Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, Jatinder Singh
Compliant & Accountable Systems Group
University of Cambridge, UK
arb229@cam.ac.uk, {michelle.sengah.lee, jennifer.cobbe, jatinder.singh}@cst.cam.ac.uk

ABSTRACT

Automated systems for detecting harmful social media content are afflicted by a variety of biases, some of which originate in their training datasets. In particular, some systems have been shown to propagate *racial dialect bias*: they systematically classify content aligned with the African American English (AAE) dialect as harmful at a higher rate than content aligned with White English (WE). This perpetuates prejudice by silencing the Black community. Towards this problem we adapt and apply two existing bias mitigation approaches: preferential sampling pre-processing and adversarial debiasing in-processing. We analyse the impact of our interventions on model performance and propagated bias. We find that when bias mitigation is employed, a high degree of predictive accuracy is maintained relative to baseline, and in many cases bias against AAE in harmful tweet predictions is reduced. However, the specific effects of these interventions on bias and performance vary widely between dataset contexts. This variation suggests the unpredictability of autonomous harmful content detection outside of its development context. We argue that this, and the low performance of these systems at baseline, raise questions about the reliability and role of such systems in high-impact, real-world settings.

CCS CONCEPTS

• **Social and professional topics → Race and ethnicity**; *Censorship*; • **Computing methodologies → Machine learning**; Natural language processing.

KEYWORDS

bias, fairness, racial disparities, dialect, content moderation

ACM Reference Format:
Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3442188.3445875

1 INTRODUCTION

As the body of social media content grows explosively, so does the problem of *harmful content*: hate speech, cyberbullying, and online abuse, among others [50]. Manual content moderation can be expensive for platforms given the volume and rates of posted content, and is both tedious and traumatic for human moderators [19, 22]. Waiting for users to flag content as inappropriate allows harmful content to proliferate with potentially widespread consequences; this has prompted interest in content moderation *ex ante*, as content is posted and before it reaches an audience [5, 10]. Recently, the COVID-19 pandemic has sent home thousands of human moderators, many of whom are not allowed to work from home [39]. For all these reasons, automated detection of harmful content by artificial intelligence (AI) systems has been the subject of considerable academic and industrial research [25, 41, 50].

1.1 Racial Dialect Bias

Large platforms such as Facebook, Google (including YouTube), and Twitter have turned to automation as a crucial element of "industrial" content moderation [9]. They use automated screening to flag content for human review even before it is flagged by users [5, 9]. One key reason for the introduction of these automated systems was to reduce the prevalence of hateful speech and incitements of violence against marginalised communities. In other words, one of their central goals is to protect historically disadvantaged groups. Yet they have been shown to perpetuate biases against various marginalised communities which arise from datasets on which they are trained [14, 17, 34, 48].

Previous research has measured racial dialect bias in harmful tweet detection systems [14, 48]. Such research focuses on Twitter due to high availability of labeled harmful tweet corpora [58] and because Twitter is an important space for Black activism and activism in general [48], but we expect this issue to exist on other online platforms as well. Tweets with high predicted *African-American English (AAE)* alignment p_{AAE} are found to be classified as harmful at a higher rate than tweets with high predicted *White English (WE)* alignment p_{WE} or low p_{AAE} [14, 48]. These alignments are calculated using a model [6] that was trained on a corpus of 60 million geolocated tweets using US Census data as topics. The model is shown to accurately follow known linguistic phenomena [6].

Twitter can take a range of actions if a tweet violates their rules, which protect against hate speech, incitements of violence, and targeted harassment of an individual or group, among other things [55]. They may hide the tweet behind an interstitial warning, limit its visibility in search results and feeds, or even require its removal and hide it in the meantime [54]. However, each of these enforcement actions, when differentially applied across populations, could amplify societal injustices [48], either by enforcing stereotypes held by users or by outright silencing minority communities. It is painfully ironic that harmful tweet detection systems may systematically diminish the voice of the Black community given that they exist, at least in part, to shield this and other marginalised

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445875

communities from harm. Furthermore, bias may reduce their ability to truly detect harmful content by rewarding them for simply associating linguistic properties of AAE with harmful labels.

1.1.1 *Dialect Alignment as a Continuous Variable*

In the fairness and bias sphere, a *protected attribute* is a characteristic that is protected against discrimination, and against which bias is suspected or known to exist [40]. Race is a widely accepted protected attribute, but in this context the race of a tweet's author is often unknown; the best we can do without additional information is to estimate probabilities of dialect alignment such as p_{AAE} [6]. AAE itself contains much regional and social variation [16], and the dialect estimation tool is probabalistic and so has some margin for error.

Most previous research dichotomises data by considering a tweet to be highly AAE aligned if the dialect model assigns it $p_{AAE} > 0.8$. An exception is Sap et al. [48], who treat p_{AAE} as continuous when evaluating dataset bias; however, they too dichotomise the data for model evaluation. Using only high-confidence dialect predictions reduces the risk of compounding errors from dialect estimation and harm prediction. However, this approach is only effective in large and diverse datasets, where an appreciable number of tweets satisfy this constraint. Harmful tweet corpora are smaller and less representative, and as such often contain too few "high-AAE" tweets for meaningful analysis: three of the datasets considered in this study contain a mere 26 [21], 14 [23], and 2 [62] instances of such tweets. In these datasets, the problems associated with dichotomisation, which include reduced statistical power and obfuscation of variation within groups [1], cannot be ignored. It is also worth noting that low dialect diversity in training datasets unsurprisingly causes bias and poor performance on dialect-aligned input in systems [6].

In this paper we treat p_{AAE} as continuous in line with precedent [48] and also assert that this is justified by empirical observation (See Appendix §A.1). This allows us to evaluate bias by comparing predictions to dataset labels while preserving statistical power, whereas existing research on racial dialect bias [14, 48] either relies on evaluation via external datasets or loses power by dichotomisation at the model bias evaluation stage. Our treatment of p_{AAE} as continuous has considerable implications: there is no *privileged* or *unprivileged* group, so our definitions of fairness and bias mitigation efforts must address p_{AAE} itself rather than group membership.

1.1.2 *Bias Mitigation*

The fairness of AI systems in general is under intense scrutiny [27, 28, 40, 57, 59, 63]. A "first wave" of algorithmic accountability is working to address known bias and discrimination in AI systems, while a "second wave" asks broader questions about the role and governance of autonomous systems [43]. Addressing racial dialect bias in harmful content moderation is particularly relevant and important of late as the world reckons with systemic racism, as recently highlighted by the death of George Floyd, and far too many other Black people, at the hands of police in the US [47]. Vigilant rejection of online content that incites or perpetuates hatred of any kind, particularly that aimed at Black people or other marginalised groups, is an important step toward dismantling institutional racism. Yet to systematically silence Black voices online in pursuit of this goal is categorically counter-productive.

It may be possible to mitigate racial dialect bias in automated harmful tweet detection using technical interventions. Automated bias mitigation has been explored in other machine learning (ML) contexts [40], though it has been criticised for being incomplete at best, and at worst for obscuring the root problem and inducing unforeseen consequences [45]. Some research has explored the problem of racial dialect bias in detecting harmful online content [14, 48], but little attention has been paid to addressing this bias. Doing so is not a straightforward task: the continuity of dialect, nuance of language processing, and challenges of defining "harm" demand new and adapted approaches.

We apply automated bias mitigation techniques to racial dialect bias in harmful tweet detection systems. In keeping with previous research, our efforts are focused only on this specific type of bias, although these systems are afflicted by many other types, some of which may have yet to be clearly identified [17, 42, 64]. We evaluate the bias mitigation approaches in terms of their ability to reduce dialect bias, measured in a variety of ways, and their effects on classification performance. While they show some promise, both the extent of bias reduction and the impact on classification performance are highly variable across *dataset contexts*: different ways data are collected, pre-processed, and annotated according to dataset creators' research goals. These differences are in turn responsible for differences in dataset size, class definitions, and distributions of linguistic features such as dialect, among other things. We argue that this variation and the associated uncertainty in harmful content detection systems raise important questions about the role of such systems in society, and underscore the difficulties of deploying opaque autonomous systems with real-world impacts.

2 BACKGROUND

2.1 Harmful Content Detection

2.1.1 *Motivation*

Although some have argued that autonomous detection of harmful content is necessary, it is a difficult task [50]. The terminology used to describe harmful content is abstract and inconsistent [15, 50] and harm itself is a subject-dependent, broadly defined category [61, 65]. Automated detection of harmful tweets presents still more challenges, given the colloquial, short, and "noisy" nature of tweets [65]; the lack of background knowledge and context in ML systems — something that is acceptable in one context may be wholly inappropriate in another [25, 50]; and differences in cultural contexts that cannot be standardised [9].

2.1.2 *Existing Approaches*

While social media content can take many multimedia forms, text posts remain prevalent, and are therefore our focus. Consistent with much literature in the space, our research focuses on Twitter, though our findings may be relevant to other text-oriented online platforms. State-of-the-art harmful content detection systems consist of neural networks which train on deep text features, but networks that train on *surface-level* (word and character) features have also been shown to perform well [2, 34, 50]. Such systems often embed word and character *n-grams* (contiguous sequences of *n* items) using either bag-of-words or term frequency-inverse document frequency (TFIDF), which normalises counts according to the frequency of the n-gram in question in the whole corpus [22, 50]. Trained on labeled harmful content corpora, these systems classify unseen text as problematic or benign.

2.2 Bias in Harmful Content Detection

ML systems for detection of harmful content, like many other ML systems, are prone to various types of bias [14, 17, 42, 48, 64].

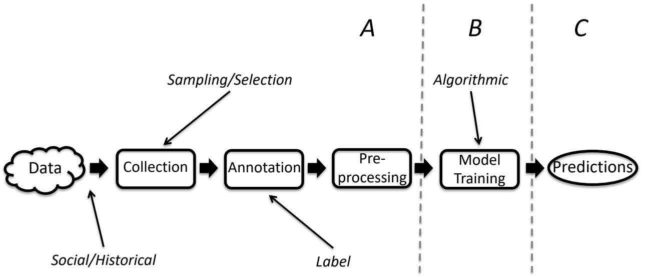

Figure 1: The ML pipeline, indicating forms of bias. Pre-processing bias mitigation intervenes in region *A*, in-processing in region *B*, and post-processing in region *C*.

Research has uncovered racial dialect bias in harmful tweet detection datasets and systems [14, 48]: findings show correlations between p_{AAE} and harmful labels in training datasets, and that models systematically classify tweets aligned with AAE as harmful at higher rates than their WE-aligned counterparts.

Different forms of bias can arise at different stages in the data processing pipeline [53]. Figure 1 shows some forms of bias that are particularly relevant in this context. *Social* [20] or *historical* [40] bias is driven by past decisions, actions, sentiments, norms and laws that we may consciously recognise as outdated or inappropriate, but which nonetheless shape the structural conditions of society. Social and historical biases may lead to differences in underlying distributions of harmful content across dialects; for instance, research has shown a correlation between profanity use on Twitter and the prevalence of AAE in a geographical region [7]. Data collection can produce *sampling* [40] or *selection* [52] biases, when non-random sampling or data cleaning lead to data that are not representative of real-world distributions. Sampling bias could lead to correlations between harmful class labels and AAE in datasets, which might not exist in random samples. At the point of annotation, *label* bias [52] may arise if annotators label tweets differently according to dialect, based on either human prejudice or cultural differences. Finally, bias can be propagated into models and even exacerbated at the stage of model training, a phenomenon called *over-amplification* [52] or *algorithmic* bias [40]. *Deployment* bias [53] arises when systems are used, or their outputs interpreted, in inappropriate ways; however, this occurs after predictions are made and is therefore beyond the scope of this paper.

2.3 Automated Bias Mitigation

Automated bias mitigation is an increasingly popular subject of ML and natural language processing (NLP) research [40, 52]. It intervenes at one of three phases: *pre-processing* bias mitigation changes the input data, *in-processing* bias mitigation changes the model itself, and *post-processing* bias mitigation changes model outputs [13]. These distinctions are shown in Figure 1. Given input X and class labels Y, these approaches aim to make predictions $\hat{Y} = f(X)$ fair with respect to a protected attribute S.

2.3.1 Defining Fairness Many different mathematical notions of *algorithmic fairness* exist [57]. In keeping with existing research in this field [6, 14, 48], we consider *group fairness*, which dictates similar treatment across demographic lines. *Individual fairness* — which dictates similar treatment for similar individuals — could potentially be assessed in this context. Though defining "similarity" here is challenging, it may be possible using synthetic datasets [17, 42] or appropriate NLP-specific metrics [18]. Consistent with the existing literature, we consider group fairness, acknowledging that exploring individual fairness may be an area for future work.

Demographic parity [57] (parity) requires independence between predictions $\hat{Y}$ and the protected attribute S — in this case that means predictions made by harm detection systems are independent of p_{AAE}. Alternatively, it may be beneficial to account for base rates. *Equal odds* [26] requires independence of $\hat{Y}$ and S conditional on ground truth labels Y — in this case it requires that each label's false positive and false negative rates are independent of p_{AAE}. This independence ensures both that benign AAE-aligned tweets are not moderated disproportionately frequently, and that harmful tweets with low AAE alignment are not overlooked disproportionately frequently. We evaluate systems using both parity and equal odds, as it is important to (respectively) consider predictions both in isolation and as they relate to real-world outcomes.

2.3.2 Existing Bias Mitigation Approaches NLP-specific bias mitigation approaches tend to intervene at the pre-processing phase. Examples include debiased word embeddings [8] and counterfactual data augmentation [68], both of which have been explored in the context of gender bias. More general pre-processing approaches aim either to modify the training dataset via massaging [31], reweighing [31, 52], or resampling [31, 35, 35, 52]; or to create intermediate representations of training data on which models then train [36, 66]. These approaches generally assume a binary protected attribute, but most can be adapted for a continuous one such as p_{AAE}.

Some in-processing approaches use a regularisation term to directly penalise any dependence bewteen S and Y within the model's training loss function [32, 38]. Others use adversarial learning, in which an *adversary* network is trained to predict S based on $\hat{Y}$, thereby coaxing the *predictor* network to make predictions that are independent of Y [4, 24, 59, 67]. Regularisation has been adapted to be compatible with continuous Y and S [38], and adversarial debiasing is theoretically shown to be as well [67]; however, we are not aware of any continuous-S applications of adversarial debiasing.

Post-processing bias mitigation relies on the use of a holdout *validation* data subset and does not require knowledge of the mapping between input data and predictions [13, 40]. This makes it an appealing choice for third parties or when classification occurs in a "black box." However, without access to input data or predictive models, post-processing approaches can only reassign some predictions according to a function [40]. We focus on strategies for mitigating bias when the data and learning process *can* be accessed and modified (i.e. pre-processing and in-processing methods) because these interventions aim to ameliorate the problem of model bias itself, rather than its symptoms.

3 RESEARCH DESIGN

3.1 Datasets

We consider four labeled harmful tweet datasets: Founta et al. [21], Davidson et al. [15], Waseem & Hovy [62], and Golbeck et al. [23]. These (along with a dataset produced by Waseem [60] that we deemed too small for reliable classification and bias evaluation in our study) comprise a prominent set of English-language datasets with tweets labeled by some type or types of harm [14].

Founta et al. *boost* the prevalence of harmful tweets by combining randomly sampled tweets with a sample of tweets that are likely to belong to harmful classes based on text analysis and results of earlier labelling rounds. Each of the other three datasets filters tweets based on the presence of words from a hate speech lexicon [15], words relating to religious or ethnic minorities [62], or a curated list of terms that correlate with harassment in an exploratory search [23].

Founta et al. define "hateful" as language that is hateful toward an individual or group; "abuse" as an insult, debasement, or violent targeted interaction; "spam" as unwanted information, and everything else "normal". Waseem & Hovy's definitions of "sexism" and "racism" are straightforward, and Golbeck et al. define "harassment" as anything that is extremely violent or offensive, threatening, hateful toward a group, or designed to upset an individual or group.

Davidson et al. define "hate speech" as targeting a specific group, "offensive language" as all other tweets that are perceived as offensive, and "neither" as everything else. It is not immediately clear whether the "offensive language" class should be considered harmful or not. However, in keeping with prior research [48] we consider it harmful for two reasons: First, the dataset's creators note that human annotators tend to consider sexist language to be offensive rather than hateful [15], but targeted sexism is clearly harmful [61] and violates Twitter's rules [55] concerning abuse. Second, we find that many tweets labeled as merely offensive contain language that can be considered as unequivocally hateful — slurs such as *'n*gger'* and *'f*ggot'* — which we feel ought to be considered harmful.

Importantly, the datasets use different annotation methods. Founta et al. and Davidson et al. crowdsource amateurs to undertake annotation (five and at least three annotators per tweet, respectively), whereas Waseem & Hovy and Golbeck et al. annotate the data themselves. Waseem & Hovy's annotations are reviewed by an outside gender studies expert, and Golbeck et al. underwent extensive training prior to annotating their data. Such differences in sampling and annotation methods, which result in different dataset sizes and class distributions, impact downstream classifiers [58] and can differentially give rise to some of the biases outlined in §2.2.

For Davidson et al.'s dataset, in line with its creators [15] we only consider the tweets for which a majority of annotators (at least half) agree on a label. We also gathered results for the subset of Davidson et al. for which there was full agreement among annotators; however, the results were very similar to the majority data, so for the sake of brevity they are excluded. The four datasets are used to train models with a label-stratified 80/20 train/test split.

Dataset bias gives an important standard against which to measure propagated bias. Previous research [48] measures dataset bias using the Pearson-r correlation between each label and p_{AAE}, which we call r_{label}. It is also useful to consider the mean value of p_{AAE} among all tweets in a certain class; this is more resilient to uneven class sizes but less so to the size and location of the overall distribution. Table 1 shows our calculations of the bias present in the full datasets (training and test sets combined) in terms of per-label $p_{AAE,avg}$ and r_{label}. Harmful labels tend to have higher $p_{AAE,avg}$ and positive r_{label}, whereas non-harmful labels tend to have lower $p_{AAE,avg}$ and negative r_{label}. This suggests the presence of bias in the sense that there is a positive relationship in the data between AAE alignment and harmful labels.[1] Dialect bias exists in all of the datasets, but it is far more extreme in the datasets produced by Founta et al. and Davidson et al. This may be due in part to their use of amateur annotation, but is likely also related to the means by which they collect and preprocess their data.

Dataset	Label	Count	$p_{AAE,avg}$	r_{label}
Founta et al.	normal	37,628	0.148	−0.198
	spam	8,232	0.157	−0.017
	abusive*	4,950	0.246	0.26
	hateful*	1,993	0.217	0.106
Davidson et al.	offensive- language*	19,097	0.442	0.393
	neither	4,119	0.214	−0.386
	hate speech*	1,410	0.322	−0.086
Waseem & Hovy	none	10,983	0.168	−0.045
	sexism*	3,359	0.194	0.135
	racism*	1,974	0.147	−0.103
Golbeck et al.	normal	14,669	0.156	−0.056
	harassment*	5,182	0.168	0.056

Table 1: Dataset bias, measured by per-label mean p_{AAE} and the Pearson-r correlation between each label and p_{AAE} (r_{label}). * denotes harmful labels.

We use two additional datasets to evaluate bias in model predictions extrinsically. Blodgett et al. [6] contains nearly 60 million tweets labeled by p_{AAE} and p_{WE} estimations. Preoţiuc-Pietro and Unger [46] contains nearly 6 million tweets labeled by the self-reported race of the author. Because these datasets are not labeled according to harm, we do not evaluate them for bias.

3.2 Bias Evaluation Metrics

3.2.1 Intrinsic Bias Metrics We measure bias *intrinsically* by examining the relationship between dialect and predictions made on held-out testing subsets of our training datasets. Because we consider p_{AAE} as continuous, we cannot easily split tweets into privileged and unprivileged groups, a prerequisite for most bias metrics. We therefore define four new metrics. The first is

$$\Delta p_{AAE} \equiv \overline{p_{AAE}}_h - \overline{p_{AAE}}_n$$

where $\overline{p_{AAE}}_h$ and $\overline{p_{AAE}}_n$ are the average values of $p_{AAE,avg}$ across all harmful and non-harmful labels, respectively. Lower Δp_{AAE} means that predictions made by a system are more fair according to demographic parity because there is a smaller difference in the dialect alignment of tweets labeled as harmful versus non-harmful.

The other three metrics consider the Pearson-r correlation between p_{AAE} and predicted labels. This technique is used to measure racial dialect bias in datasets [48], and other types of correlation are used to measure unfairness in other settings with continuous protected attributes [38]. For each label, we measure the overall correlation r, the correlation r_T for subsets of tweets that carry that label in the dataset, and the correlation r_F for subsets of tweets that do not carry that label. That is, for a given label ℓ,

$$r_T \equiv r(p_{AAE}(X), g(\hat{Y})) \text{ among tweets with } Y = \ell$$

$$r_F \equiv r(p_{AAE}(X), g(\hat{Y})) \text{ among tweets with } Y \neq \ell$$

where $g(\hat{Y}) = 1$ if $\hat{Y} = \ell$ and $g(\hat{Y}) = 0$ if $\hat{Y} \neq \ell$.

[1] In keeping with existing research, we assume *a priori* that the average tweet is not inherently more or less toxic in a particular dialect [48].

We define three bias metrics derived from these correlations:

$$\Delta r \equiv \overline{r_h} - \overline{r_n}$$
$$\Delta r_T \equiv \overline{r_{T,h}} - \overline{r_{T,n}}$$
$$\Delta r_F \equiv \overline{r_{F,h}} - \overline{r_{F,n}}$$

where $\overline{r_h}$ and $\overline{r_n}$ are the average values of r across all harmful and non-harmful tweets, respectively, and this convention is extended to $\overline{r_{T,h}}$, $\overline{r_{T,n}}$, $\overline{r_{F,h}}$, and $\overline{r_{F,n}}$.

Like Δp_{AAE}, Δr measures violations of parity as unconditional dependencies between p_{AAE} and label predictions. By conditioning on a given true label, Δr_T provides information on false negative classifications for that label, and by conditioning on the absence of a given true label, Δr_F provides information on false positive classifications. Therefore, Δr_T and Δr_F together tell us to what extent equal odds is upheld.

Across all four intrinsic bias evaluation metrics, positive values indicate bias against AAE, negative values indicate bias in favour of AAE, and values of zero indicate no bias.

3.2.2 Extrinsic Bias Metrics We measure bias *extrinsically* by examining patterns in predictions made on tweets in external datasets that are labeled either by dialect alignment or self-reported author race. Because Blodgett et al.'s dataset [6] is sufficiently large, we define groups based on dialect alignment in keeping with existing research: AAE-aligned if $p_{AAE} > 0.8$ and WE-aligned if $p_{WE} > 0.8$. In Preoţiuc-Pietro and Unger's dataset [46], tweets are naturally grouped by the self-reported race of their authors.

For predictions made on [6] we define the following metrics:

$$\Delta h_{AAE,WE} \equiv p(\hat{y} \text{ is harmful}|AAE) - p(\hat{y} \text{ is harmful}|WE)$$
$$\Delta h_{AAE,AD} \equiv p(\hat{y} \text{ is harmful}|AAE) - p(\hat{y} \text{ is harmful})$$

$\Delta h_{AAE,WE}$ measures the gap in proportion of harmful label predictions between AAE- and WE-aligned tweets, and $\Delta h_{AAE,AD}$ measures the gap in proportion of harmful label predictions between AAE-aligned tweets and all tweets (*AD* for "all dialects").

For predictions made on [46] we define the following metrics:

$$\Delta h_{black,white} \equiv p(\hat{y} \text{ is harmful}|black) - p(\hat{y} \text{ is harmful}|white)$$
$$\Delta h_{black,all} \equiv p(\hat{y} \text{ is harmful}|black) - p(\hat{y} \text{ is harmful})$$

These measure equivalent "gaps" in predicted harmful porportion across self-reported author race groups.

All four extrinsic evaluation metrics measure violations of parity fairness: positive values against AAE and negative values in its favour. Neither external dataset labels the tweets by type of harm, so we cannot compare predictions to real-world outcomes; therefore, we cannot evaluate equal odds fairness extrinsically.

3.3 Baseline Classifier

As a baseline model, we train a convolutional neural network (CNN) with two hidden layers, based on the observation that CNN are well-suited to this task [34] and because they are easily compatible with adversarial debiasing [59]. Hyperparameters were chosen to optimise performance within the bounds of reasonable training times, based on five-fold stratified cross-validation on Founta et al.'s dataset. This dataset was chosen because it contains four somewhat vague and overlapping labels but is large enough to train reasonably accurate models. We experimented with tweet embedding via bag-of-words vectors [2] and TFIDF, using both character and word features. Ultimately, TFIDF embeddings of 10,000 character 1-, 2-, and 3-grams were chosen. The network trains for 50 epochs using batches of 64 tweets. It uses an Adam optimiser with a decaying learning rate initially set to 0.001. The number of units per hidden layer is calculated based on an analysis by Huang et al. [29].

Dataset	Precision	Recall	F1	F1 (BK)
Founta et al.	0.74	0.76	0.75	0.81 [34]
Davidson et al.	0.87	0.87	0.87	0.91 [37]
Waseem & Hovy	0.81	0.81	0.81	0.93 [44]
Golbeck et al.	0.69	0.71	0.69	0.67 [14]

Table 2: Baseline weighted average performance per dataset

Dataset	Label	Precision	Recall	F1	F1 (BK)
F	abusive	0.74	0.69	0.71	0.89 [34]
F	hateful	0.35	0.18	0.24	0.31 [34]
D	offensive-language	0.92	0.93	0.93	-
D	hate speech	0.35	0.26	0.30	-
W&H	sexism	0.72	0.67	0.70	1.00 [44]
W&H	racism	0.71	0.69	0.70	0.71 [44]
G	harassment	0.43	0.33	0.37	-

Table 3: Performance for the harmful label(s) in each dataset. Per-label BK performance is not available for all datasets.

Table 2 shows weighted average precision, recall, and f1-score for the baseline model, evaluated in-domain on held-out test sets. Importantly, the performance is close to, though not quite as good as, the best known (BK) performance that we could find in the literature for each dataset [14, 34, 37, 44]. The classifier is reasonably able to identify harmful content in most datasets; however, like even the most discerning human annotators, it struggles to differentiate between different types of harm, and like the BK models it also performs poorly in classification of some specific harmful labels (as Table 3 shows). Recall tends to be lower than precision for harmful labels, which suggests that the baseline system under-classifies content as harmful. A weakness of weighted average f1-score as a measure of system performance is that systems can have high weighted average f1-scores while performing poorly on harmful labels as a result of the relatively low prevalence of harmful labels in all datasets except that of Davidson et al. It is nonetheless a useful way to quantify a system's performance across labels.

3.4 Bias Mitigation Pre-Processing (Preferential Sampling)

Preferential sampling [30] mitigates dataset bias by resampling datapoints with high classification uncertainty: it duplicates anti-stereotypical points and drops pro-stereotypical ones. This changes the dataset distribution to reduce discrimination while preserving much of the information for training. It has been shown to effectively reduce bias while maintaining reasonably high performance in classification settings with binary Y and S, and is less intrusive than similarly effective dataset bias mitigation techniques [30, 31].

Harmful tweet datasets often contain more class labels than simply *harmful* and *not harmful*, and differences between different types of harmful label can be important. We are tasked, then, with a *multi-class* classification problem. In the context of multi-class

Y and continuous S, a preferential sampling implementation must differ from the original algorithm [31] in three ways:

(1) The measure of uncertainty should be compatible with multiple classes.
(2) Because resampling data with more extreme (high or low) p_{AAE} is likely to have a greater impact, p_{AAE} should be a factor in considering which data objects to duplicate or drop.
(3) For the same reason, we cannot pre-calculate an optimal number of datapoints to duplicate or drop.

The first point is addressed by defining a measure of predictive uncertainty similar to margin sampling [49]. We argue that generally speaking, misclassifications of harmful tweets as non-harmful and vice versa are more consequential than misclassifications within either harmful or non-harmful label sets: classifying harmful tweets as non-harmful allows harm to go undetected and unmoderated, and classifying non-harmful tweets as harmful may stifle benign sentiments of a population or individual. Therefore, the margin of confidence $m(x)$ is defined as the margin between the highest predicted label probability and the highest predicted probability of a label of the opposite harm value, where predictions are made using a basic logistic regression classifier. This encourages the system to resample tweets that are most likely to be misclassified as harmful when they are not, and vice versa.

The second point is addressed by adding a term to a data object's resampling candidacy $C(x)$ that encapsulates the extremity of its AAE alignment. We use $|a(x)|$, the absolute value of a tweet's normalised p_{AAE} rank: $a(x) = -1$ for the tweet with lowest p_{AAE}, $a(x) = 1$ for the tweet with highest p_{AAE}, and $a(x) = 0$ for the median. Overall, then, a tweet's resampling candidacy is defined as

$$C(x) \equiv |a(x)| - w_p m(x)$$

where w_p is a hyperparameter. As a result, tweets are the strongest candidates for resampling when they both have extreme p_{AAE} and lead to a low margin of confidence between a harmful and non-harmful label prediction.

The third constraint is addressed by performing resampling iteratively. Rather than pre-calculating an ideal number of duplications and deletions, we resample N data objects at a time until bias B is reduced to below a theshold T. We define B as the average of either $|r|$ if fairness is defined as demographic parity, or $\frac{|r_T|+|r_F|}{2}$ if fairness is defined as equality of odds, across all labels in the dataset. This penalises any bias, even if the bias favours AAE alignment. Importantly, different fairness definitions impact the algorithm only to the extent that they differently determine when the threshold for termination is reached — the resampling process itself is the same.

Because there is a limit to bias that can be removed by resampling data objects this way, B eventually begins to increase if T is set too low. As a result, the algorithm terminates if at any point B is greater than or equal to its value from the previous iteration.

In this implementation, w_p and T are tunable hyperparameters. Intuitively, w_p sets the importance of predictive uncertainty relative to extremity of p_{AAE} in determining a tweet's candidacy for resampling; it is a hyperparameter because the relative importance of these factors is not immediately obvious. T determines to what extent the data are resampled: how much bias should be reduced. A standardised batch size $N = 1,000$ is used.

For each dataset and each definition of fairness, we train models using all combinations of the following hyperparameter values:

$$w_p \in \{0.1, 0.32, 1, 3.2, 10\},\ T \in \{0.05, 0.1, 0.15, 0.2\}$$

We choose values that lead to the lowest prediction bias B — defined according to the fairness definition used for resampling — without reducing weighted average f1-score by more than 5% of baseline.

We find that the values of w_p and T impact classification performance and propagated bias, but they do not do so consistently across datasets. Each value of w_p is best for at least one dataset and fairness definition, as are all values of T except 0.2, which may be too high a threshold to force substantial resampling.

3.5 Bias Mitigation In-Processing (Adversarial Debiasing)

Adversarial debiasing [67] unintrusively attempts to remove any bias, whether against or in favour of any demographic group. In-processing bias mitigation by regularisation has been adapted to the case of a continuous protected attribute [38], but adversarial debiasing appears not yet to have been explored in this case.

We adapt the adversarial debiasing code developed by IBM [3] to be compatible with multi-class classification,[2] and add a hidden layer of size 100 to the adversary network in keeping with [59].

According to the original implementation [67], the predictor network updates its prediction weights W according to

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

where L_P is its loss, L_A is the adversary network's loss, and α is a tunable hyperparameter. The magnitude of α determines the adversary's strength in debiasing classifications made by the predictor, presumably at the expense of higher L_P.

In order to enforce demographic parity the adversary network receives as input only the predictions $\hat{Y}$ during training. For equal odds enforcement the adversary receives both $\hat{Y}$ and true labels Y.

For each dataset, a model with adversarial debiasing was trained using each fairness definition, and with the following values of α:

$$\alpha \in \{0.1, 0.32, 1, 3.2, 10, 32, 100\}$$

We use the same criteria as with w_p and T to choose α in each case. Generally, higher values of α lead to greater bias reduction: all chosen values are either 10, 32, or 100.

Finally, we combine preferential sampling and adversarial debiasing by training models using adversarial debiasing on resampled data. To select hyperparameter values w_p, T, and α for these models, we explore all combinations of the three best sets of w_p and T and the three best values of α. It is rarely the case that the best values for both preferential sampling and adversarial debiasing alone are also best when the two are combined, and we find substantial variation in which values are most effective between datasets.

4 EXPERIMENTAL RESULTS

Using each bias mitigation approach independently and combined, and using both parity and equal odds definitions of fairness, we train seven types of models, using the following notation:

- B — baseline model
- Pr (p) — preferential sampling, parity fairness

[2] Our code is available at https://github.com/arb7/adv-deb-multi.

- Pr (e) — preferential sampling, equal odds fairness
- Adv (p) — adversarial debiasing, parity
- Adv (e) — adversarial debiasing, equal odds
- Pr+Adv (p) — both preferential sampling and adversarial debiasing, parity
- Pr+Adv (e) — both preferential sampling and adversarial debiasing, equal odds

For each model type and training dataset, five individual models were trained. All reported results in this section represent averages across these sets of five models, as this smooths the variation due to random variable initialisation in the neural network. That said, we find variation is generally low between models. We evaluate models based on their performance in-domain and cross-domain, and on the bias they propagate measured intrinsically and extrinsically. Performance and intrinsic bias evaluations were performed using a single University of Cambridge Computer Lab GPU machine, and the extrinsic using the University of Cambridge Research Computing Services Wilkes2 GPU cluster.[3]

4.1 In-Domain Classification

Table 4 shows the weighted average f1-score of each model, evaluated in-domain on the held-out test subset of the dataset on which it was trained. Bias-mitigated models tend to perform nearly as well in-domain as baseline for Founta et al. and Davidson et al.'s datasets. On these datasets, we see that adversarial debiasing yields slightly higher classification performance than preferential sampling or a combination of approaches. Bias-mitigated f1-score is roughly equal to baseline for Waseem & Hovy and Golbeck et al.'s less biased datasets.

	F	D	W&H	G
B	**0.75**	**0.87**	**0.81**	0.69
Pr (p)	0.73	0.83	**0.81**	0.69
Pr (e)	0.74	0.86	**0.81**	0.69
Adv (p)	0.74	0.86	**0.81**	**0.70**
Adv (e)	**0.75**	0.86	**0.81**	0.69
Pr+Adv (p)	0.73	0.84	**0.81**	**0.70**
Pr+Adv (e)	0.73	0.85	**0.81**	0.69

Table 4: Weighted average f1-score for baseline and bias mitigated models, evaluated in-domain. The highest score for each dataset appears in bold.

4.2 Cross-Domain Classification

One measure of a system's generalisability is its cross-domain classification performance [64]. In this case, it is impossible to perform multi-class classification because each dataset uses different labels. Therefore, in keeping with previous research [64] we restrict cross-domain classification to a binary task between *harmful* and *non-harmful* tweets as defined in Table 1.

For each model, we predict labels for the entire (training plus test subsets) datasets on which the model was not trained. We calculate weighted average f1-score across harmful and non-harmful labels. Because each model can be evaluated on each of the three datasets on which it was not trained, twelve dataset permutations exist in total. As before, the reported f1-score for each dataset permutation represents an average over five models. Figure 2 shows these results (full tabular data appear in the Appendix, §A.2).

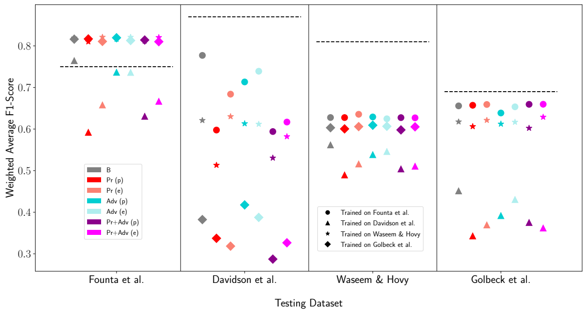

Figure 2: Weighted average f1-score organised by testing dataset. Marker shape indicates training dataset, and colour indicates bias mitigation approach. Dashed lines indicate baseline *in-domain* f1-score for each (testing) dataset.

4.2.1 Performance Suffers Cross-Domain In line with previous findings [64], models perform worse cross-domain than in-domain. This suggests that dataset bias — including but not restricted to dialect bias or other discriminatory biases — may cause harmful tweet detection systems to perform worse in the real world than in their development contexts.

4.2.2 Differences Between Datasets As Figure 2 shows, models trained on Founta et al.'s data perform best when evaluated on other datasets, followed closely by Waseem & Hovy, then Golbeck et al., and then Davidson et al. Given that the best and worst performing datasets are both the largest and the most biased against AAE (per Table 1), we observe that neither dataset size nor dataset bias relates clearly to cross-domain classification performance.

We observe that the baseline models often perform better than bias-mitigated models; however, these performance differences are not very consistent. At least one bias-mitigated model performs at least as well as baseline in a majority of dataset permutations, including all six permutations that train on Waseem & Hovy or Golbeck et al.'s datasets. Generally, adversarial debiasing yields higher cross-domain performance than preferential sampling or a combination of the two, but training on Waseem & Hovy's dataset provides a notable exception. Similarly, equal odds generally yields higher performance than parity, but this result is also inconsistent.

The level of dialect bias present in a dataset appears to impact the extent to which bias mitigation affects cross-domain performance. The variance in f1-score, and the decrease in performance when bias mitigation is applied, are greatest when the two more biased datasets, produced by Founta et al. and Davidson et al., are used for both training and testing. The most consistent f1-scores occur when the less biased datasets, produced by Waseem & Hovy and Golbeck et al., are used for both training and testing. Combinations of the two fall in the middle, though classification performance on Founta et al.'s data is preserved through bias mitigation when models are trained on one of the less biased datasets. This again underscores the importance of the data collection and annotation.

4.3 Intrinsic Bias Evaluation

Figure 3 compares bias metrics for predictions made by models in-domain on each dataset. Each graph represents a dataset. Within each graph, the two left bar groups show Δp_{AAE} and Δr, which measure violations of demographic parity. The two right bar groups

[3] See https://www.hpc.cam.ac.uk/systems/wilkes-2 for technical specifications.

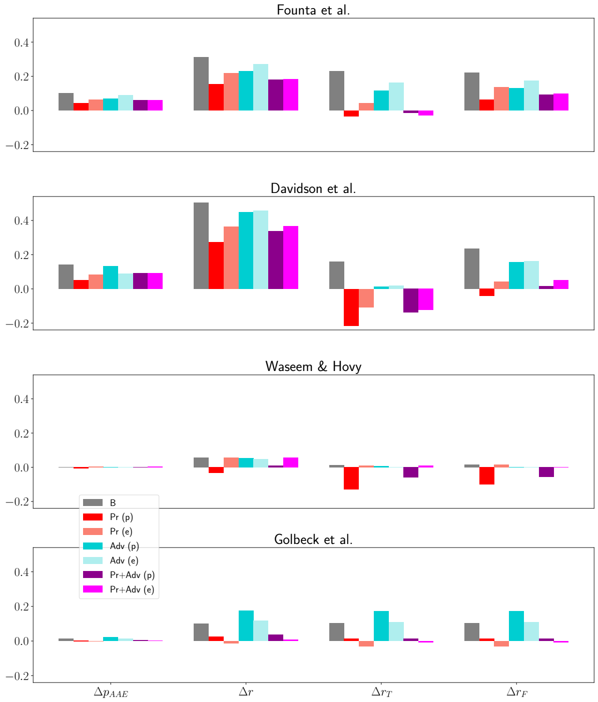

Figure 3: Intrinsic bias results from in-domain classification. From left to right: Δp_{AAE}, Δr, Δr_T, and Δr_F.

show Δr_T and Δr_F, which measure violations of equal odds. As described in §3.2, higher bars indicate higher bias against AAE.

For each dataset there exists some bias mitigation approach that substantially reduces bias against AAE at a small or nonexistent performance cost. However, the reduction in bias is not perfect.

4.3.1 Differences Between Datasets In most cases, we observe that preferential sampling decreases bias, particularly Δp_{AAE} and Δr, to a greater extent than adversarial debiasing. That said, no single approach is consistently best for reducing all types of bias while maintaining high performance, and the success of different approaches varies across datasets. For instance, on Davidson et al. and Waseem and Hovy's data, preferential sampling with fairness defined by parity creates negative Δr_T with a higher magnitude than baseline, which suggests a greater violation of equal odds, but with misclassifications *favouring* AAE: tweets with high p_{AAE} are more likely to be misclassified as benign and less likely to be misclassified as harmful. For these datasets, adversarial debiasing with a parity definition of fairness reasonably reduces bias without reversing it. Yet on Founta et al. and Golbeck et al.'s data, preferential sampling leads to less bias of all types, and in the case of Golbeck et al. adversarial debiasing exacerbates bias above baseline.

These differences indicate variation in the types and nature of bias present. For example, each dataset uses a different sampling method (before and distinct from our resampling) to increase the prevalence of harmful content. Sampling biases arising from these different methods may be differentially amenable to "correction" by resampling. Meanwhile, adversarial debiasing targets propagated algorithmic bias. Because social, label, and sampling biases each affect bias propagation differently, and because they are each present to different extents in these datasets, it seems natural that the impact of adversarial debiasing is varied.

4.3.2 Tension Between Fairness Types Across datasets, those bias mitigation approaches that best improve demographic parity fairness occasionally do so at the expense of equal odds. This phenomenon appears strongest when preferential sampling is employed, whether in combination with adversarial debiasing or not. On all datasets except Golbeck et al., training models on resampled data using a parity definition of fairness produces negative values for Δr_T, Δr_F, or both. This is, unsurprisingly, usually accompanied by the largest decreases in predictive performance.

In fact, Figure 3 does not even tell the whole story: in some cases where Δr_T and Δr_F are small, individual harmful labels have high-magnitude negative correlations with p_{AAE}. This indicates bias in favour of AAE for those labels, but can also hide substantial bias *against* AAE for other labels in secondary bias metrics. An advantage of adversarial debiasing over preferential sampling is that the tradeoff between parity and equal odds fairness is less pronounced, making it easier to better balance the two. This makes sense given the mechanisms by which the two approaches work: preferential sampling seeks always to advantage AAE-aligned tweets in its duplications and removals, whereas adversarial debiasing works to minimise any relationship between p_{AAE} and predictions.

It is important for those who implement and deploy harmful content detection systems to consider the extent to which they wish to enforce different types of fairness. In some cases the solution might be straightforward. For example, preferential sampling of Golbeck et al.'s dataset enforces both parity and equal odds in models: each of Δp_{AAE}, Δr, Δr_T, and Δr_F is nearly zero. However, the data show that generally, some degree of either parity or equal odds fairness must be sacrificed in order to optimise the other. This is consistent with the observation [33] that parity and equal odds are incompatible given different base rates. That is, if S is dependent of Y, it is impossible for $\hat{Y}$ to be both unconditionally independent of S and conditionally independent of S given Y. In all but very equal datasets, differences in base rates may make it impossible to enforce a satisfactory degree of both parity and equal odds.

4.4 Extrinsic Bias Evaluation

Figure 4 compares bias metrics for predictions made by models on external datasets labeled by either dialect alignment [6] or self-reported author race [46]. Each graph represents a training dataset. Within each graph, the two left bar groups show $\Delta h_{AAE,WE}$ and $\Delta h_{AAE,AD}$, which measure gaps in the proportions of tweets predicted to be harmful along dialect lines. The two right bar groups show $\Delta h_{black,white}$ and $\Delta h_{black,all}$, which measure gaps in the proportions of tweets predicted to be harmful along racial lines. Recall that all of these metrics measure bias only as violations of parity because tweets in the external datasets are not labeled according to harm. Once again, higher bars indicate higher bias against AAE.

As with intrinsic bias, for each dataset there is some form of bias mitigation that appears capable of reducing extrinsic bias somewhat. In some cases, there is tension between the ability of bias mitigation approaches to promote equality along dialect lines versus along author race lines, but this is not a consistent trend.

4.4.1 Differences Between Datasets Consistent with the intrinsic evaluation, Figure 4 shows that not only are the baseline bias levels

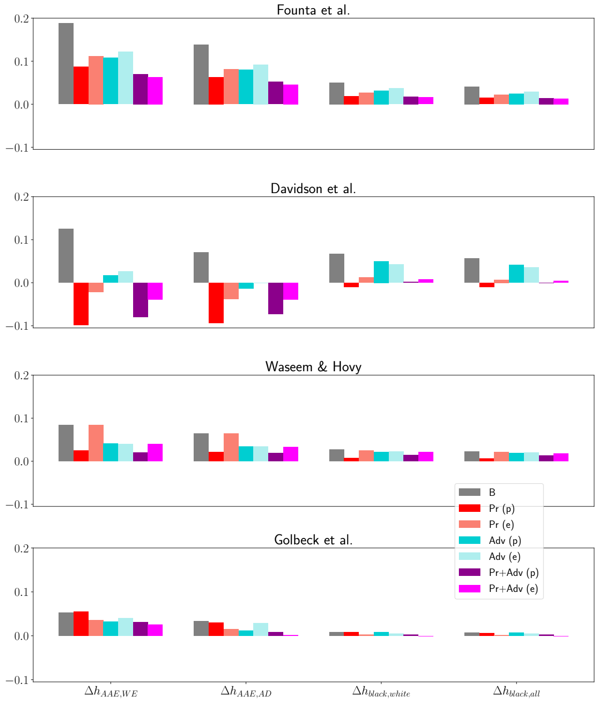

Figure 4: Extrinsic bias evaluation results. From left to right: $\Delta h_{AAE,WE}$, $\Delta h_{AAE,AD}$, $\Delta h_{black,white}$, and $\Delta h_{black,all}$.

different across datasets, but also that bias mitigation has different effects in different contexts. Individual approaches substantially reduce bias for some datasets, but are ineffective or create additional biases for other datasets. For instance, using a combination of preferential sampling and adversarial debiasing effectively reduces bias against both AAE-aligned tweets and tweets written by Black authors in the datasets produced by Founta et al., Waseem & Hovy, and Golbeck et al. However, while this approach reduces bias against tweets written by Black authors to near zero in the two Davidson et al. datasets, it creates substantial negative $\Delta h_{AAE,WE}$ and $\Delta h_{AAE,AD}$, which suggests new imbalance in favour of AAE. Once again, no one approach is consistently superior.

This contextual dependence suggests that the link between experimental results and real-world behaviour may be more tenuous than one would hope: context clearly matters, but its impact is a much more complex problem. So while there is reason to be optimistic that bias mitigation strategies can reduce differential moderation of online content, we must aim to better understand these relationships before and as we deploy such technologies.

5 DISCUSSION

Our results indicate that both preferential sampling and adversarial debiasing can substantially reduce — though not completely eliminate — bias against AAE in the task of harmful tweet detection, at little to no performance cost. However, bias and performance impacts vary between datasets in ways that are not always straightforward. This complex context dependence, and performance inconsistencies in different settings, raise questions about the fairness of automated content moderation systems and their use.

5.1 Bias Rooted in Datasets

The context dependence of bias and performance responses to bias mitigation suggests the importance of the entire data collection, annotation, and processing pipeline. Classification performance may be improved by filtering and boosting training data to increase the prevalence of harmful tweets, and by using scalable annotation methods to allow for larger datasets. However, our results suggest that may be at the cost of more severe dialect bias.

Rather than identifying a single best approach to mitigating bias against AAE in datasets and models, our results further suggest that those who implement and deploy harmful content detection systems are best placed to undertake bias analyses. They can then explore options for minimising dialect bias that align with their specific contexts, goals, and values (and can be held to account based on these). Other bias mitigation techniques may be more effective than those we implemented and are worth exploring, especially when the bias is introduced by human labellers who may not have a nuanced understanding of subconscious biases against AAE. For instance, multi-task learning has been shown [56] to substantially reduce marginalised identity bias in harmful online comment detection. We also recommend that further research continue to seek a deeper understanding of the sources and nature of racial dialect bias in this and other NLP tasks.

Importantly, racial dialect bias is not the only form of bias that impacts harmful content detection systems. Research [17, 42] has exposed bias that discriminates against marginalised identity mentions, and other forms of bias such as topic and author bias [64] can hurt system accuracy metrics. A deeper understanding of the range of bias that can exist might enable more complete mitigation of biases, including those not yet identified.

5.2 Towards "True" Fairness

5.2.1 Competing Definitions of Fairness The task of building a "fair" harm detection model is made challenging by biased datasets and by conflicting perspectives on how to define fairness. In this paper, we have proposed equal odds: systems should not systematically misclassify tweets as more harmful the more they align with AAE and less harmful the less they do. However, bias and ambiguity in the datasets limit our ability to calculate equal odds. In contrast with many other domains, such as true recidivism or loan default rates, underlying "objective truth" class labels are unknown in the harmful tweet context for a variety of reasons including vague and overlapping class definitions, annotation bias, and the subjectivity of harmful content. Therefore, calculated false positive and negative rates may themselves have issues of bias, which weakens evaluations of equal odds fairness. For this reason, it may be valuable to also consider demographic parity fairness, which enforces the same proportion of classification as harmful across groups, even though this may be problematic if there are true differences in distributions.

What is the most appropriate metric of fairness depends on the context. On one hand, freedom of expression is at stake, with the risk of disproportionately silencing an already-marginalised group. On the other hand, there is a risk that truly harmful content could go undetected. Efforts towards greater fairness in correctly classifying AAE tweets should take seriously the need to prevent harmful content from reaching too wide an audience, and any tensions or trade-offs between these risks must be considered. Ultimately, our aspiration toward fairness is grounded in a belief that harmful content detection systems, and AI and technology more generally, should reflect, advance, and support a society to which we aspire, even if that means a slight dissonance with society's current state.

5.2.2 *Bias-aware Use*

The harms averted by automated harmful tweet detection systems, and those caused by their biases, ultimately depend on the way they are deployed. For instance, at present automated systems are used to screen tweets for review by human moderators [9]. *Dialect priming* has been shown [48] to reduce dialect bias at the point of dataset annotation; similarly, priming human moderators for dialectal differences might reduce the extent to which Black voices are silenced online. Platforms could choose to never remove content outright based on automated predictions, but rather to hide it behind interstitial warnings. While this would not fully *silence* communities per se, it might perpetuate societal biases by priming readers to expect harm when they encounter AAE. It is critical that the merits and shortcomings of these systems be evaluated and debated not in isolation, but in relation to the ways that they are used and impact people.

5.3 Implications of Uncertainty

The often-tenuous connections between in-domain and cross-domain performance, and between intrinsic and extrinsic bias, demonstrate the difficulty of predicting real-world impact based on experiments performed only in limited research and development contexts. The bias mitigation approaches explored seem capable of reducing bias along some axes while maintaining high in-domain classification performance, but this cannot be guaranteed in the complex real-world. A certain level of uncertainty inevitably accompanies applying systems outside of their development settings.

This uncertainty raises the question of what role such technologies — which for better or worse impact our social and political systems worldwide — can and should play in society. It is clear that harmful tweet detection systems are far from perfect, and that there is no silver bullet to solve their problems. However, these systems can serve an important role protecting people from online abuse and hate. Further, human decisions are also afflicted by bias, which is reflected in the label bias we observe in datasets. Should AI systems be held to a higher standard than their human counterparts should we choose to deploy them? Although structural issues afflict both humans and autonomous systems, individual human bias is just that, individual, whereas algorithmic bias can have systemic effects by crystallising and reperpetuating the bias at-scale.

It has been argued [45] that we ought to answer important questions about the role and scope of technological interventions before they are implemented, and that the reformist nature of bias mitigation research can distract from these deeper issues. However, content classification and moderation systems are already in place. It is important to simultaneously interrogate their role *and* make them as fair as possible. In fact, as we have demonstrated, attempting to mitigate bias using technical interventions can shed light on new facets of the problem, or at the very least reveal how incomplete our understanding is.

5.4 Social Media's Political Economy

More fundamentally, there are questions of whether it is possible to produce "fair" systems for commercial platforms in societies that are systemically unfair. Other questions are raised by the commercial nature of platforms such as Twitter. These platforms are now the technical infrastructure on which parts of society rely, but are also sites of power, control, and profit [10–12, 69]. Research on improving systems that produces more effective control of social infrastructure by platforms — by, for example, improving their moderation systems — may contribute to increasing those companies' ability to influence communications according to their commercial priorities [10] while at the same time offloading some responsibility for, and cost of, getting their systems right themselves. Moreover, concerns of platformised "predatory" inclusion are also raised, whereby greater inclusiveness ultimately works to increase marginalised groups' exposure to forms of control and revenue extraction rather than addressing structural disadvantage [51]

This is not to say that such research is not important; on the contrary, ensuring that Black people and others can access and use social platforms without being subject to harm *or* discriminatory moderation is a societal imperative. But, for platforms to be truly inclusive of marginalised or minoritised communities, research also needs to address the business models and structural features of those platforms — such as their design and affordances — that can contribute to the prevalence of hate speech in the first place.

6 CONCLUSION

In this paper we have explored mitigating racial dialect bias in a neural network for harmful tweet detection by adapting two approaches: preferential sampling pre-processing and adversarial debiasing in-processing. These techniques tend to reduce systematic bias against AAE, measured both intrinsically and extrinsically, while maintaining a high degree of performance for in-domain prediction. However, we observe the extent to which bias and performance are impacted by our interventions is extremely dependent on dataset context. A cross-domain performance evaluation further reveals the differences in the behaviour of harmful tweet detection systems within and outside of their training contexts.

These unavoidable uncertainties raise important questions regarding the role of automated harmful content detection and other AI technologies in society. There is value in attempting to mitigate bias; however, the inconsistencies and shortcomings of our bias mitigation strategies indicate how complex these biases can be.

This research is inherently limited in that it attempts to address a social problem — though admittedly one that has been exacerbated by technology — through purely technical means. Quantitative representations of bias can illuminate and mitigate critical and unforeseen challenges, and computational interventions can relieve their symptoms. However, this work represents only a starting point. We hope that our research will promote a continued conversation on societal and personal biases, fair AI, technology's political economy, and the broader role and risks of technology.

ACKNOWLEDGMENTS

Our group acknowledges the financial support of the UK EPSRC, Aviva, Alan Turing Institute (Strategic Priorities Fund (TPS)), and Microsoft through the Microsoft Cloud Computing Research Centre. We also acknowledge the compute resources provided by the Cambridge Service for Data Driven Discovery (CSD3).

REFERENCES

[1] Douglas G Altman and Patrick Royston. 2006. The cost of dichotomising continuous variables. BMJ 332, 7549 (2006), 1080.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion). International World Wide Web Conferences Steering Committee, 759–760.

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In Proceedings of the Conference on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017).

[5] Hannah Bloch-Wehba. 2019. Automation in Moderation. Cornell International Law Journal, Forthcoming (2019).

[6] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Association for Computational Linguistics, 1119–1130.

[7] Eszter Bokányi, Dániel Kondor, László Dobos, Tamás Sebők, József Stéger, István Csabai, and Gábor Vattay. 2016. Race, Religion and the City: Twitter Word Frequency Patterns Reveal Dominant Demographic Dimensions in the United States. Palgrave Communications 2 (2016).

[8] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS '16). Curran Associates Inc., 4356–4364.

[9] Robyn Caplan. 2018. Content or context moderation. Data & Society Research Institute (2018).

[10] Jennifer Cobbe. 2020. Algorithmic Censorship by Social Platforms: Power and Resistance. Philosophy & Technology (2020).

[11] Jennifer Cobbe and Jatinder Singh. 2019. Regulating Recommending: Motivations, Considerations, and Principles. European Journal of Law and Technology, 10(3) (2019).

[12] Nicole Cohen. 2011. The Valorization of Surveillance: Towards a Political Economy of Facebook. Democratic Communiqué 22 (2011).

[13] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. Big Data 5 (2017), 120–134.

[14] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online. 25–35.

[15] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017. AAAI Press, 512–515.

[16] M. Di Paolo and A.K. Spears. 2014. Languages and Dialects in the U.S.: Focus on Diversity and Linguistics. Taylor & Francis.

[17] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In AAAI/ACM Conference on AI, Ethics, and Society. Association for the Advancement of Artificial Intelligence.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (Cambridge, Massachusetts) (ITCS '12). Association for Computing Machinery, 214–226.

[19] Elizabeth Dwoskin, Jeanne Whalen, and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web - and suffer silently. The Washington Post. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.

[20] UK Department for Digital, Culture, Media & Sport. 2018. Data Ethics Framework. Government Guideline.

[21] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM 2018). 491–500.

[22] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. arXiv:1809.08651 [cs.CL]

[23] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A Large Labeled Corpus for Online Harassment Research. In Proceedings of the 2017 ACM on Web Science Conference (WebSci '17). Association for Computing Machinery, 229–233.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS '14). MIT Press, 2672–2680.

[25] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7, 1 (2020).

[26] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., 3323–3331.

[27] High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI. Report. European Commission.

[28] Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 591–598.

[29] Guang-Bin Huang. 2003. Learning Capability and Storage Capacity of Two-Hidden-Layer Feedforward Networks. Trans. Neur. Netw. 14, 2 (2003), 274–281.

[30] F. Kamiran and T.G.K. Calders. 2010. Classification with no discrimination by preferential sampling. In Informal proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'10, Leuven, Belgium, May 27-28, 2010). 1–6.

[31] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. Knowl. Inf. Syst. 33, 1 (2012), 1–33.

[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD '12). 35–50.

[33] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23.

[34] Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative Studies of Detecting Abusive Language on Twitter. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics, 101–106.

[35] Yi Li and Nuno Vasconcelos. 2019. REPAIR: Removing Representation Bias by Dataset Resampling. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[36] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The Variational Fair Autoencoder. In Proceedings of the 4th International Conference on Learning Representations (ICLR 2016).

[37] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. PLOS ONE 14, 8 (2019), 1–16.

[38] Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. 2019. Fairness-Aware Learning for Continuous Attributes and Treatments. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97). PMLR, 4382–4391.

[39] Louise Matsakis and Paris Martineau. 2020. Coronavirus Disrupts Social Media's First Line of Defense. Wired. https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/.

[40] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[41] UK Home Office and UK Department for Digital, Culture, Media & Sport. 2020. Online Harms White Paper. Government Guideline.

[42] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2799–2804.

[43] Frank Pasquale. 2019. The Second Wave of Algorithmic Accountability. LPE Blog. https://lpeblog.org/2019/11/25/the-second-wave-of-algorithmic-accountability/.

[44] Georgios Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting Offensive Language in Tweets Using Deep Learning. (2018).

[45] Julia Powles and Helen Nissenbaum. 2018. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. Medium OneZero. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

[46] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In Proceedings of the 27th International Conference on Computational Linguistics (COLING). Association for Computational Linguistics, 1534–1545.

[47] The Associated Press. 2020. Solidarity with U.S. protesters: People around the world march and speak out against racism. Canadian Broadcasting Corporation. https://www.cbc.ca/news/world/protests-world-floyd-1.5595135.

[48] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 1668–1678.

[49] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active Hidden Markov Models for Information Extraction. In Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA '01). Springer-Verlag, 309–318.

[50] 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. Association for Computational Linguistics, 1–10.

[51] Louise Seamster and Raphaël Charron-Chénier. 2017. Predatory Inclusion and Education Debt: Rethinking the Racial Wealth Gap. Social Currents 4, 3 (2017).

[52] Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2019. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. arXiv:1912.11078 [cs.CL]

[53] H. Suresh and J. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. ArXiv abs/1901.10002 (2019).

[54] Twitter. [n.d.]. Our range of enforcement options. https://help.twitter.com/en/rules-and-policies/enforcement-options.

[55] Twitter. [n.d.]. The Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules.

[56] Ameya Vaidya, Feng Mai, and Yue Ning. 2019. Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection. arXiv:1909.09758 [cs.AI]

[57] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In Proceedings of the International Workshop on Software Fairness (FairWare '18). Association for Computing Machinery, 1–7.

[58] Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data: Garbage In, Garbage Out. arXiv:2004.01670 [cs.CL]

[59] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. In Proceedings of the Conference on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018).

[60] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science. Association for Computational Linguistics, 138–142.

[61] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics, 78–84.

[62] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics, 88–93.

[63] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Technical Report. Nuffield Foundation.

[64] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 602–608.

[65] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12). Association for Computing Machinery, 1980–1984.

[66] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML '13). JMLR.org, III–325–III–333.

[67] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, 335–340.

[68] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, 15–20.

[69] Shoshana Zuboff. 2019. Surveillance Capitalism and the Challenge of Collective Action. New Labor Forum 28, 1 (2019), 10–29.

A APPENDIX

This appendix presents support for our treatment of p_{AAE} as a continuous variable, and shows full tabular data for the cross-domain performance evaluation.

A.1 Continuous Dialect Alignment

Figure 5 shows normalised count histograms of p_{AAE} distributions for tweets written by authors who self-identify as Black and non-Black in a user-level race dataset [46]. Relative to non-Black authors, Black authors wrote fewer tweets with low (below about 0.2) p_{AAE} and more tweets with high p_{AAE}. Welch's t-test yields $p < 0.001$ for these two distributions: the expected value of p_{AAE} is significantly higher for Black authors. Furthermore, the Pearson-r correlation between p_{AAE} and Black authors in the dataset is positive and significant ($r = 0.205$, $p < 0.001$). Of course, dialect does not correspond perfectly to race; however, this supports the use of p_{AAE} as a continuous protected attribute because the difference persists across its entire range, not only above a threshold value.

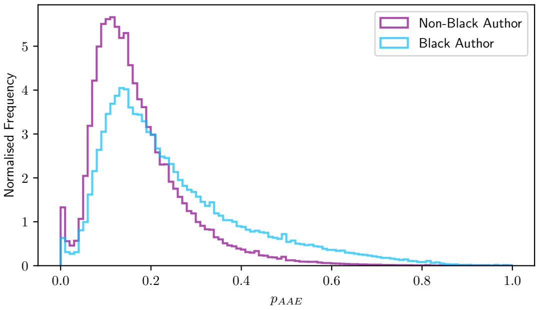

Figure 5: Normalised p_{AAE} distributions for tweets written by authors who do and do not identify as Black.

A.2 Full Cross-Domain Performance Evaluation Data

		Testing Dataset			
Train.		F	D	W&H	G
F	B	-	**0.78**	0.63	**0.66**
	Pr (p)	-	0.60	0.63	**0.66**
	Pr (e)	-	0.68	**0.64**	**0.66**
	Adv (p)	-	0.71	0.63	0.64
	Adv (e)	-	0.74	0.62	0.65
	Pr+Adv (p)	-	0.59	0.63	**0.66**
	Pr+Adv (e)	-	0.62	0.63	**0.66**
D	B	**0.76**	-	**0.56**	**0.45**
	Pr (p)	0.59	-	0.49	0.34
	Pr (e)	0.66	-	0.52	0.37
	Adv (p)	0.74	-	0.54	0.39
	Adv (e)	0.74	-	0.55	0.43
	Pr+Adv (p)	0.63	-	0.50	0.38
	Pr+Adv (e)	0.67	-	0.51	0.36
W&H	B	**0.82**	0.62	-	0.62
	Pr (p)	0.81	0.51	-	0.61
	Pr (e)	**0.82**	**0.63**	-	0.62
	Adv (p)	**0.82**	0.61	-	0.61
	Adv (e)	**0.82**	0.61	-	0.62
	Pr+Adv (p)	0.81	0.53	-	0.60
	Pr+Adv (e)	**0.82**	0.58	-	**0.63**
G	B	**0.82**	0.38	0.60	-
	Pr (p)	**0.82**	0.34	0.60	-
	Pr (e)	0.81	0.32	**0.61**	-
	Adv (p)	**0.82**	**0.42**	**0.61**	-
	Adv (e)	0.81	0.39	**0.61**	-
	Pr+Adv (p)	0.81	0.29	0.60	-
	Pr+Adv (e)	0.81	0.33	**0.61**	-

Table 5: Weighted average f1-score (binary classification) for baseline and bias mitigated models, evaluated cross-domain. The highest score for each dataset permutation in bold.