following the outbreak of COVID-19 and the identification of its origins in China, there has been a surge in anti-Asian hate crimes ranging from microaggressions to physical and verbal assault. This type of toxic language can trigger harmful real-world events and have long-term consequences. He et al [2] address this problem by creating an anti-Asian hate and counterspeech dataset containing over 206 million tweets. This group was able to train classifiers with BERT embeddings to identify between hate speech, counter speech, and neutral speech with an F1 score of 0.832. Our goal was to expand on the work by He et al [2] by using transfer learning. This project seeks to understand if the general form of hate speech has patterns that could be used to identify COVID-19 hate speech. We asked ourselves if there was something unique about the COVID-19 anti-Asian hate speech such that transfer learning would hinder the performance of He et al's model [2]. We found two datasets containing hate speech tweets from before the COVID-19 pandemic. We trained several models on these two datasets before fine-tuning on the dataset created by He et al. We also analyzed the results of our models and visualized where these tweets were coming from geographically in order to assess whether there are any areas that have disproportionately high anti-Asian tweets.

II. Datasets

For this project, we used three different pre-existing datasets containing tweets. The first two datasets were the Hate Speech and Offensive Language dataset from Davidson et al. [3] and the Hate and Abusive Speech on Twitter dataset from Founta et al. [4] Both of these datasets contain tweets that were collected before the COVID-19 pandemic. They were used for the initial training. The third dataset was the COVID-HATE dataset from He et al. This was used to fine-tune our models.

A. Hate Speech and Offensive Language

The Hate Speech and Offensive Language dataset contains 24,802 annotated tweets collected in 2017. These tweets were annotated as either hate, offensive, or neither. Each tweet was annotated by 3 annotators on average. The label distribution of these tweets is shown in Table I.

This dataset has a class imbalance towards tweets labeled as offensive. We had to change the labels of this dataset in order to match with the others. We combined the offensive label with the hate speech label.

TABLE I
HATE SPEECH AND OFFENSIVE LANGUAGE LABEL DISTRIBUTION

Label	Label Count
Offensive	19190
Hate speech	1430
Neither	4163

B. Hate and Abusive Speech on Twitter

The Hate and Abusive Speech on Twitter dataset contains 100,000 annotated tweets collected in 2018. These tweets were annotated as either normal, abusive, spam, or hateful. Each tweet was annotated by 3 annotators on average. The label distribution of these tweets is shown in Table II.

This dataset has a class imbalance towards the normal label. We removed all data labeled as spam since that label was unimportant to our model. We also combined the abusive and hateful labels.

TABLE II
HATE AND ABUSIVE SPEECH ON TWITTER LABEL DISTRIBUTION

Label	Label Count
Normal	53851
Hateful	4965
Abusive	27150
Spam	14030

C. COVID-HATE

The last dataset we are using is the COVID-HATE dataset. This dataset has over 206 million tweets taken between January 15, 2020, and March 26, 2021. The majority of these tweets were annotated using BERT. 3355 of these tweets were hand-annotated as hate speech, counterspeech, or neutral. We used the hand-annotated tweets as our test set. There are 206,348,565 tweets labeled by BERT. The full dataset with tweets labeled by BERT does not contain the tweets, only the tweet IDs. We collected 26,759 tweets from this dataset using the provided tweet IDs. This is the subset we used to finetune our model. The label counts for the hand-annotated labels, BERT labels, and our subset of the BERT labels are shown in Table III.

To address the class imbalance, we made sure to include more data with the counter hate and hate speech labels. In the original BERT-labeled data, 98.7% of the tweets were labeled as neutral, while in our subset of the tweets, 79.1% of the tweets are labeled as neutral.

TABLE III
COVID-HATE LABEL DISTRIBUTION

Label	Hand Annotated Label Count	BERT Annotated Label Count	BERT Subset
Neutral	1344	203,857,160	21,155
Hate Speech	429	1,337,116	3205
Counter Hate	517	1,154,289	2399

III. METHOD

The models were developed on local machines and mostly trained on GPUs provided by Google Colaboratory. To model this multi-class classification task, we used PyTorch and Hugging Face frameworks. For the model architecture, we used a single-layer BERT encoder transformer and ‘bert-base-uncased’ for the tokenizer. For the model optimizer, we chose the AdamW optimizer and Cross-Entropy Loss as the loss function. Since labels were unbalanced, we chose precision, recall, and Micro F1 score as our primary metrics.

IV. EXPERIMENTS

We implemented the code provided by He et al[2] as our baseline for this project. To improve upon this baseline, we ran 3 experiments. One model each trained on each of the datasets and one model was trained on both datasets. Models trained on

each dataset independently were fine-tuned on the COVID-specific hate speech dataset with all three labels (hate, anti-hate, neutral). The model trained on both datasets was fine-tuned on the same dataset but with only two labels (hate, neutral). For this, the 'anti-hate' label was reassigned to the 'neutral' label. The reason for this was to improve performance in hate speech detection, at the cost of not being able to detect 'anti-hate'.

A. Model trained with Hate and Abusive Speech on Twitter (HAST) dataset

The first model we tried was with the Hate and Abusive Speech on Twitter dataset. To prepare this dataset for our fine-tuning task, we removed all tweets with the spam label because this was not relevant. In this dataset, they classify the rest of the tweets as either hateful, abusive or neutral. For the purposes of our later task, we have combined all tweets labeled as hateful or abusive into the same category.

B. Model training with Hate Speech and Offensive Language (HSOL) dataset

The second model was trained with the Hate and Offensive Language dataset. This dataset had labels for hate, offensive, and neither hate nor offensive.

C. Model trained with both Hate and Abusive Speech on Twitter dataset and Hate Speech and Abusive Language Dataset

The third model was trained with both the Hate and Abusive Speech on Twitter dataset and Hate Speech and Abusive Language Dataset sequentially.

All three models were then independently fine-tuned with the same parameters and training data on the COVID-specific hate speech. Experiment 1 and Experiment 2 were fine-tuned on all three labels, Experiment 3 was fine-tuned on only two labels, as mentioned earlier.

To further explore the effects of the model architecture on performance, we performed the two more experiments.

D. LSTM Architecture

We embedded the sentences to a 300-dimension space, with a total vocabulary size of 5002. This was then fed into a two-layer LSTM model, with a hidden layer size of 500, with a learning rate of 0.1. The loss was calculated using NLL Loss.

E. RoBERTa Architecture

We also implemented an optimized version of the BERT model called RoBERTa. The rest of the parameters were kept constant from the original experiments using BERT.

V. Results

The results for the first three experiments are documented in Table IV.

TABLE IV
Results from Test Dataset

Experiment	Model Pre-train Dataset	F1 Score (Micro)
A	Hate Speech and Offensive Language (HSOL) dataset, fine-tuned on 3 labels	0.62
B	Hate and Abusive Speech on Twitter (HAST) dataset, fine-tuned on 3 labels	0.74
C	HSOL + HAST, fine-tuned on 2 labels, 10% data	0.92 (0.95 in detecting hate)
C	HSOL + HAST, fine-tuned on 2 labels, 5% data	0.88 (0.93 in detecting hate)

On the HSOL dataset, we got a score of 0.62 which is quite low. On the HAST dataset, we get a score of 0.74, which is an improvement over the HSOL dataset, but this is most probably because the HAST dataset is almost four times larger than the HSOL dataset and therefore simply has more data. On the final model that was trained on both the HSOL and HAST datasets, we were not expecting to get a score much better than the first two experiments. Therefore, we decided to merge the

'anti-hate' label into the 'neutral' label resulting in two labels (hate and neutral) for fine-tuning.

This indeed did improve our scores significantly. More so, on fine-tuning with just 10% of the COVID-specific hate speech data, we got an F1 score of 0.92, with 0.95 in detecting hate speech. Out of curiosity, we fine-tuned the model on just 5% of the COVID-specific hate speech data and got an F1 score of 0.88, with 0.93 in detecting hate.

For the next two experiments we performed to explore the effects of the model architecture on performance, the results are documented in table V.

TABLE V
RESULTS FROM MODEL ARCHITECTURE

Experiment	Model Architecture	Micro F1 Score
D	LSTM	0.51
E	RoBERTa	0.82

With the LSTM architecture, we got a score of 0.51 which was worse than our original scores from the BERT model. With the RoBERTa architecture, we got a score of 0.82 which is also lower than our original BERT architecture.

VI. RELATED WORK

In order to prepare for this project, we looked at several other related works. The first area we focused on was previous work identifying COVID-19 specific hate speech. He et al [2] create a dataset specifically for the classification of hate and counterspeech for anti-Asian hate and use BERT embeddings to create a classification model. From their work, we were able to get a silver dataset created from their and also use their hand-annotated data as our test set. Vishwamitra et al [5] also created a COVID-19 dataset that consists of tweets targeting older people and supplemented with a dataset targeting the Asian community. Their goal was to discover specific keywords used in hate speech against these groups. Supplementing the original dataset with the dataset targeting the Asian community allowed the model to produce more confident keywords. From this work, we can see that supplementing a dataset with another in the domain of COVID-19 hate speech can be helpful. In order to prepare for this project, we wanted to look at different works using transfer learning. Benítez-Andrades et al [6] used BERT embeddings along with attention mechanisms. This group first trained a model on an English dataset containing more general hate speech and then used transfer learning with a second Spanish dataset to achieve better classification results. We can see here that using transfer learning in the domain of hate speech can ameliorate results. We even looked at transfer learning concerning sentiment tasks. A study by Boy et al [7] uses transfer learning for sentiment tasks by taking models with parameters trained for emoji-based tasks. We also looked into the specifics of the architecture within transfer learning. Wei et al [8] used a pre-trained BERT model and used transfer learning by freezing all the layers of the model, attached a few neural layers of their own, and trained that part. This array of previous work makes us confident that we will be able to achieve good results by transferring what the model learned in the generic hate speech dataset to the COVID-19-specific one.

VII. DISCUSSIONS, CHALLENGES, AND LIMITATIONS

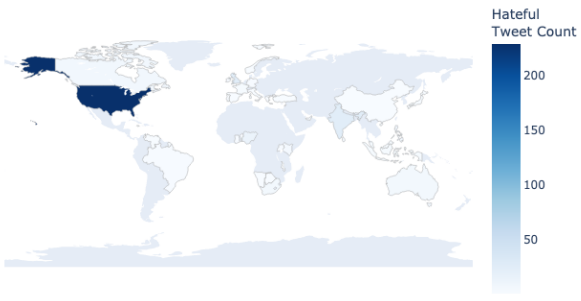
As seen in the results, the BERT model trained on both generic datasets and a subset of the COVID-19 dataset using hate speech and neutral labels performed the best. When training the models on the 3 labels with each dataset separately, we did not achieve satisfactory results. We suggest that the generic data helps the model learn COVID-19 hate speech more effectively but may not help the model learn counter speech. We also find it interesting that training on all the COVID-19 data hurts the performance of the test set. If given more time, we would've liked to explore model interpretability and discover exactly why this phenomenon occurs.

When exploring other models, we found that BERT still outperformed LSTM and RoBERTa. We were not surprised by these results as this was in line with the results of the previous works that we looked up. We were unable to try more models at the time but we give future consideration to variants of BERT such as BERTweet and BERT with CNN and LSTM architecture.

We want to discuss how transfer learning could potentially hinder the performance of our model as it is important to acknowledge the potential downfalls of our approach. Transfer learning will not be effective if there is a mismatch in the domain. If our model has never seen counterspeech before the fine-tuning, we would not be helping the performance of our model in identifying it. In our task, we must also be cautious not to assume that the hate speech seen in the generic dataset will be the same as the ones we saw in the COVID-19 dataset. This could result in the negative transfer where the patterns learned in our generic dataset cause our model to incorrectly label the COVID-19 text.

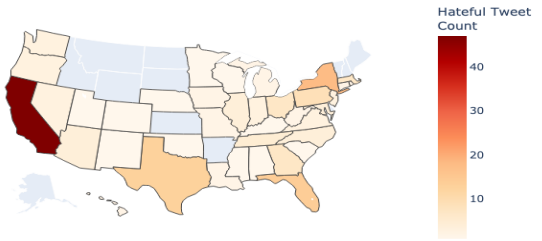
We were also curious to know the geographical extent of hate speech. To do this, we first queried a subset of tweets using Twitter's API to collect the location data, if available. Surprisingly, a majority of the tweets had their location disabled or not included by the user. We decided to only collect locations for tweets marked as hateful and were able to create a choropleth as shown in Figure I to map the global extent.

FIGURE I
GLOBAL EXTENT OF HATE SPEECH



We noticed a majority of the tweets were located in the United States of America. This could be attributed to the fact that we are only working with English language tweets and there is a high possibility that a majority of the English language tweets originate from the USA. Furthermore, we decided to plot the distribution of hate speech in the USA as well. This is shown in Figure II.

FIGURE II
EXTENT OF HATE SPEECH IN USA



We observe a large proportion of hate speech comes from the state of California. This could potentially be a result of a large number of Twitter users originating from that particular state.

Finally, during this project, we faced several challenges that we had to overcome. First, we had to figure out why our models were not performing the way we expected. When we tried combining counter-speech and neutral speech, we saw an amelioration in our scores. By addressing this issue in this project, we also open a discussion of how we could help identify counter-speech. If given more time, we would have liked to research previous work in counter speech and try to address this deficit in our model. We also tried different subsets of the COVID-19 dataset in order to see the effect and found that using a subset of 8% gave us the best scores. Another challenge we had was data collection. Since we had to source some of the data and metadata from Twitter, we ran into some barriers. The Twitter API only allows 900 API calls every 15 minutes which was not very efficient. Additionally, some of the tweets have been deleted since the time of data collection which was

problematic because we would still have to make an API call to check to see if they had been deleted.

VIII. Conclusions

This project aimed to improve upon the BERT model implemented by He et al by using transfer learning. We explored models such as BERT, LSTM, and RoBERTa and found that the BERT model performed the best. We created three different models using two datasets. Our final model was trained on both datasets and fine-tuned with only the hate and neutral labels from the COVID-HATE dataset. This model achieves a micro F1-score of 0.92 when fine-tuned on 10% of the data. The model from He et al. achieved an F1-score of 0.83. It is important to note that their model predicted three labels, while ours only predicted two. Finally, we looked at the geographic extent of the hate tweets in our dataset and found many of the tweets were from the United States, specifically from California.

References

[1] N. Montemurro. (2020) "The Emotional Impact of Covid-19: From Medical Staff To Common People", *Brain, behavior, and immunity*

[2] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, S. Kumar. (2021, November). "Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis". *Proceedings 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* pp. 90-94

[3] T. Davidson, D. Bhattacharya, I. Weber. (2019). "Racial Bias in Hate Speech and Abusive Language Detection Datasets". *Proceedings Third Workshop on Abusive Language Online.* pp. 25–35

[4] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis. (2018). "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior". *Proceedings International AAAI Conference on Web and Social Media*. pp. 491-500

[5] N. Vishwamitra, R. Hu, F. Luo, L. Cheng, M. Costello, Y. Yang. (2020, December). "On Analyzing Covid-19-Related Hate Speech Using BERT Attention". *In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* pp. 669-676.

[6] J. A. Benítez-Andrades, Á. González-Jiménez, Á. López-Brea, J. Aveleira-Mata, J. M. Alija-Pérez, M. T. García-Ordás. (2022). "Detecting Racism and Xenophobia using Deep Learning Models on Twitter data: CNN, LSTM and BERT." *PeerJ Computer Science, 8, e906*

[7] S. Boy, D. Ruiter, D. Klakow. (2021). "Emoji-based Transfer Learning for Sentiment Tasks." *Proceedings 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop,* pp. 103–110

[8] B. Wei, J. Li, A. Gupta, H. Umair, A. Vovor, N. Durzynski. (2021) "Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning."