Detecting East Asian Prejudice

Catherine Jang
cjang@g.hmc.edu

Yuki Wang
yukwang@g.hmc.edu

Abstract

We investigate the problem proposed by Vidgen et al. (2020) to create classifiers that detect prejudice against East Asians in tweets. We specifically compare four RoBERTa-based classifiers that differ in the number of epochs trained and whether or not they were trained with original or augmented data. We find that although longer epochs and having data augmentation do not significantly improve the model, they do help correctly identify tweets from categories that do not occur frequently in the data.

1 Introduction

As concerns of COVID-19 have risen in the last year, unfortunately, so has anti-Asian sentiment. Prejudice against Asians has been spreading on social media platforms like Twitter and creating a toxic online community. A robust classifier that identifies hurtful posts can help human moderators prevent the emergence and spread of such prejudice and aid studies to analyze the effect of language on anti-Asian sentiment.

Recently, notable work has been done to study hate speech and anti-Asian sentiment that occur online. Ziems et al. (2020) study the spread of hate speech in web communities and how bots help instigate the use of more hateful language. Tahmasbi et al. (2020) analyze how the use of Sinophobic language online has changed since the spread of COVID-19. Davidson et al. (2019) present how offensive speech classifiers may propagate racial bias. Specifically, classifiers tend to more often identify posts written by African-American people as offensive speech. In another paper, Davidson et al. (2017) emphasize the importance of identifying the difference between hate speech and offensive speech. They find that classifiers struggle with differentiating hate speech that targets a specific community from general offensive speech.

In hopes to contribute to this field of work, we expand upon the study by Vidgen et al. (2020). We use their dataset of 20,000 cleaned tweets and build RoBERTa-based neural network models to classify a given tweet into one of the following classes: "Hostility against an East Asian Entity" ("Hostility"), "Criticism of an East Asian Entity" ("Criticism"), "Discussion of East Asian Prejudice" ("Discussion"), "Counter Speech," and "None of the Above." Specifically, the tweets in the "Counter Speech" class denounce prejudice or hostility towards East Asian entities, while the tweets in the "Discussion" class mention those prejudices without taking a specific stance (i.e. acting neither hostile towards nor supportive of East Asians).

We investigate how using data augmentation to increase the amount of training data can affect the performance of the model. Specifically, our data augmentation technique adds "attacks" to the training data that could potentially make a tweet harder to classify. We also vary the total number of epochs trained to study the effect of training time. Our results show that training with data augmentation and training for more epochs do not significantly improve the classifier's overall performance. However, these techniques do improve classifications of categories that originally have a sparse amount of training data. Specifically, the model with data augmentation and longer training time improves the F1 score from 0.0 to 0.31 for the "Counter Speech" class, which only makes up 0.6% of the overall dataset.

2 Related Work

2.1 Dataset

Vidgen et al. (2020) collect 20,000 tweets that use hashtags related to COVID-19 and East Asia. Before asking human annotators to label these tweets, they pre-process the data by removing usernames and replacing specific hashtags with more general ones. The five general hashtag categories are #EASTASIA (only referring to an East Asian entity: #China), #VIRUS (only referring to Corona Virus: #covid19), #EASTASIAVIRUS (referring to both an East Asian entity and the virus: #wuhanflu), #OTHERCOUNTRYVIRUS (referring to non-East Asian countries and the virus: #italycovid), and #HASHTAG. The annotators classify 67.6% of the 20,000 tweets as "None of the Above," 19.5% as "Hostility," 7.2% as "Criticism," 5.1% as "Discussion," and 0.6% as "Counter Speech."

2.2 Data Augmentation

Inspired by data augmentation techniques in computer vision, Wei and Zou (2019) explore simple methods to augment training data for text classification tasks. Even with simple modifications such as replacing a word with synonyms, randomly inserting or deleting words, and randomly swapping words, their result shows that the technique helps boost the classification model's performance and reduce over-fitting.

2.3 Classifier Models

Research on detecting hate speech often formulate a classification problem that assigns a particular category to a given text. BERT, which is a language representation model, has been a popular and effective option for text classification tasks. Isaksen and Gambäck (2020) apply transfer learning by repurposing an already trained BERT language model to classify tweets into one of three categories: "Hate Speech," "Offensive Language," and "Neither." Because BERT only provides the pre-trained language representation, the authors extend the model by passing the output of BERT to fully connected layers, which act as the actual classifier. Koufakou et al. (2020) also use BERT as a part of their model, but they focus on including additional lexical information by utilizing HurtLex, a lexicon of offensive words. The final classification layer receives both the BERT encoding and the frequency of words that belong to a HurtLex offensive category. Their result shows that the additional lexical information improves the BERT model's performance.

3 Methods

We focus on using 20,000 annotated tweets provided by Vidgen et al. (2020) and improving their best performing model (RoBERTa). Specifically, we want to apply data augmentation, inspired by Gröndahl et al.'s research, to the original training dataset then explore whether this technique would produce a more robust classifier.

3.1 Model Architecture

Vidgen et al. (2020) do not include detailed information on the implementation of their best performing RoBERTa model. Because RoBERTa and BERT are similar in architecture, we refer to Isaksen and Gambäck's BERT transfer learning architecture. Specifically, we follow Roberti's tutorial to implement our classifier by using Python's Fastai library and HuggingFace library, which offers the pre-trained RoBERTa model.

RoBERTa is an improved BERT language model that modifies the pre-training procedure. RoBERTa is trained on five English-language corpora with over 160GB of uncompressed text. By training the model for a longer time, with larger datasets, over bigger batches, and so on, Liu et al.'s RoBERTa model achieves state-of-the-art results in several classical tasks.

Originally, the RoBERTa language model is trained to predict the masked token in a given sentence. Thus, we repurpose the RoBERTa model to perform multi-class classification for our dataset. Following the neural network structure by Isaksen and Gambäck (2020), we pass the final RoBERTa encoder's output into a fully connected network that acts as our classifier.

3.2 Data Augmentation

Our data augmentation method is based on Gröndahl et al.'s research. Their study shows that many hate speech classifiers' performance gets hindered by "attacks," which are simple text modifications to introduce noise. Their result shows that adding typos and appending non-offensive words decrease the accuracy for models, while inserting or deleting whitespaces does not affect the performance as much. Furthermore, Grondahl et al. suggest "adversarial training," which trains the model with input that contains the aforementioned "attacks" to make the classifier more resilient to noise.

We apply this "adversarial training" technique by augmenting the original training data provided by Vidgen et al. (2020). Specifically, we first randomly split the dataset of 20,000 cleaned tweets into 80% for training, 10% for validation, and 10% for testing. Then, for each training instance, we generate four more slightly modified tweets that retain the original label by doing the following:

- We add typos by randomly changing 5% to 25% of the characters.
- We first randomly remove at most 25% of the whitespaces. Then, we randomly insert at most 5% of the tweet's length number of whitespaces.
- We insert at most 5% of the tweet's length number of words selected from the top 100 most common words. We create the list from the 10,000 most common English words provided by Kaufman et al. (2019).
- We apply all three aforementioned modifications to the tweet.

Thus, the model trained with data augmentation has 80,000 training instances, while the model trained without data augmentation has 16,000 training instances. All models have the same validation dataset and testing dataset (2000 tweets each).

3.3 Training

We train 4 different models: (M1) without data augmentation for 5 epochs (M2) without data augmentation for 20 epochs (M3) with data augmentation for 5 epochs (M4) with data augmentation for 20 epochs. For M1 and M3, we use the same learning rate and training procedure specified in the tutorial by Roberti (2020). Then, considering that data augmentation has increased the training data by 5 times, we train M4 as long as possible while monitoring the validation loss. We stop after 20 epochs because the validation loss starts significantly increasing, which indicates overfitting. Finally, we train M2 with the same procedure for 20 epochs to better evaluate the effect of data augmentation on performance. During training, we optimize the model using the Adam optimizer and use cross-entropy loss because this is a classification task. (Isaksen and Gambäck, 2020)

3.4 Evaluation

We evaluate our classifiers using four evaluation metrics: recall, precision, F1, and accuracy. Since we do not have enough information to reproduce the same RoBERTa model as Vidgen et al. (2020), we cannot make a definite conclusion by comparing our models' evaluation metrics with their models'. Nevertheless, comparing M3 and M4 with M1 and M2 can indicate whether data augmentation has improved or worsened the performance of the classifiers.

4 Results

Model	Precision	Recall	F1	Accuracy
M1	0.82	0.83	0.82	0.83
M2	0.82	0.83	0.83	0.83
M3	0.82	0.83	0.82	0.83
M4	0.82	0.83	0.83	0.83

Table 1: M1: original data, trained for 5 epochs, M2: original data, trained for 20 epochs, M3: augmented data, trained for 5 epochs, M4: augmented data, trained for 20 epochs

We present a summary of our results for the four models in Table 1. Since our objective is to consider the effect of adding "attacks" into training data for classifiers, we can treat M1 as a baseline. Overall, there is very little to no difference in the metric scores between all four models.

Considering the effects of training time, we find that training for more epochs results in a very slightly higher F1 score. However, the 0.01 difference in scores is too small to be conclusively significant. The main difference between models trained for 5 epochs and models trained for 20 epochs occurs in the "Counter Speech" classification. M1 and M3 never predicted "Counter Speech," neither correctly nor incorrectly. M2 predicted only one tweet to be "Counter Speech," but the tweet's actual label was "Hostility." M4 was the only model to correctly predict two tweets to be "Counter Speech." Considering that M2 and M4 were both trained for 20 epochs and were the only models to ever predict "Counter Speech," the results suggest that a longer training period slightly improves the classifier's performance.

We also analyze the difference between training on the original data and on the augmented data. Again, the overall scores do not show any significant difference. However, there are some noticeable differences in category-specific scores between M1 and M3 as well as between M2 and M4. We focus on the difference between M2 and M4, which is presented in Table 2. By looking at this data, we can identify some trade-offs in precision and recall between the two models as well as a higher F1 score for M4 than M2 for "Criticism."

Label	Precision		Recall		F1	
	M2	M4	M2	M4	M2	M4
None of the Above	0.90	0.90	0.93	0.93	0.92	0.92
Hostility against an East Asian Entity	0.69	0.70	0.74	0.73	0.71	0.71
Criticism of an East Asian Entity	0.63	0.56	0.31	0.40	0.41	0.47
Discussion of East Asian Prejudice	0.62	0.70	0.70	0.61	0.65	0.65
Counter Speech	0.00	0.33	0.00	0.29	0.00	0.31

Table 2: Breakdown comparison of the two models trained for 20 epochs (one trained on original data, one trained on augmented data)

For all categories, M4 had a significantly higher set of precision, recall, and F1 scores than M2 for "Counter Speech." In reality, M4 only classified two out of seven "Counter Speech" tweets in the testing dataset correctly. However, considering that M4 is the only model to ever correctly predict "Counter Speech," we conclude that data augmentation also has some positive effect on classifier's performance. Considering that only 0.6% of the 20,000 tweets are labeled as "Counter Speech," we also believe that this technique especially helps improve the model's performance in categories that originally lack training data.

5 Error Analysis

5.1 Effects of Replaced Hashtags

A data preprocessing step defined by Vidgen et al. (2020) is replacing specific hashtags with general thematic hashtags. This step introduces some difficulties for our classifiers. First, for some tweets, hashtags are responsible for defining the overall topic, the targeted entities, and the sentiment of the tweet, while the actual texts do not clearly indicate the writer's position. For example, this tweet, which is labeled as "Hostility" by the annotators, includes the actual text "this shown how useless carrie and our government are ......" and hostile hashtags such as #BoycottChina and #boycottCCP. However, because those offensive hashtags all get replaced by the uniform #HASHTAG, all four models could not identify the main subject nor the opinion of this tweet from just the text, thereby all confidently misclassifying this text as "None of the Above." Even a human annotator would be unlikely to classify this as "Hostility" given the same generalized input that the classifier receives. Thus, these tweets show that hashtags play an important role in reinforcing hostility or criticism towards East Asians.

Second, all models sometimes have trouble differentiating a "Discussion of East Asian Prejudice" tweet from "None of the Above" tweets. This type of mistake is surprising because determining whether a tweet discusses topics related to East Asians should be easier than determining whether a tweet contains hostility or criticism. Even general hashtags such as #EASTASIA and #EASTASIAVIRUS should help the models identify the topic of the tweets. After analyzing the tweets labeled as "None of the Above" in the training data, we notice that about 79% of them contain #EASTASIAVIRUS, and about 50% of them contain #VIRUS. These tweets could have tagged #wuhanflu, which tends to get used prior to identifying the virus, or #covid19 (under #VIRUS) to advocate for safety protocols without targeting any East Asian entities. Thus, it is reasonable that the models could not rely on the hashtags in their classification.

5.2 Importance of Social Context

We find that the models all generally struggle with classifying texts that require some level of social context. More specifically, the models have a difficult time when a specific entity is not referred to within the text or when a sarcastic tone is used.

For instance, one tweet from the test data includes: "what a marketing level!! first spread a virus in the world and then sell masks to that world." This was classified as "Hostility" by the human annotators. However, the classifier predicted that it was "None of the Above." Here, humans can infer that this statement is sarcastic and conveying negative sentiment toward the Chinese government. However, the classifier does not understand that this text is talking about the COVID-19 pandemic, nor does it have any way of knowing that the tweet is directed toward the Chinese government. Therefore, with no clear reference to the Chinese government within the text, the classifier is unlikely to correctly decipher if this text is at all related to an East Asian entity. In short, classifiers sometimes cannot differentiate between "None of the Above" and other labels due to sarcasm or indirect references within some tweets. By introducing more instances of sarcasm and indirect reference to an East Asian entity to the training data, the classifier may be trained to sometimes predict hostility or criticism against an East Asian entity when the text seems like it may be "None of the Above" (i.e. when the text has a sarcastic tone or when it is hostile toward or criticizing East Asian identities without explicitly mentioning terms like "China").

5.3 Difficulty in Differentiating between Hostility and Criticism

Vidgen et al. (2020) point out that "misclassifying Hostility as Criticism (and vice versa) was the largest source of error." Our models run into the same issue. While the model that gets trained with data augmentation for 20 epochs (M4) has the highest F1 score (0.47) for the "Criticism" category, this model still performs the second-worst in this category compared to other categories. Overall, on one hand, all four models most frequently misclassify a "Criticism" tweet as a "Hostility" tweet (and "None of the Above" is the second-most frequently misclassified category). On the other hand, these models most frequently misclassify a "Hostility" tweet as "None of the Above," while Vidgen et al.'s models most frequently misclassify it as "Criticism." The predilection for predicting a tweet as "None of the Above" might be caused by the uneven distribution of data. Our training data contains 67.7% of "None of the Above" tweets and 19.5% of "Hostility" tweets, so it is reasonable that our models tend to classify tweets as "None of the Above" then as "Hostility."

5.4 Error in Data Annotation

Another important source of error is the error in data annotation. Vidgen et al. (2020) attribute about 17% of the error from the model to annotator error, which is, where upon review, the classification by the model seems more accurate than the actual annotations in the dataset. Note that it is not the case that 17% of the dataset is incorrectly annotated, but rather, 17% of the incorrect classifications can be attributed to annotator error. We similarly observe annotator error in our results. For instance, one text contains the following: "it isn't racist to name a virus after the location it started in, your disinformation isn't working." Here, the label by the annotator is "None of the Above," while the prediction by M2 is "Discussion." Because the content of the text is about whether or not calling COVID-19 the "wuhan virus" or "kung flu" is racist or not, it would more accurately be classified as "Discussion" instead of "None of the Above."

6 Conclusion

We presented work on creating a more robust classifier for detecting East Asian prejudice by introducing noise to the training data in the form of three different attacks: typos, whitespace insertion/deletion, and appending of neutral words. Our results show insignificant differences between models trained with and without data augmentation on the training data in terms of precision, recall, F1, and accuracy. However, in inspecting subtle differences in the behavior of the models, we found that, with data augmentation and a sufficient training period, the classifier was more likely to correctly classify "Counter Speech." Considering the lack of "Counter Speech" in the training data, we conjecture that data augmentation can be especially useful for classifying categories with insufficient training data because data augmentation increases cases of underrepresented categories in the training data.

In future work, data augmentation methods should be refined such that neither too few nor too many attacks are introduced into the data. Our data augmentation was done randomly in terms of the amount of augmentation, the location of augmentation, and the type of augmentation. Controlling and optimizing this process may lead to better results. In addition, we find that creating a more balanced dataset may help improve the performance of classifiers. If all labels are represented relatively equally in the dataset, the model will more likely train with a more similar emphasis on each class, improving the overall accuracy of the models. Finally, it is imperative that further work inspects how bias, perhaps most notably racial bias, manifests in these classifiers.

Acknowledgments

We thank our course instructor Professor Alexandra Schofield and student advisor Serenity Wade as well as all peer reviewers who have helped guide the work presented in our paper.

References

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate-speech detection.

Vebjørn Isaksen and Björn Gambäck. 2020. Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online. Association for Computational Linguistics.

John Kaufman, Peter Novid, Thorsten Brants, and Alex Franz. 2019. Google 10000 most common english words.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

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.

Maximilien Roberti. 2020. Fastai with huggingface transformers (bert, roberta, xlnet, xlm, distilbert).

Fatemeh Tahmasbi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2020. "go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19. *arXiv e-prints*, pages arXiv–2004.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting east asian prejudice on social media. *arXiv preprint arXiv:2005.03909*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.