Anti-Asian Hate Speech Classification

Ruian Shi

Department of Computer Science, University of Toronto, ian.shi@mail.utoronto.ca

Charita Koya

Department of Computer Science, University of Toronto, charitakoya@cs.toronto.edu

Kopal Garg

Department of Computer Science, University of Toronto, kopal@cs.toronto.edu

The COVID-19 pandemic has resulted in an increase in anti-East Asian hate speech in online platforms. While moderation and removal of offensive content is one approach to address this issue, we propose the development of an anti-Asian hate speech classifier for integration into a platform which notifies users of potential hateful content prior to posting. While current hate speech classifiers have been shown to accurately detect hateful content, the majority of hate speech classification literature does not focus on anti-Asian content and may result in under-representation of anti-Asian content in training datasets. In order to address these issues, we investigate whether classifiers without specific representation of anti-Asian hate can effectively classify anti-Asian hate speech by training baseline and state-of-the-art hate speech classification methods (LR, RF, BERT-family models) on pre-COVID-19 Twitter datasets with binary labels (offensive, non-offensive). We also apply a post-hoc explainability method, SHAP, on the trained anti-Asian hate speech detector to investigate the ability for explainability methods to provide sensible outputs. Our results showed that all the hate speech classifiers performed well in-domain but performed very poorly out-of-domain. In our augmented out-of-domain dataset, which consisted of a combination of general and Anti-Asian tweets, the BERT model performed the best in comparison to the other models, but when compared to the baseline and in-domain results, the performance of all models was diminished. We suggest taking a decolonial approach to hate speech classification by developing context-specific datasets, curated with specific ethnicities, cultures and problems in mind.

Additional Keywords and Phrases: Anti-Asian Hate speech, Language annotation, BERT models, Post-hoc explainability, Decolonization theory, NLP annotator bias

1 INTRODUCTION

The progression of COVID-19 has resulted in an increase in anti-East Asian hate speech in online platforms [1]. In 2020, approximately 1 in 5 hashtags containing #covid19 were anti-Asian [1]. While moderation and removal of offensive content is one approach to address this issue, we hypothesize that a preventative approach will yield better results. In this paper, we propose the development of an anti-Asian hate speech classifier for integration into a platform which notifies users of potential hateful content prior to posting. While hate speech classifiers have been shown to accurately detect hateful content, several issues exist in their deployment. The majority of hate speech classification literature does not focus on anti-Asian content and may result in under-representation of anti-Asian content in training datasets. Hate speech classifiers have also been shown to have an undesirable bias against groups and their colloquialisms [2, 3].

It is thus critical to develop an explainable hate speech classifier, which allows users to understand why their content is seen as hateful. This additional functionality provides several ethical advantages. Firstly, the adoption of explainable models enables decentralization of this hate speech classifier. By removing the black-box nature of the classifier, users are empowered to both educate themselves against hate speech and correct cases of erroneous classification. Education about hate speech has been identified by UNESCO as the most effective method to combat hate speech, while explainable predictions allow end-users to improve classification where biases have resulted in incorrect output [4, 5].

As hate speech detectors (and NLP methods in general) become widely deployed, we believe that explainable methods can aid in avoiding colonial frameworks of computing dominated by the centralization of knowledge [6], where the definition of hate speech captured in training datasets becomes universalized. We hope that removing the black-box nature of explainable methods will allow users to directly engage in model development by correcting erroneous aspects of model prediction. Re-training hate speech detection methods on community sourced data will gradually replace biases in training data with localized knowledge of hate speech. The interaction between the human and hate speech detection system can lead to increased trust and comprehensibility for users. Ideally, the transparency of explainable hate speech classifiers will prevent it from becoming a form of control through quantification [6] and allow communities to reduce hate speech on their own terms.

1.1 Contributions

We explore several interesting avenues in the general field of anti-Asian hate speech detection, with specific focus given to datasets of Twitter content.

- We conduct topic modeling and explore annotator bias in hate speech annotation tasks using multiple exploratory data analysis techniques.
- We investigate whether current hate speech methods are biased against detection of anti-Asian hate. Due to the sharp rise in anti-Asian hate speech content after the COVID-19 pandemic, we investigate whether classifiers without specific representation of anti-Asian hate can effectively classify anti-Asian hate speech. This experiment involves training baseline and state-of-the-art hate speech classification methods (LR, RF, BERT-family models) on pre-COVID-19 Twitter datasets with binary labels (offensive, non-offensive) [12, 13]. We then evaluate how the model performs on a dataset of anti-Asian hate tweets. We evaluate the performance gap between detection accuracy on held-out general hate tweets and held-out anti-Asian hate tweets. Observed performance gap supports the need to develop context-specific datasets for hate speech detection tasks.
- We apply a post-hoc explainability method called SHAP [18], on the trained anti-Asian hate speech detector to investigate the ability for explainability methods to provide sensible outputs [5, 17].

2 LITERATURE REVIEW

Numerous tools have been developed to detect varying hate speech content in online platforms. Vidgen et al. created a classifier to detect East Asian prejudice in social media data using a training dataset of 20,000 tweets. The classifier differentiated the dataset into four categories: hostility against East Asia, criticism of East Asia, meta-discussions of East Asian prejudice and a neutral class. The authors then conducted extensive data analysis, including 40,000 annotations with the aforementioned categories and additional flags for hostility, East Asian slurs and pejoratives. Of multiple models tested, the authors reported the RoBERTa model performed well across all categories with some misclassification errors in conceptually similar categories [7].

In another study by Dhamija et al., a comparative analysis of ML and deep learning algorithms was conducted to detect online hate speech. They used two datasets, one comprising 25000 tweets and the other a hate speech and personal attack dataset from zenodo.org. The tweets were labeled as hate, offensive or neither, with the latter two considered as non-hate for binary classification purposes. The data was cleaned using various feature engineering techniques and then classified using a number of algorithms, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Recurrent Neural Networks (RNN). Of these feature engineering techniques and classification models, it was concluded that BERT sentence embeddings with a Decision Tree algorithm had the most accurate results and had the potential to be used as a strong practical model [8].

Miok et al. investigated using Monte Carlo Dropout (MCD) in neural networks as a regularization approach for hate speech detection. Specifically, they analyzed Bayesian Attention Networks (BAN) and MCD-enhanced BERT models to conclude an improved calibration and prediction performance on hate speech detection in multiple languages. Compared to BAN, the MCD BERT model significantly and reliably improved the prediction performance in hate speech detection [9].

Putri et al. compared various ML algorithms and classification methods to determine the most accurate model for hate speech detection in a dataset of 4002 Indonesian tweets. They collected data using the Twitter API, classified the data against Naïve Bayes, Multi-Level Perceptron, AdaBoost Classifier, Decision Tree and Support Vector Machine algorithms, and conducted a comparison between models that did and did not use Synthetic Minority Oversampling Technique (SMOTE) for data balancing. The study concluded that the best model for hate speech classification was the Multinomial Naïve Bayes algorithm with unigram features without SMOTE [10].

The Salminen et al. study evaluated the performance of numerous classification algorithms, including LR, NB, SVM, XGBoost and Neural Network, using data collected from YouTube, Reddit, Wikipedia and Twitter. They also used a keyword-based classifier as the baseline model. They observed that XGBoost with BERT features outperformed the other models while LR, NB and SVM were the least accurate on all feature subsets [11].

3 DATA EXPLORATION

3.1 Data Sources and Processing Pipeline

We used 3 different datasets, as described in Table 1. The first dataset was obtained from Kaggle (KGEN), and it contained pre-COVID general hate speech-related tweets. It had 3 primary labels (hate speech, offensive language, neutral), which were re-encoded to 2 (hate speech, and neutral) by combining two categories, in order to facilitate a binary classification task [13]. The second dataset was obtained from a study by Vidgen et al., that investigated abusive and prejudiced content against East Asians on Twitter [14]. The dataset originally contained 4 labels (entity directed hostility, entity directed criticism, East-Asian prejudice, and neutral) which were re-encoded to 2 main labels (hate speech, or neutral). The final dataset contained general COVID-19-related tweets as well as tweets containing sensitive keywords. Each tweet was labeled as non-stigmatizing or stigmatizing along with information on the degree of perceived stigma. The target label, which originally contained 5 main categories (stigmatizing - low, stigmatizing - medium, stigmatizing - high, neutral, unknown/irrelevant), was again re-encoded to 2 main categories (stigmatizing content, or neutral) by combining the stigmatizing labels, and removing tweets labeled as unknown/irrelevant.

We wrote a standard tweet cleaning and preprocessing pipeline which involved eliminating textual and non-textual components, such as twitter handles, hyperlinks, non-ASCII characters, punctuation, numeric values, and lower-casing all tweets. We filtered out stop-words found in the *Natural Language Toolkit (NLTK)* suite, as well as words with little lexical content (e.g. the, a, also, etc.). Using the *emot* package, we replaced emojis with their Unicode CLDR short-name (e.g. ☺ replaced by 'happy face smiley') and using *pyspellchecker*, we corrected misspelled words and filtered out non-English words. We used standard lemmatization, and tokenization functions from *NLTK* to convert text to word tokens and replace words with their canonical forms (e.g. inflected forms of words like 'settling' and 'settled' were replaced with 'settle').

We experimented with 3 text encoding/vectorization techniques, namely: Unigram Bag-of-Words (uBoW), Bigram Bag-of-Words (bBoW) and Term Frequency-Inverse Document Frequency (TF-IDF) but displayed final results for TF-IDF, as that had the best performance across the board. We implemented this using *TfidfVectorizer* from *scikit-learn*, setting *min_df* to 0.2 and *ngram_range* to (1,1) to include unigrams. For experimentation with baseline models implemented using *keras*, like Support Vector Machine (SVM), Decision Trees (DT), Multinomial Naïve Bayes (MNB) and XGBoost, we used 10-fold cross validation for hyperparameter tuning. Finally, we computed model evaluation parameters like accuracy, precision, recall, but for brevity, only present F1-scores in this paper, as it is the most representative of model performance in our case.

Table 1: Dataset Summary

Dataset	Attribute	Summary	Theme
KGEN	#Unique tweets	24,783	Pre-COVID general hate speech
	#Unique labels	3	
EAP	#Unique tweets	19,884	Pre-COVID Anti-Asian and general
	#Unique labels	4	hate speech
STIG	#Unique tweets	11,263	COVID-19 Anti-Asian and general
	#Unique labels	5	hate speech

3.2 Topic Modeling

Since hate speech datasets are topic-specific, we analyzed the topics contained in the publicly available hate-speech dataset (KGEN) and compared them to hate-speech datasets relevant to the COVID-19 pandemic (EAP and STIG). We used an unsupervised topic modeling approach based on Latent Dirichlet Allocation (LDA), that discovers word-groups and similar expressions that best characterize the dataset. We explored the top 5 topics in each dataset and visualized the top-2 topics in the form of word clouds, as seen in Figure 1. We found that the top-5 topics in each dataset were vastly different. Since KGEN is a general hate-speech dataset obtained from Twitter, it contained expected topics revolving around abusive language or violence. Since EAP and STIG were collected during COVID-19, certain events and topics dominated social media at that time. This is clearly reflected in the top-5 topics. We hypothesized that because of this difference, models trained on KGEN would not generalize well on datasets that captured unobserved forms of hate and East-Asian racism that were perhaps exacerbated by the pandemic.

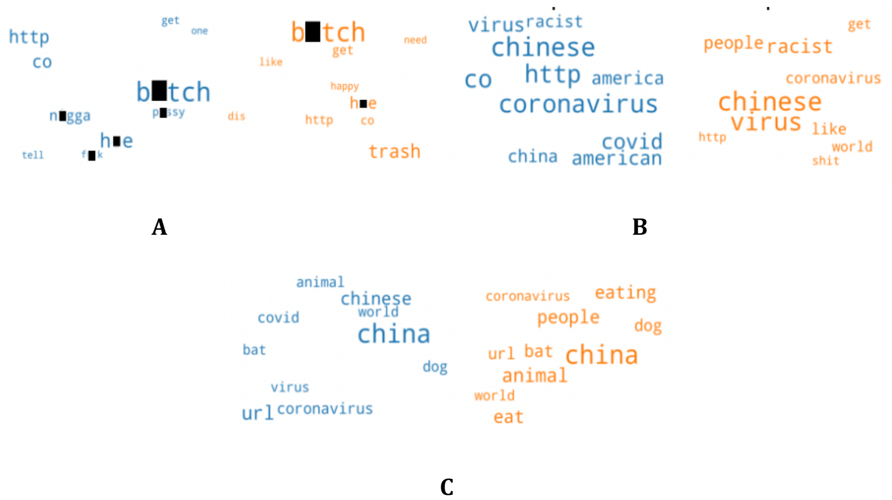


Figure 1: Datasets Topic Modeling. LDA modeling uses probability distributions over words to represent topics, where words with a higher probability take a larger form on the word cloud as they are more representative of the topic. *A.* KGEN. *B.* EAP. *C.* STIG

3.3 Annotator Agreement Analysis

To explore annotator bias and inconsistencies in data labeling, each author manually labeled 50 tweets from the KGEN datasets. Each tweet was given a label of either 0 for hate speech, 1 for offensive, and 2 for neither. We then used confusion matrices to summarize the count values for degree of concordance between the provided labels and our own annotations, as shown in Figure 2, with the horizontal and vertical axes representing our labels and the provided labels, respectively. Based on these results, we found that the datasets we fine-tuned contained many ambiguities, as the meaning of the tweets were highly dependent on the context and the combination of words, rather than individual words themselves. As such, due to this ambiguity, the annotators' own knowledge and biases significantly influenced the labels, demonstrating how the process of labeling is highly subjective and how annotator bias could potentially influence model performance.

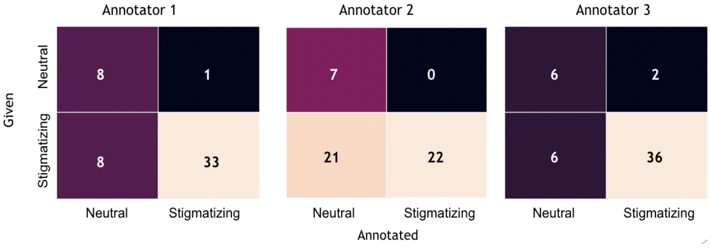


Figure 2: Datasets Topic Modeling. LDA modeling uses probability distributions over words to represent topics, where words with a higher probability take a larger form on the word cloud as they are more representative of the topic.*A.* KGEN. *B.* EAP. *C.* STIG

4 EXPERIMENTAL DESIGN

Given the established inconsistencies in hate speech classification datasets, we aimed to demonstrate the performance drop associated with applying a hate speech classifier (HSC) trained on generic hate tweets to the domain of COVID-19 anti-Asian hate tweets.

Using the previously described datasets, we set up four experimental tasks which evaluated cross-domain performance. We provide a visualization of the evaluation tasks in Figure 3. We first split all datasets into train, validation, and test splits in a (0.75, 0.15, 0.15) ratio, respectively. In the **Baseline** task, we trained HSCs on the train split of the STIG dataset, and then evaluated on the test split. This task served to establish the baseline ability for HSCs to perform, given access to the true distribution of COVID-19-related anti-Asian hate speech. In **Task 1**, we trained on the KGEN train split, and evaluated on the KGEN test split. This task represented the ability for HSCs to perform when applied on in-domain data. In **Task 2**, we trained on the KGEN train split, and evaluated on the STIG test split. This task represented the ability for HSCs to generalize their representation of hate tweets out-of-domain, specifically in this case to COVID-19-related anti-Asian hate tweets. Finally, in **Task 3**, we trained HSCs on the combined training splits of the KGEN and EAP datasets and evaluated on the test split of the STIG dataset. This task aimed to investigate the effects of augmenting generic data with context-related data.

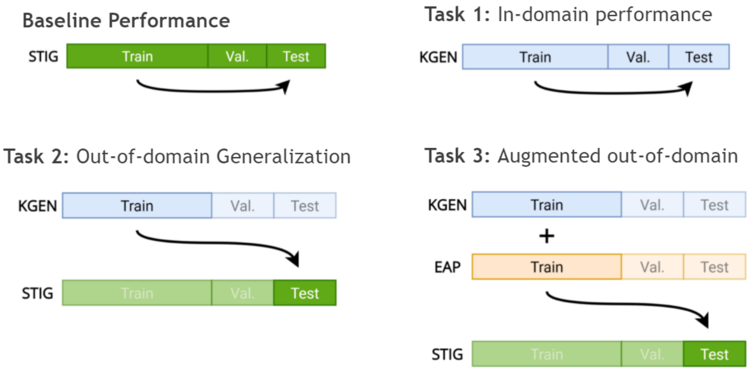


Figure 3: Experimental setup for evaluation of hate speech classifier generalization across types of hate speech. Each bar represents a dataset split into training / validation / testing subsets. The start of the arrow represents the training subset used by a HSC for a task, and the end of the arrow represents the subset the HSC is tested on. In Task 3, the train set of KGEN and EAP are concatenated.

In addition to the previously established baseline methods (SVM, DT, MNB, and XGBoost), we also evaluated the performance of the BERT language model [15]. BERT is a deep-learning model based on the Transformer architecture [16], which has recently become ubiquitous in the field of natural language processing. BERT is pre-trained on a large corpus of unlabelled English text (from unpublished books and Wikipedia) to learn embeddings for English words which are dependent on their context within a sentence. These embeddings are then fed into a feed-forward neural network to output classifications. Pre-trained BERT models are typically fine-tuned for a specific context. For our experimental tasks, we used the 'bert-base-uncased' pretrained model, and then fine-tuned using the training data specified under each evaluation task. Due to the specific tokenization requirements for BERT models, our data pre-processing varied slightly from baseline methods, but efforts were made to keep them comparable by applying similar transformations when possible.

All HSCs were evaluated using the F1 score on binary classification of detecting whether a specific tweet contains hateful content. We did not perform extensive hyper-parameter search for the BERT model, only performing a grid search for initial learning rate from the grid [1e-4, 1e-5, 1e-6], using validation F1 score for model selection. We trained for a maximum of six epochs and performed early stopping when validation loss plateaued.

5 RESULTS

5.1 Generalization Gaps Incurred by Non-Specific Models

We evaluated all baseline methods and pre-trained BERT models on the four evaluation tasks defined above. We selected the two best performing baseline models and the BERT model; the results are reported in Table 2.

Table 2: Performance of HSCs on previously defined experimental tasks which approximate cross-domain performance. Rows Train DS and Test DS re-iterate the train and test subsets used by each task. After these rows, reported numbers correspond to model performances evaluated by the F1 score. The performance gap column computes the difference in F1 score between the baseline task and the out-of-domain task. The best method per task is bolded.

Dataset / Model	Baseline	Task 1: In-domain	Task 2: Out-of-domain	Performance Gap	Task 3: Augmented Out-of-domain
Train DS	STIG	KGEN	KGEN	-	KGEN+EAP
Test DS	STIG	KGEN	STIG	-	STIG
Naïve Bayes	0.678	**0.982**	0.552	-0.430	0.247
XGBoost	0.630	0.925	**0.645**	**-0.395**	0.572
BERT	**0.719**	0.918	0.009	-0.828	**0.692**

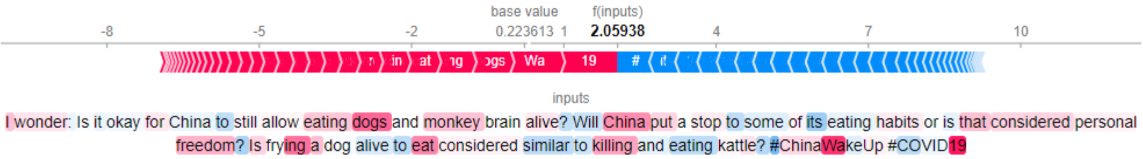
The experimental results in Table 2 confirmed our hypothesis that HSC trained on generic datasets perform poorly when exposed to specific hate tweet contexts. In the Baseline column, we see that all methods performed reasonably well at COVID-19-related anti-Asian hate speech classification when provided with similar training data. Similarly, HSCs trained on generic datasets performed well when classifying generic hate tweets, demonstrated in Task 1. However, all HSC methods trained on generic data suffered a drop in performance when encountering unseen COVID-19 related anti-Asian hate tweets, quantified in the performance gap column.

We believe Task 2 to approximately represent the actual performance gap associated as the COVID-19 pandemic emerged. Although hate speech classifiers in deployment are likely to be more sophisticated than those in our experiments, they would similarly not have access to COVID-19-related anti-Asian hate speech data during the early stages of the pandemic, as training data would not yet have been collected. Thus, the observed drop in performance when generalizing to unseen categories of hate tweets would have meant that any hate speech detection system would have largely missed COVID-19-related anti-Asian hate tweets. In our experiments, the ubiquitous BERT model suffered the worst drop in performance. This would have resulted in harm to affected communities, all the while maintaining a false impression of successful hate tweet classification.

As new contexts for hate speech arise during emergent situations, the lack of training data poses a challenge for developing context-specific HSCs. As shown by Task 3, augmentation of generic data with general, non-COVID-19-related anti-Asian tweets provides a method to boost generalization performance on COVID-19-related anti-Asian hate tweets. The BERT model benefits most from this augmentation, performing on par with the baseline task, while augmenting baseline methods causes them to perform worse.

5.2 Explainability Methods for Hate Speech Classification

Post-hoc explainability methods such as SHAP [18] can be applied to trained methods to obtain attributions on which input features resulted in the predicted output. In the context of hate tweet detection, SHAP is able to identify the words which cause models to label tweets as hateful. In Figure 4, we provide an example of SHAP values applied to the BERT model from the baseline task, visualized on a hateful tweet from the STIG dataset. Words which contribute towards a positive (hateful) classification are highlighted in red, while those which contribute towards a negative classification are highlighted in blue. The numerical value on the top represents the output of the BERT model. As shown in Figure 4, sensitive words which are typically associated with anti-Asian stigmatizing speech are correctly highlighted in red, while words not related to stigmatizing speech are shown in blue. While some of the SHAP values correctly correspond with stigmatizing language, we see that some of the discovered relations are less clear in correctness, resulting in lower interpretability.



I wonder: Is it okay for China to still allow eating dogs and monkey brain alive? Will China put a stop to some of its eating habits or is that considered personal freedom? Is frying a dog alive to eat considered similar to killing and eating kattle? #ChinaWakeUp #COVID19

Figure 4: SHAP values for the BERT model trained using the baseline experimental setup. The evaluation tweet is obtained from the STIG dataset. Words highlighted in red contribute towards a positive (hateful) classification, while words in blue contribute towards a negative classification. The intensity of the color represents the size of the contribution.

Although the accuracy of these explainability methods is an area for future investigation, we see two benefits to explainability methods. First, explainability methods allow the "black-box" of machine learning methods to be opened. By revealing the reasons why a prediction is made, users of the method are better able to understand logic used by the method, and identify discrepancies between the model's potentially flawed concept of hate speech and their own definition. Ultimately, this could allow the community to identify when HSCs are failing to generalize to specific contexts. For example, the failure of models to correctly attribute stigmatizing language could prompt an investigation into model generalization performance.

Furthermore, the output of these explainability methods could be deployed in educational contexts. As previously discussed, education has been identified as one of the best preventative measures against hate speech. Hate speech classifiers could be used to detect hate speech prior to users posting hateful tweets, and explainability methods can specifically identify the parts of the tweet the HSC believes to be stigmatizing or hateful. This could support educational methods against hate speech, or alternatively, if the model is biased to erroneously detect hate speech when none exists, allow for identification of false classifications and recourse for affected users.

6 DISCUSSION

6.1 Decolonization Theory

As demonstrated by our findings, the models performed well on the KGEN dataset, which consisted of general hate speech and offensive tweets, but performed very poorly outside of the training domain, i.e., when applied to the Anti-Asian dataset; this demonstrates how training the models with the general dataset was ineffective when tested against a culture-specific dataset.

Datasets are typically constructed based on Western language and biases, which is representative of a centralized source of knowledge. The one-truth assumption that this can be applied in other contexts serves to marginalize non-Western groups, as the meaning of certain terminology and tweets cannot be generalized outside of these general datasets. For instance, within our datasets, there were terms that were specific to Asians and could be classified as offensive or hate speech, but out of context, they took on a completely different meaning. Furthermore, the knowledge and biases of the annotators also influenced the manual labeling of the general dataset.

Building models based on this one centralized knowledge source highlights the colonial bias and assumption that Western interpretations of text and classification are assumed to be universal and applicable to all groups. However, our experiments show that such models and results cannot be generalized, which is why there is a need for the development of context-specific datasets, curated with specific ethnicities, cultures and problems in mind. By using culture-specific datasets, we are ensuring a decolonial approach to hate speech classification, in which the hegemonic Western knowledge system is no longer utilized as a centralized source of knowledge.

6.2 Annotator Bias

In our work, we briefly investigated annotator bias in hate speech annotation tasks where we randomly selected 150 tweets from the KGEN dataset and split the data in such a way that each annotator had to independently label a disjoint set of 50 tweets. We used confusion matrices to picture inter-annotator agreement, and through this analysis, we noticed that offensive data had moderate levels of agreement. In general, this makes the analysis of subjective data like stigmatizing tweets highly challenging. The performance and reliability of classification models is dependent on training labels, and inconsistent annotations can lead to false confidence in its performance.

It is difficult to develop better definitions and guidelines around consistency in annotations, as people perceive stigmatizing content differently and strict guidelines would inevitably end up hindering the annotators' freedom of decision making. There is a lack of gold standard when it comes to hate speech annotations but relying on tweet content alone might be insufficient in classifying it equivocally as offensive or not.

Techniques like quantification of inter-annotator agreement are typically used to resolve inconsistencies in disagreed-upon instances. We argue that this might be an inefficient technique, as it operates on a one-truth assumption, i.e., a single annotation towards which opinions converge, as this clearly does not hold for subjective annotation tasks. If anything, such metrics serve as quantitative indices that measure the degree of polarization in judgment of stigmatizing content and instance-level task difficulty. A single metric fails to capture and leverage a divergence of opinions and it is, therefore, difficult to generate a high-quality reference source and build models that can encode multiple perspectives.

7 CONCLUSION

Online hate speech is a challenging issue for minorities and other targeted groups, particularly Asians during the COVID-19 pandemic. In this study, we investigated popular NLP models, including LR, RF, and BERT-family models, and determined that classifiers without specific representation of anti-Asian hate are ineffective in classifying anti-Asian hate speech online. We also concluded that the performance of all models significantly decreased in out-of-domain and augmented-out-of-domain datasets in comparison to in-domain, as the models were unable to effectively recognize anti-Asian hate speech when trained against general datasets. In order to target online hate speech, we proposed developing an anti-Asian hate speech classifier for integration into a platform which notifies users of potential hateful content prior to posting. We suggest taking a decolonial approach for development of this classifier by building context-specific datasets, keeping specific cultures and languages, as well as researcher and annotator bias in mind, in order to effectively represent and combat anti-Asian hate speech online.

REFERENCES

[1] Hswen, Y., Xu, X., Hing, A., Hawkins, J., Brownstein, J. and Gee, G., 2021. Association of “#covid19” Versus “#chinesevirus” With Anti-Asian Sentiments on Twitter: March 9–23, 2020. American Journal of Public Health, 111(5), pp.956-964.

[2] Mozafari, M., Farahbakhsh, R. and Crespi, N., 2020. Hate speech detection and racial bias mitigation in social media based on the BERT model. PLOS ONE, 15(8), p.e0237861.

[3] Sap, M., Card, D. and Smith, N., 2019. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[4] Unesdoc.unesco.org. 2020. Education as a tool for prevention: addressing and countering hate speech, Expert meeting: 13-18 May 2020. [online] Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000379146?locale=en> [Accessed 15 October 2021].

[5] Ribeiro, M. and Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16.

[6] Dourish, Paul, and Scott D. Mainwaring. "Ubicomp's colonial impulse." *Proceedings of the 2012 ACM conference on ubiquitous computing*. 2012.

[7] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, and R. Tromble, “Detecting East Asian prejudice on social media,” *Proceedings of the Fourth Workshop on Online Abuse and Harms*, May 2020.

[8] T. Dhamija, Anjum, and R. Katarya, “Comparative analysis of machine learning and deep learning algorithms for detection of online hate speech,” *Advances in Mechanical Engineering*, pp. 509–520, Jun. 2021.

[9] K. Miok, B. Škrlj, D. Zaharie, and M. Robnik-Šikonja, “To ban or not to ban: Bayesian attention networks for reliable hate speech detection,” *Cognitive Computation*, 2021.

[10] T. T. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean, “A comparison of classification algorithms for hate speech detection,” IOP Conference Series: Materials Science and Engineering, vol. 830, p. 032006, 2020.

[11] J. Salminen, M. Hopf, S. A. Chowdhury, S.-gyo Jung, H. Almerekhi, and B. J. Jansen, “Developing an online hate classifier for multiple social media platforms,” *Human-centric Computing and Information Sciences*, vol. 10, no. 1, 2020.

[12] Kaggle.com. 2021. Hate Speech and Offensive Language Dataset. [online] Available at: <https://www.kaggle.com/mrmorj/hate-speech-and-offensive-language-dataset> [Accessed 15 October 2021].

[13] Kaggle.com. 2021. Twitter hate speech. [online] Available at: <https://www.kaggle.com/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv> [Accessed 15 October 2021].

[14] Vidgen, B. and Hale, S., 2020. Detecting East Asian Prejudice on Social Media. [online] Available at: <https://arxiv.org/abs/2005.03909> [Accessed 15 October 2021].

[15] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems pp. 5998-6008.

[17] Ziems, C. and Kumar, S., 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. [online] Available at: <https://arxiv.org/abs/2005.12423> [Accessed 15 October 2021].

[18] Lundberg, S.M. and Lee, S.I., 2017, December. A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777)