MITIGATING RACIAL BIASES IN TOXIC LANGUAGE DETECTION

A Thesis
Presented to
The Academic Faculty

By

Matan Halevy

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science
School of Interactive Computing
Department of Computer Science

Georgia Institute of Technology

December 2021

MITIGATING RACIAL BIASES IN TOXIC LANGUAGE DETECTION

Thesis committee:

Dr. Ayanna Howard, Advisor
College of Engineering
The Ohio State University

Dr. Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology

Dr. Diyi Yang
School of Interactive Computing
Georgia Institute of Technology

Date approved: December 10, 2021

To my grandparents whose bravery, resilience, and love for knowledge and humanity inspire me everyday.

ACKNOWLEDGMENTS

I would like to begin by thanking the members of my thesis committee for all the support and assistance in the preparation of this thesis. First, thank you to Dr. Ayanna Howard, who agreed to supervise an online master's student she never met and provided me with the opportunity to pursue my goal of publishing a paper and participating in academic research. The advice, support, and knowledge you have shared with me has helped me grow as a person, engineer, and a researcher - thank you for the countless lessons you have provided. Thank you Dr. Amy Bruckman and Dr. Diyi Yang who spent hours of their time to guide this research, advise my work, and provide the help to publish our work. Thank you as well to Camille Harris whose collaboration in our research provided insights and knowledge essential to this work's completion.

I would like to thank my parents, Yoram and Anat Halevy, whose endless love and support has provided me the freedom to pursue my own interests and goals. Your love for knowledge, compassion in humanity, and the strength you have displayed throughout my life is a source of many traits that I am thankful to have inherited from you two. To my older sister Lotem, thank you for always pushing me to be the best version of myself and providing the peace of mind of knowing that I always have someone who has my back. Thank you to my younger sister Netta, for always providing the love and motivation I need to overcome obstacles and challenges I face.

Thank you to my coworkers at Hopper, who provided mentorship, opportunities to grow as a leader and engineer, and flexibility as I was working through the OMSCS program.

I would like to thank the many friends who have been my cheerleaders throughout this process, whether it is understanding when I bail to study or listening to me complain about being stressed. Thank you especially to Emmy Egulu and Abu Kamat, who spent their time proof-reading my paper and encouraging me as I pursued this research.

Thank you to the members of HumAnS Lab who welcomed me into their lab group and

were always willing to offer advice and feedback regarding this research. Also thank you to the many people who made OMSCS possible, without this program the opportunity of accessing such high quality education may not have been possible for me. This program taught me many lessons that I will be carrying with me for the rest of my life.

I would also like to thank the anonymous ACM EAAMO reviewers whose helpful feedback allowed this work to reach its current state. Thank to Cisco Systems, Inc. who provided the grant and discussions that allowed our research group to pursue this topic.

Lastly, thank you Solly for being the best dog anyone can ask for and keeping me company on late nights studying.

SUMMARY

Recent research has demonstrated how racial biases against users who write African American English exists in popular toxic language datasets. While previous work has focused on a single fairness criteria, we propose to use additional descriptive fairness metrics to better understand the source of these biases. We demonstrate that different benchmark classifiers, as well as two in-process bias-remediation techniques, propagate racial biases even in a larger corpus. We then propose a novel ensemble-framework that uses a specialized classifier that is fine-tuned to the African American English dialect. We show that our proposed framework substantially reduces the racial biases that the model learns from these datasets. We demonstrate how the ensemble framework improves fairness metrics across all sample datasets with minimal impact on the classification performance, and provide empirical evidence to its ability to unlearn the annotation biases towards authors who use African American English.

** Please note that this work may contain examples of offensive words and phrases.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

In response to the rise of hateful and toxic language common in online communities, its detection has become a growing field of interest for both researchers and industry professionals [1]. Social media companies have increased their automated moderation efforts in order to promote healthier discourse and reduce toxicity [2]. However, the problem space is riddled with challenges and human biases that pose many open questions in both classifier performance and practical applications. Issues such as class imbalances, label subjectivity, and annotation biases can cause these algorithmic models to encompass and propagate human biases against the very minority groups they are designed to protect [3, 4, 5].

The challenge of using machine learning systems to automate hateful language detection can be traced to a high-degree of subjectivity in the human-labeled datasets on which the algorithms are trained. These datasets rely on annotators' familiarity with cultural and historical contexts and ever-changing societal forms of bigotry [1]. Sap et al. [3] documented the existence of annotation bias against users of African American English (AAE) in commonly used toxic language datasets. As a well-studied English dialect, AAE exhibits distinct grammatical rules and syntax and can serve as a proxy for racial identity when a user's race is not reported. The potential of falsely moderating AAE users' speech comes at a time of increased online racial harassment towards African Americans [2].

In Chapter 2, we review existing literature, covering previous research that explores topics of AI biases, African American English, hate speech, and the associated biases in toxic language detection. We further discuss the existing bias mitigation strategies for reduction of racial biases in toxic language detection and note that they are measured using false-positive rate (FPR) - the probability of classifying non-toxic samples as toxic conditional on the samples being non-toxic as the only criteria for correcting models' biases

against AAE authors. We observe that AAE samples in popular toxic-language datasets are mainly annotated as toxic, leaving a very small sample of AAE instances that are true-negatives (non-toxic). Hence, a fairness criterion based solely on FPR has a very limited scope.

Based on this preliminary research, we employ more descriptive fairness metrics, described in Subsection 4.1.3, in addition to FPR, to evaluate how annotation biases against AAE authors propagate through hate-detection models. We then perform experiments on commonly used models and examine the results from two bias-mitigation strategies described in Subsection 4.1.1 that have been proposed to reduce algorithmic bias towards group-identifiers and between group disparities using FPR [6, 7]. We demonstrate how all models continue to propagate and encompass racial biases from four toxic-language datasets that we describe in Section 3.1, in spite of the proposed bias-mitigation techniques.

To address this issue, we propose an ensemble model architecture (Subsection 4.1.2) that uses a general toxic language classifier coupled with a specialized AAE classifier. We demonstrated in Chapter 5 the results that show that this framework reduces the effects of annotation biases towards AAE users without impacting classifier performance. We then conduct error analysis on misclassified AAE tweets from this framework (Section 5.2) to better understand further challenges in debiasing and classifying AAE instances for toxic language detection.

This work heavily borrows from our research paper that explored the same topic which was published at ACM EAAMO 2021 [8].

CHAPTER 2
LITERATURE REVIEW

As machine learning systems are integrated as a standard tool of society, more researchers have begun to document and demonstrate how these systems learn and amplify biases from the data. Hate speech and toxic language classification has become a growing subfield of natural language processing that has been receiving attention for the fairness challenges it poses. In this chapter, we discuss previous research that demonstrates how these AI systems encompass harmful biases. We also explore computational hate speech detection and the challenges that exist around that problem space. Lastly, we discuss bias mitigation techniques that have been used in machine learning, specifically natural language processing and hate speech detection domains.

In this work we make use of a few terms that can have varied definitions. For the purpose of our work we define these terms explicitly here. We use the generalized understanding of *Artificial Intelligence* (AI) to refer to algorithms and systems that learn and act on their learning [9]. This use of AI encompasses methods referring to computational statistics, machine learning, deep learning, etc.

Bias is often used interchangeably with terms such as stereotypes, prejudice, implicit, or subconsciously held beliefs. While biases are often referred to in negative contexts, positive biases also exist that refer to "favoritism". However, in this work we largely focus on the negative biases that are tied to unjust discrimination. We define bias to be the influence in a decision-making process that prevents an objective consideration of an issue, decision, or situation [10]. Bias includes a taxonomy of features and sources that results in favoring or discriminating behaviors between things, people, or groups. We note that bias does not have to be judgements a person makes on other people but can include systems and practices that make decisions.

Lastly, in our work we utilize *Toxic Language* as a hypernym that includes hateful, abusive, or offensive language. Hate speech, identity based attacks, online bullying, trolling, threats of violence, and sexual harassment are all examples of language that is defined as toxic language [11].

2.1 AI Biases

Bias is a human behavior that guides decision processes in our daily life. When these biases emerge in the data that represents these human decisions, the biases can systematically be captured in automated decision processes that rely on this data. In recent years, interest and attention towards biases embedded in AI systems have grown dramatically. More attention in AI research fields is being devoted to the data and processes that are the sources of biases, how the systems perpetuate the biases, and the mitigation of such biases.

Biases present in AI systems may be derived from a multitude of sources, such as general assumptions in the problem space, ambiguous or prejudice task definition, and the data itself [12]. Biased data comes from both explicit and implicit sources. For example, annotation biases can stem from using crowd-sourced workers who lack the cultural context needed for data labelling [3, 10, 13]. Nobata et al.'s work includes an experiment that compares annotations conducted by crowd-sourced workers and expert annotators. They document low agreement between crowd-sourced annotators and the expert annotators [14]. Implicitly, bias in the data does not necessarily require the introduction of human biases. For example, within hate speech detection identity and lexical biases hinder model performance. Identity biases in this domain create AI systems that are biased to the presence of group identifiers (i.e. terms such as women, Jews, immigrants, etc.) that can create false-positives in classification.

To address some of the human-introduced biases, techniques such as racial or stereotyping priming have been utilized. In Sap et al.'s [3] investigation of racial biases in toxic language datasets they re-annotated the samples using dialect priming with crowd-sourced

workers. They found that annotators that were primed with race/dialect information for AAE tweets were significantly less likely to label those tweets as offensive to anyone compared to the non-primed annotators. Similarly to racial priming, Patton et al.'s Contextual Analysis of Social Media (CASM) framework introduced techniques to address the introduction of biases during data annotation [15]. The CASM technique utilizes a multi-step process in which the annotators examine the cultural and contextualization of the data. The authors observed improved classification results with reduction in annotation biases.

To analyze the effects of bias in the AI systems, a concept of algorithmic fairness is often pursued. Quantitatively, fairness metrics are used to measure and explain the effects of the bias and algorithmic fairness in these systems. In Verma and Rubin's [16] exploration of fairness definitions and metrics, they demonstrate how for the same cases, different definitions of fairness can produce conflicting results of whether a system is fair or unfair. This work explores the definitions and application of these metrics, with groups of fairness metrics categorized as general statistical measures, misclassification rates, to metrics of predicted outcomes, or metrics for both predicted and actual outcomes. A common fairness metric based on predicted outcomes is the statistical parity that measures whether subjects in a protected and unprotected group have the same probability of being in the positive predicted class. For example, assuming the groups are split as m and f, $P(\hat{d} = 1|G = m) = P(\hat{d} = 1|G = f)$. Similarly, a fairness metric that uses the predictive and actual outcome is predictive parity (PPV). Predictive parity is satisfied when both protected and unprotected subjects have equal probability of the positive predicted value belonging to the positive class. Verma and Rubin also explore the use of similarity-based measures that help compare fairness when the non-protected attributes of samples are similar and casual reasoning that represents relations between attributes and predicted outcomes.

As more attention is being devoted to biased AI in the media and research more effort is being exerted to audit datasets and AI systems. Buolamwini and Gebru [17] demonstrated how commercial facial analysis machine learning (ML) systems perpetrate colorism by

misclassifying dark-skinned women at higher rates than lighter-skinned individuals. This study found that by evaluating three commercial facial analysis algorithms by the phenotypic subgroups, there was a dramatic increase in error rates between the dark and light-skinned groups per gender. To correct the under-representation of gender and different skin types in existing facial recognition datasets they also released a balanced and representative facial analysis dataset.

AI biases have been demonstrated across many domains and applications, including facial recognition, voice recognition, recommendation systems, and search engine applications. These bias systems have been reported to be used in practice to favor male software engineers' applicants over females in large technology companies, judge inmates for likelihood of recidivism, and labelling Black individuals as "gorillas" in photo applications [10, 18]. For the rest of this section, we will narrow the scope of AI biases to biases in natural language processing (NLP).

With so much of our information being stored in text form, the biases present in language based systems can enable, enforce, and propagate social hierarchies, inequalities, and harmful stereotypes. Much of the work being done in identifying and mitigating biases in NLP has been concentrated in word-embeddings and language modelling tasks [19]. Bolukbasi et al. [20] demonstrated how widely used pre-trained word embeddings that are trained on Google News articles still capture sexist stereotypes. These gender biases are shown to exist as directions in the word embeddings, such as man $-$ woman is roughly equivalent to computer programmer$-$homemaker. In their work they also introduced an algorithm that debiases these embeddings while preserving the gender association of certain words (i.e. Female and Queen).

In Lu et al. they benchmark gender bias in language modelling and coreference resolution tasks [21]. Language modelling is a task that attempts to model the distribution of word sequence and is commonly used as a pre-training task in Transformer-based architectures [22]. A coreference system is a mention-ranking model that finds words and

expressions referring to the same entity in a natural language text. In their evaluation, they found that bias mirrors stereotypical gender occupations in these systems and bias mitigation strategies of debiased word embeddings and counterfactual data augmentation were not sufficient at mitigating the biases without dramatically impacting predictive results.

Blodgett et al.'s survey of 146 papers related to bias in NLP demonstrates how the field is growing [19]. In this work, the researchers critique the inconsistencies between motivations and quantitative techniques for measuring bias. They also note that bias in NLP tends to not engage in multi-disciplinary literature, as present by 32% of the papers they surveyed were motivated by system performance rather than normative reasoning. Along with an inconsistent range of motivations and assumptions around definitions of "racial bias," "gender bias," or even social injustice imply in these different contexts show that there is still much work to be done in this field of research. This work showcases the difficulties in quantifying system biases and provides recommendations for how future work can better detail, discover, and address these biases.

2.2 Hate Speech

In recent years interest in hate speech detection has grown as a subtopic of natural language processing research and industry practice [1]. To respond to the rise in hateful and toxic language, social platform companies have increased their automatic moderation efforts that detects and removes such language [2]. Issues regarding hate speech detection emerge as many other machine learning disciplines have high agreement of their true labels, while defining toxic and hateful language classification requires expertise in both cultural and historical contexts. Additionally, subjective definitions of what constitutes hate speech, ever-changing forms of societal bigotry, and concerns of suppression of free speech induce challenges that make the automatic hate speech detection a difficult problem to solve.

Due to the lack of an international legal definition of hate speech, we first explore some common definitions of hate speech in Table 2.1. In the survey paper conducted by Fortuna

and Nunes [1], they compare these definitions and note the differences each definition entails. Features of the hate speech definitions differ as an incitement to violence or hate, a call to attach or diminish, and the status of humor have some inconsistencies. Based on their analysis of other papers and policies, the definition they use, and we adopt for the remainder of this thesis is:

"Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used." [1]

With the increased attention on automatic hate speech detection, more datasets have become available for researchers to utilize. However, in the literature there is still no widely accepted benchmark dataset, therefore it is important to note that the available datasets are derived from different sources, annotation guidelines, and domains. Twitter is the most used platform for data collection, being used in several abusive language and hate speech, anti-refugee, and cyber-bullying datasets. Other datasets are derived from Facebook content, far-right and white supremacy forums such as GAB and Stormfront, and from comment sections on websites such as Yahoo! or Wikipedia. More details regarding these and other datasets are available in Table 2.2.

Table 2.1: Definitions of Hate Speech

Source	Definition
United Nations	Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor [23].
Scientific Paper	Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender [24].
Facebook	... as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as code words for PC groups [25].
Twitter	Promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease [26].
YouTube	... content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status [27].

Table 2.2: Toxic Language Datasets

Dataset	Domain	Source	Size
Davidson et al. (2017) [28]	HateBase Terms	Twitter	22050
Founta et al. (2018) [24]	HateBase + Offensive Terms	Twitter	87371
Waseem & Hovy (2016) [29]	Sexist + Racist Terms	Twitter	16849
Golbeck et al. (2018) [30]	Offensive Hashtags + Phrases	Twitter	19715
Warner & Hirschberg (2012) [31]	Anti-semitic	Yahoo + AJC Data	9000
Burnap and Williams (2015) [32]	Aftermath of Terror Attack	Twitter	1878
Silva et al. (2016) [33]	Sample for Hate Speech	Whisper + Twitter	27.55M + 512M
Xiang et al. (2012) [34]	Sample for Hate Speech	Twitter	696M
Ross et al. (2017) [35]	German Anti-Refugee	Twitter	13766
Kennedy et al. (2018) [36]	Far-right Forum	GAB	27665
Gibert et al. (2018) [37]	White Supremacy Forum	Stormfront	9916
Del Vigna et al. (2017) [38]	Italian Hate Speech	Facebook	1687
Zhong et al. (2016) [39]	Cyberbullying	Instagram	3000 pictures + comments
Wulczyn et al. (2017) [40]	Toxic Comments	Wikipedia	100k
Mathew et al. (2020) [41]	Hate Targets	Twitter	20148
ElSherief et al. (2021) [42]	Implicit Hate Speech	Twitter	22056

Earlier work in automatic hate speech detection relied on classical machine learning methods such as logistic regression, SVMs, and random forests [1]. More recently, many of the works exploring hate speech detection leverage deep learning models. In Badjatiya et al.'s work using Waseem and Hovy's dataset they demonstrated how using deep learning approaches of a convolutional neural networks (CNN), long short-term memory (LSTM), and FastText in hate speech detection are able to significantly outperform classical machine learning models trained on bag of words and n-gram embeddings [43, 29]. Qian et al.'s work proposed the usage of intra-user and inter-user representations along with the single tweet in their model. This inclusion improves the classification score when trained on a baseline LSTM model [44]. Additionally, in Zhang et al.'s research they used a combination of a CNN and gated recurrent units (GRU) architecture to capture implicit features of hateful content and computed state-of-the-art results in hate speech detection [45].

More recent work has begun to explore the intricacies of hate speech detection challenges. Matthew et al. has released a benchmark dataset for explainable hate speech detection and used state-of-the-art models (CNN-GRU, BiRNN, BiRNN with attention and BERT) to evaluate performance on this dataset [41]. They found that even while achieving high classification metrics, the models did not score well on explainability metrics. Another dataset focused on implicit hate speech was released to address an underserved type of hate speech by ElSherief et al. [42]. With the dataset they benchmarked their results for classification and generation of intended target and implied meanings tasks using state-of-the-art baselines (BERT, GPT, and GPT-2).

2.3 African American English (AAE)

African American English (AAE), also known by the names of African American Vernacular English, Black English, or Ebonics, is a well studied, rule-bound and grammatical dialect of the English language [46, 47]. AAE is characterized by unique grammatical, pronunciation, and lexical features that distinguish it from Standard American English (SAE)

on which many natural language processing applications are based. AAE's origin are unknown, one hypothesis for its origins is that the dialect was developed as a common language shared among slaves taken from different language backgrounds which developed into a creole language. Another is that slaves in the Southern United States worked along indentured servants of Scottish/Irish descent and learned English through them. Support for these theories are rooted in the linguistic similarities AAE shares with other English dialects and creole languages that persist today [48].

The AAE dialect is rule-based and grammatical, such that it's syntactic and lexical characteristics can be used to develop parts of speech and dialect detection models [49, 50, 51]. In Table 2.3 we present these characteristics and examples.

Jørgensen et al. investigated performance differences for Part-of-Speech (POS) models on SAE and AAE texts. Additionally, they created a POS for AAE associated subtitles, lyrics, and tweet texts. To develop this model, they mined tag dictionaries from various websites to have partially labeled data then manually annotated it. They released the model and accompanying dataset that improved on the state-of-the-art POS model for AAE texts, reducing the disparity in prediction performances between dialects of AAE and SAE [49].

In Blodgett et al.'s research they developed a corpus of Tweets and a dialect estimation ensemble model [50]. In their collection of this corpus they utilized the demographic census data and the geolocation from Twitter's API to define covariates that act as a proxy for the probability a user belongs to a census group. The groupings they chose were non-Hispanic whites, non-Hispanic Blacks, Hispanics, or Asian. Using their collected corpus, they trained a model that outputs the posterior probability of a tweet's author being AAE (non-Hispanic Black), SAE (non-Hispanic white), Hispanic, or Asian. In their analysis, they found that only the first two topics correlated with the census data and recommended discarding the classification results of Hispanic and Asian. This model was verified in its ability to detect AAE texts by using Jørgensen et al.'s [49] POS model to show well-known AAE characteristics exist in the AAE predicted texts.

Similar work was done by Pietro and Ungar who created a dialect estimation model that relies on user-level race and ethnicity predictors on Twitter text samples [51]. This contrasted the existing methodologies that relied on distantly supervised ethnic predictors with census data. Instead, the models developed rely on self-reported user race data and the participants' tweets to predict demographic information. They achieved better results for out-of-sample accuracy when predicting the demographics of the four largest racial/ethnic groups in the United States. The dataset that they used to create this model was also distributed and made available to the research community.

2.4 Bias Mitigation Techniques and Hate Speech

Hate speech and toxic language classification are riddled with challenges due to a low agreement in annotation, complexity and ambiguity in what constitutes hate speech, and a lack of expertise that is required to understand the social and cultural structures that underlay different types of bigotry [1]. These challenges induce biases to the data and models that detect hate speech that unfairly impact certain demographics or lower model performance.

One such issue is the "false-positive" biases associated with identity terms. Work by Dixon et al. attempted to reduce the biases associated with identity terms by re-balancing the dataset with an unsupervised approach [52]. They first demonstrated that class imbalances in the training data leads to unintended bias in the models trained on them. They found that rebalancing their dataset helped reduce the biases without impacting the model's performance quality.

To tackle a similar issue with group-identifiers, Kennedy et al. used a post-hoc explanation regularizer to encourage the classifiers to learn the context around hate speech rather than the models over-reliance on the presence of group-identifiers [6]. This post-hoc explanation regularization uses Occlusion (OC) and Sampling and Occlusion (SOC) explanations over BERT to score how the identifiers contribute to the classification. Dur-

ing training it penalized the model for weighing the presence of a group identifiers heavily, with the intention that the model will instead learn the context surrounding the group identifiers. The results of this work improved the vanilla-BERT's model performance and demonstrated a decrease of the group-identifier biases in hate speech detection.

Work by Park et al. attempted to reduce gender bias in the toxic language datasets that exist against women [4]. They did so by experimenting with debiased word embeddings, data augmentation to swap gender pronouns in the training data, and the usage of a larger corpus. These methods proved very effective at mitigating the bias that discriminates against women and recommend usage of similar approaches in similar scenarios. In Zueva et al.'s work that reduced identity bias in Russian hate speech detection by employing similar principles as previous works to reduce these biases. They utilized language models to generate a larger training set and experimented with random word drop-outs during training such that a protected identity term would be replaced with an unknown token to help the model learn the context surrounding the identifiers [53].

More recently, limitations with the construction of toxic-language datasets have become a source of research in this field. For example, Awal et al. showed that semantically similar samples in hate and abusive language datasets have issues with label consistency [54]. As Fortuna and Nunes discussed, a consistent definition of hate speech does not exist [1]. This subjectivity in label definitions can introduce annotation biases as demonstrated by the graph-based approach developed by Wich et al. that groups annotators to identify annotated biases [55]. They demonstrated this by building a graph based on annotations from different annotators and applied a community detection algorithm. They then trained data from the group of annotators to demonstrate the effects of annotation biases in these datasets.

In another study, Wich et al. demonstrated that politically biased abusive language datasets impair the performance of hate speech classifiers [5]. They constructed a politically biased dataset by collecting samples from the right-wing, left-wing, and center of the

political spectrum. Using this dataset they demonstrated that these political biases negatively affect the performance of the hate speech detection models that are trained on these datasets.

Sap et al. revealed a high correlation between annotators' perception of toxic labels and tweets predicted to be AAE using Blodgett et al.'s dialect-prediction model [3, 50]. Furthermore, by using AAE dialect as a proxy for race, they showed that by relabeling a sample of the dataset with racial priming, the annotation bias towards AAE authors was significantly reduced. In both Sap et al.'s and Davidson et al.'s research, they demonstrated how using the toxic language datasets for training, ad-hoc machine learning models propagate and amplify the racial biases against AAE speakers [3, 56].

Xia et al. and Zhou et al. have proposed approaches to minimize the racial biases that are propagated from these datasets [57, 58]. Xia et al. introduced an adversarial model architecture to reduce the false-positive rates for AAE samples while reducing the impact on classifier performance. Zhou et al. evaluated pre-processing and in-processing debiasing techniques and introduced an experiment of relabeling the dataset by translating the AAE sample to SAE, in order to have the same toxicity label. Their works reported improvements on model biases measured by the false-positive rates. Our work builds on these findings and adds to the state of knowledge by evaluating new debiasing techniques with additional fairness metrics, demonstrating challenges in using a larger corpus for bias-mitigation in low-resource contexts, and introducing an ensemble framework that increases fairness while minimizing classification degradation.

Table 2.3: Characteristics of AAE and Sample Phrases

Characteristic	Definition	Example(s)
Verbal Auxiliaries (inversion, reduction, and singularity) [50].	To form tenses, moods, and voices of other verbs – this includes aspectual markers. In AAE the usage includes a habitual be, future gone, and a completive done, and remote past of bin (been).	"Fees be looking upside my head.", "Now we gone get fucked up.", "damnnn I done let a lot of time pass by.", "I BIN knowing that."
Null Copulas [50]	Removing the link between the subject of a clause to the subject complement. In AAE this occurs when the copula is present, not first person, accented, negative nor conveying present tenses (Green 2002).	"If u with me den u pose to RESPECT ME."
Preverbal markers	In AAE these are words that precede a verb, common uses of ain't in the past tense or steady/stay to indicate an action is done in a consistent matter.	"I ain't want him to know." "Them students be steady trying to make a buck."
Syntactic Properties	Double negatives, existential it, Ass camouflage construction (ACC)	"Ain't nobody can beat me." "They should've fired her ass." "It's some coffee in the kitchen."
Unmarked Possessive [48]	Omitting the -s with verbs following a third person singular subject.	"He jump high."
N-word usage	Usage of the N-word ending with an "a" to indicate another person.	"I know that n***a."

CHAPTER 3

THE DATA

Our work uses four publicly available toxic language datasets derived from Twitter that have compatible definitions of hate speech and toxic language. The four datasets were chosen due to their popularity in toxic language research and their corpus being sampled from Twitter. Our proposed Hierarchical Ensemble (HxEnsemble) model architecture (described in Subsection 4.1.2) includes a specialized AAE language model for which we utilized the TwitterAAE dataset that Blodgett et al. made available in their work to train. [50].

3.1 Toxic Language and AAE Datasets

3.1.1 DWMW17 [28]

Davidson et al. randomly sampled 24,802 tweets that contained words and phrases from Hatebase.org. The tweets were annotated by 3 or more crowd-sourced annotators and assigned labels of: "Hate Speech," "Offensive," or "Neither." The definition for hate speech provided to the annotators was: "language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." The annotators achieved 92% intercoder agreement and the final label distribution were 77.4% offensive language, 5.8% hate, and 16.8% were neither. Using Blodgett et al.'s dialect estimation model, we note that 98% of the AAE tweets in the dataset are labeled as Toxic (either "Hate Speech" or "Offensive"), relative to the 80% of SAE tweets. The "Hate Speech" label had 4% of AAE tweets and 6% of SAE, noting a lower disparity between dialect groups when using a stricter annotation definition.

3.1.2 FDCL18 [24]

Founta et al. created a 79,996-sample dataset consisting of tweets annotated as either "Abusive, Hateful, Normal, or Spam." These tweets were collected from a stream of tweets and then filtered using sentiment analysis and phrases from Hatebase.org. The authors defined hate speech as: "language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender." Inter-coder label agreement when holding out at most one of five annotators was 55.9%. The final label distribution showed 66% of the labels were normal, 16.8% were spam, 12.6% were abusive, and 4.5% were hateful. On this dataset we observe a larger disparity in Toxic and Hate labels in dialect estimation of the tweets, with 84% of AAE tweets labelled as "Toxic" (Abusive or Hateful) compared to 26% of SAE tweets. For this dataset we also observe a disparity between the dialect of tweets with the hateful label – with 21% of AAE tweets labelled as Hateful compared to 4% of SAE tweets.

3.1.3 Golbeck [30]

Golbeck et al.'s binary dataset of tweets labeled as "Harassment" and "Not Harassment" was created from sampling tweets that contained hashtags and phrases that were present in their exploration of offensive tweets. They used two annotators in the first round and when there was disagreement, a third would be brought in to determine a majority label. The "Harassment" label ended up including sub-topics of racism, misogyny, homophobia, threats, hate speech, directed harassment. Non-harassment included potentially offensive and non-harassing tweets. Their definition of hate speech used was: "hate or extreme bias to a particular group. Could be based on religion, race, gender, sexual orientation, etc. Generally, these groups are defined by their inherent attributes, not by things they do or think." The Harassment label accounted for 15.7% of the tweets in the corpus. This dataset is only used in the *Toxic* aggregate as it does not contain enough AAE samples (92)

individually. With the small AAE sample size, we still observe a disparity in annotation, where 50% of AAE samples are labeled as "Harassment" compared to 24% of SAE tweets.

3.1.4 WH16 [29]

The Waseem and Hovy (WH16) 16,849 sample dataset was collected by sampling tweets containing at least one of the phrases or words they deemed to be hateful. The authors labeled the tweets as racist, sexist, or neither using guidelines inspired by critical race theory and had a domain expert review their labels. However, this dataset has received significant criticism from scholars [59, 60], who deride it for most of the racist tweets being anti-Muslim and the sexist tweets relating to a debate over an Australian television show. Additionally, this dataset can introduce author bias as it is noted that two users wrote 70% of sexist tweets and 99% of racist tweets were written by another single user. Due to these limitations, we only include the positive instances of this dataset in the aggregated datasets *Toxic* and *Hate* as its usage in an aggregate largely addresses the author and topic biases.

3.1.5 TwitterAAE [50]

Blodgett et al.'s TwitterAAE dataset was developed in adjacency with their dialect estimation model that we used throughout our work. This dataset contains 1,045,467 samples of Tweets that their dialect model predicted as being AAE. The associated data includes the geo-location and the census block that location corresponds to and the other posterior probabilities for the prediction of race.

3.2 Data Preparation

To estimate the dialect of the tweet, we used Blodgett et al.'s [50] dialect estimation model that outputs the posterior probability of the text sample belonging to the dialect of AAE or SAE. Due to the low incidence of AAE tweets within the toxic language datasets, we used

a lower threshold of $Pr(AAE \geq 0.6)$ relative to Blodgett et al.'s use of $Pr(AAE \geq 0.8)$ in order to have a larger sample of AAE tweets. In the Appendix A, we also showed the fairness analysis of the hate-detection algorithms using the $Pr(AAE \geq 0.8)$ threshold in order to provide equal comparisons and observed similar patterns. In Section 5.2 we also discussed how the lower threshold may lead to some misclassified SAE instances being evaluated in the specialized AAE model.

A common first step in addressing biases in datasets is using a larger corpus to increase instances of the minority class [4]. We evaluated the *DWMW17* and *FDCL18* datasets individually and created two aggregate datasets to better understand the racial biases and annotation agreement across the toxic language datasets. These experiments informed the challenges and feasibility of using a larger corpus or complimentary AAE toxic language data to address the racial biases present in available datasets. We chose four commonly used toxic language datasets that are taken from Twitter and annotated with compatible definitions of either `toxic` or `hate`. The low AAE totals in the individual datasets of *Golbeck* and *WH16* made it difficult to assess the racial biases on their own. When evaluated as part of the aggregated dataset, they informed cross-corpus label agreement and how the use of a larger corpus impacts classification and fairness metrics.

For evaluation, we segmented the datasets based on binary labels of `hate` or `toxic`. The *Hate* dataset is aggregated on the `hate` labels from *DWMW17* and *FDCL18* and only the subset of positive instances of *WH16* that were labeled as `racist` or `sexist`. The *Toxic* aggregate dataset included the `hate` and `offensive` instances of *DWMW17*, the `hate` and `abusive` instances from *FDCL18*, the positive instances of *WH16* `racist` or `sexist`, and the `harassment` instances from *Golbeck17*. We evaluated the outcomes from the hate-detection algorithms applied to *DWMW17* and *FDCL18* datasets based on the `toxic` label due to the strong association with AAE text being marked as toxic by annotators and the lower sample count of hate in AAE texts. Additionally, using the aggregate datasets of *Hate* and *Toxic*, we gain insight into the effects of using a larger corpus to

reduce biases and compare the effects of annotation consistency on classifier performance.

Table 3.1: Datasets Characteristics and Dialect Label Comparisons

Dataset (n=)	AAE[1]	Toxic	Hate	Total	$\text{Toxic}_{\text{AAE}}, \text{Toxic}_{\text{SAE}}$[2]	$\text{Hate}_{\text{AAE}}, \text{Hate}_{\text{SAE}}$[3]	Annotated By[4]
DWMW17	4878	20620	1430	22050	0.98, 0.80	0.04, 0.06	3+ CSW
FDCL18	1265	24821	4119	87371	0.84, 0.26	0.21, 0.04	5 CSW
Golbeck	92	4760	-	19715	0.50, 0.24	-	2+ GRA
$WH16_{subset}$	12	3360	3360	3360	-	-	Authors + DE
Toxic	6205	52679	-	134457	0.94, 0.36	-	-
Hate	6120	-	8710	123886	-	0.08, 0.07	-

[1] Number of samples in the dataset Blodgett et al.'s [50] dialect model predicts as $Pr(AAE \geq 0.6)$
[2] Proportion of the AAE vs. SAE dialect samples labeled as `Toxic`
[3] Proportion of the AAE vs. SAE dialect samples labeled as `Hate`
[4] Crowd-Sourced Worker (CSW), Graduate Research Assistant (GRA), Domain Expert (DE)

CHAPTER 4

RESEARCH DESIGN

In prior work and Table 3.1, a strong correlation was found between instances of AAE dialect and the associated annotation of toxic labels in abusive language datasets. This leads to racial bias against African American authors in models trained on them, introducing a higher false-positive rate (FPR) for instances that are AAE [3, 56]. A higher FPR means that non-toxic AAE texts are more likely to be misclassified as toxic by hate-detection algorithms. This lexicographical and identity bias can then become embedded and further propagated through the classifiers that are trained with them.

In our work, we benchmarked classifier performance and fairness indicators across the datasets based on a series of different hate-detection algorithms including baselines, logistic regression models with unigram and bigram encodings, TF-IDF, and GloVe embeddings [61]. We also evaluated a vanilla-BERT classifier [22], a commonly used language model in NLP classification tasks. We then evaluated the effectiveness of two in-process debiasing algorithms that use BERT as a base model. Finally, we introduce a new ensemble framework as a proposed method to remediate biases that may be attributed to low-resource contexts.

4.1 Models and Bias Remediation Techniques

4.1.1 In-Process Debasing Algorithms

Explanation Regularization

Kennedy et al. [6] used the Occlusion (OC) and Sampling and Occlusion (SOC) explanation over BERT to generate hierarchical explanations for a prediction and use it to score how a phrase contributes to the classification. This score is then used to regularize the

model during learning, with the intention of mitigating the compositional effects of a phrase and the context around it. AAE's characteristic of using the n-word to indicate another person is an example of the in-group reclaiming a group identifier that was used to oppress and dehumanize Black individuals. In-group usage of the term is not considered hateful, only out-group usage is deemed hateful. Ass Camouflage Construction (ACC) is another characteristic of AAE that has "ass" or "butt" usually proceeded by a possessive pronoun and is an equivalent to the reflexive self. With the purpose of reducing the classifier's bias towards the in-group identifiers and AAE pronoun altercations, we applied this regularization algorithm to help the model to learn the context surrounding the pronouns and identifiers in AAE [62, 63, 64].

MinDiff Framework

Prost et al. [7] introduced a regularization technique that penalized models for dependence between the distribution of predicted probabilities and a protected subgroup, such as AAE. This framework attempts to minimize the difference between the protected subgroup and the majority (unprotected) group distributions. This algorithm minimizes the differences in FPR across the two slices with the intention of a minimal impact on classification performance. This algorithm has been discussed as an effective manner of reducing biases in language modeling tasks but is limited by the available data samples in the group slices.

4.1.2 Hierarchical Ensemble Framework (HxEnsemble)

This paper proposes a hierarchical ensemble framework that minimizes potential classifier performance degradation while mitigating biases that are a result of training data that does not effectively represent the target population. To achieve closer to equality results across groups (AAE and SAE authors), we make use of the general classifier that contains biases to the "protected group," AAE authors. For instances predicted as positive (toxic) in the general model and in the dialect estimation model (as AAE), the ensemble will pass them

to a specialized classifier that is pre-trained to the AAE dialect and fine-tuned on only AAE samples in the toxic language datasets. Effectively, in order to achieve a classifier that has closer to equal-outcomes across groups, we make use of an equity-based framework that is better able to predict positive instances of the protected group which the general model has been shown to exhibit bias against.

A related technique was first presented by Howard et al. [65], where it was shown that an ensemble framework achieved better classification performance and reduced FPR for misclassified emotions by using a combination of a generalized model and a specialized learner that is trained on the classes that are most commonly misclassified. As demonstrated in Sap et al. [3], annotator biases towards the AAE dialect may be minimized during labeling when the annotators are racially primed of the potential race of the author. Since AAE has different syntactical and lexicographical characteristics, we hypothesize that similar to racially primed annotators, a classifier that is pre-trained on non-toxic AAE samples is better able to distinguish between AAE samples in the toxic language datasets that are true-positives and those that are true-negatives. While the underlying language model remains a black-box, we conjecture that the AAE-BERT will have a better understanding of the AAE dialect. Therefore, when the classifier for AAE toxic language classification is fine-tuned on it, there will be less biases propagated than using a base language model that is trained on a corpus of mainly SAE samples.

The motivation behind using a general and specialized learner as a debiasing tool is that we hypothesize that the general learner and language model does not properly represent the AAE dialect. As there exists an ethical trade-off in bias mitigation strategies between model performance and bias reduction, when a specialized learner is used for the under-represented group we expect better model performance for the group that has biases against them. As a general learner is prone to biases for smaller sample groups, combining them together to reduce the false-positive cases should address the bias issues without degrading the overall model performance. Additionally this allows samples that belong in the under-

represented group to have the correct context in the model that is evaluating them. In the HxEnsemble case, this means that the specialized learner has a better understanding of the AAE dialect. Therefore able to better distinguish between AAE samples that are truly toxic and those that are not compared to a model that is mostly trained on language data outside of AAE's distribution.

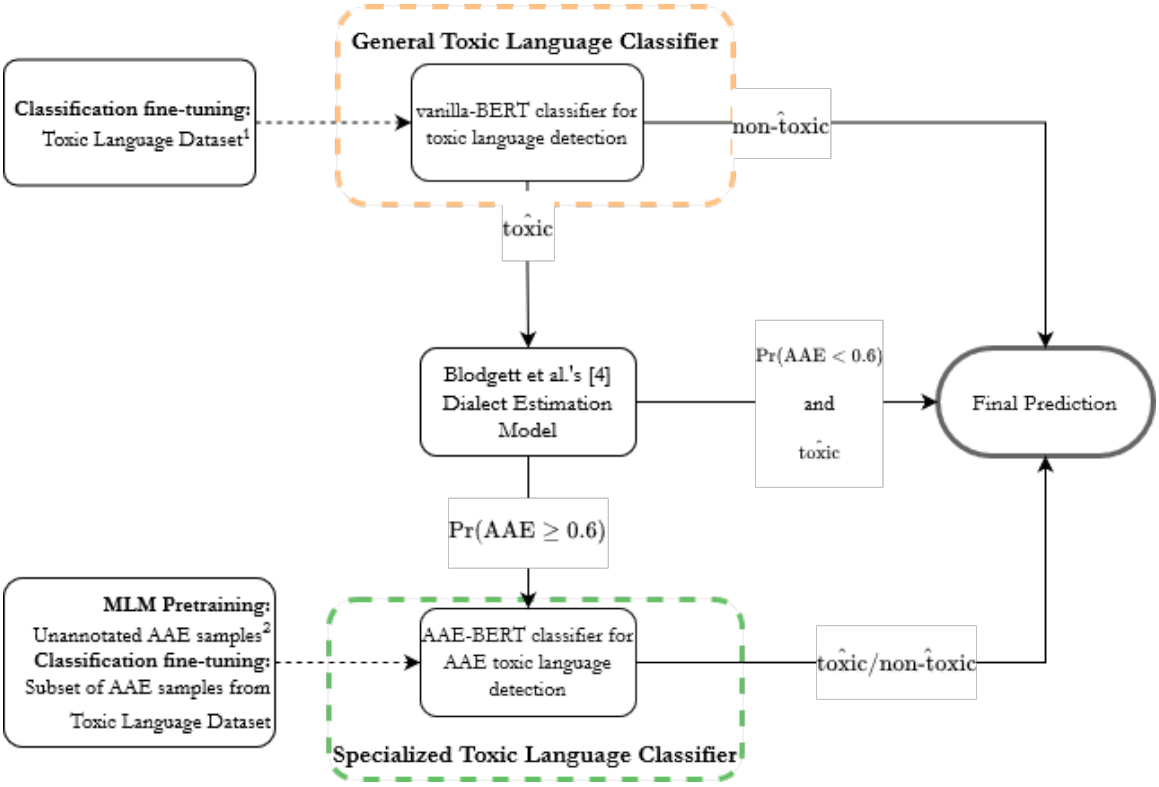

Figure 4.1: The Hierarchical Ensemble (HxEnsemble) Framework
[1] Training set of DWMW17, FDCL18, Toxic, or Hate is used. More details about these datasets can be found in Table 3.1.
[2] The unannotated AAE samples used for the MLM pre-training come from Blodgett et al. [50]

The proposed HxEnsemble (Figure Figure 4.1) uses a general toxic language detection model, in our case a vanilla-BERT classifier trained on the original dataset. If the general model predicted a positive instance (`hate` or `toxic`), we then used the out-of-box Blodgett et al. dialect estimation model [50] to predict the probability the sample is AAE. If the sample is not AAE, we returned the predicted result from the general classifier. However, if the text is AAE we then passed it through the specialized AAE classifier and the en-

semble model returned that prediction. The specialized classifier is created by fine-tuning the vanilla BERT on a masked language model task with Blodgett et al.'s corpus of demographic based Twitter data that belongs to African American authors [50]. We then use this BERT-AAE language model to fine-tune a classifier head on the AAE samples in the same train-validation-test splits that the general model was trained on. The final prediction for the positive AAE sample is the prediction of the specialized AAE classifier.

4.1.3 Fairness Metrics

Fairness metrics are used to statistically evaluate notions of fairness in classifier performance, where certain metrics can reflect different definitions of fairness [16]. In previous work by Zhou et al. [58] and Xia et al. [57] that explore bias remediation techniques for toxic language detection, the FPR is used as the single fairness metric for reporting how the biases in the training data propagate to the models. As seen in Table 3.1, the positively-skewed distribution of AAE texts in the datasets means that there is a very limited number of true-negatives in the data splits that were evaluated and the insights to the effects of biases is limited. As such and in order to address the low occurrences of AAE samples, we also compute a fairness metrics based on the disparate impact (DI) metric [16], also commonly referred to as adverse impact. DI is a fairness metric that evaluates the predictive parity ratio to compare predicted outcomes across groups, which does not rely on the annotations in its computation. In practice, the acceptable fairness range for this metric is limited to $\{0.8 - 1.2\}$ [66], and a $DI < 1$ is interpreted as bias against the protected group, in our case the AAE authors. Conversely, if $DI > 1$ there is said to be bias towards the protected group.

$$DI_{\text{fav}} = \frac{Pr(\hat{Y} = 0|D = \text{AAE})}{Pr(\hat{Y} = 0|D = \text{SAE})}, DI_{\text{unfav}} = \frac{Pr(\hat{Y} = 1|D = \text{AAE})}{Pr(\hat{Y} = 1|D = \text{SAE})} \tag{4.1}$$

In toxic language detection, the notion of a favorable prediction is subjective. Therefore we analyzed the DI for when the prediction is non-toxic (DI_{fav}) and toxic (DI_{unfav}) to

gain more insight into how the classifier treats the different dialect groups for both outcomes. Lastly, we looked at false negative rates (FNR) to provide insight into whether the bias remediation techniques are causing disagreement with the positively-skewed toxic annotations for AAE authors.

4.1.4 Experiment Implementation

To evaluate the hate-detection algorithms as applied to the aforementioned datasets, we randomly split all our datasets stratified on the positive-cases into 80% training, 10% validation, and 10% for testing. We used the training and validation splits to run a grid-search in order to fine-tune the algorithms we benchmarked. We chose the hyperparameters that resulted in the best validation F1-score and trained the model across 10 random seeds to report our results below (Chapter 5).[1]

For the logistic regression-based models, we tune the hyperparameters on learning rates $\{2\text{e}{-3}, 2\text{e}{-5}, 5\text{e}{-5}\}$, epochs $\{10, 100, 1000\}$, and batch sizes $\{16, 32, 64\}$. For all BERT-based models, we tuned the hyperparameters on learning rates $\{2\text{e}{-3}, 2\text{e}{-5}, 5\text{e}{-5}\}$, epochs $\{1, 2, 3\}$, and batch sizes $\{16, 32, 64\}$. For the Explanation Regularization model, we searched the regularization strength $\{0.1, 0.3, 0.5\}$, and for the MinDiff Framework we fine-tuned MinDiff weight to $\{0.5, 1, 1.5\}$. With the specialized AAE learner in the hierarchical ensemble, we ran the parameter search on both the masked language model task and classification task together.

[1] All our code is available at https://github.com/matanhalevy/DebiasingHateDetectionAAE

Table 4.1: Hyperparameters used in Grid-Search

Algorithms	Learning Rates	Epochs	Batch Sizes	Regularization Strengths
Logistic Regression Models	$\{2e{-}3, 2e{-}5, 5e{-}5\}$	$\{10, 100, 1000\}$	$\{16, 32, 64\}$	-
Vanilla BERT	$\{2e{-}3, 2e{-}5, 5e{-}5\}$	$\{1, 2, 3\}$	$\{16, 32, 64\}$	-
Explanation Regularization	$\{2e{-}3, 2e{-}5, 5e{-}5\}$	$\{1, 2, 3\}$	$\{16, 32, 64\}$	$\{0.1, 0.3, 0.5\}$
MinDiff	$\{2e{-}3, 2e{-}5, 5e{-}5\}$	$\{1, 2, 3\}$	$\{16, 32, 64\}$	$\{0.5, 1, 1.5\}$
HxEnsemble	$\{2e{-}3, 2e{-}5, 5e{-}5\}$	$\{1, 2, 3\}$	$\{16, 32, 64\}$	-

CHAPTER 5

RESULTS

5.1 Experiment Results

All metrics reported below include the subscript of their standard deviation across the 10 randomized trials.

5.1.1 DWMW17

In Table 5.1, we present the results of the different classifiers trained on *DWMW17*. Individually, this dataset is the most difficult to analyze for fairness due to the fact that 98% of AAE samples in the complete dataset are annotated as toxic, and only 2.5% of the 502 AAE samples in the test set are labeled as true-negatives. The classifiers of BERT, BERT with OC and SOC, and our HxEnsemble method have the best classification performance. After evaluating the fairness metrics across the best performing models, we verify that the HxEnsemble model achieved the best disparate impact scores for prediction of non-toxic and toxic outcomes. We theorize this result is due to the slight increase in the FNR_{AAE} and decrease in the FNR_{SAE} compared to the other models. We observe that BERT+OC achieves the lowest FPR_{AAE}. However, for BERT+OC, BERT+SOC, and HxEnsemble, the FNR_{AAE} and FPR_{AAE} are within each others' error bounds. Overall we note that for *DWMW17*, all classifiers are biased towards being more likely to predict AAE text as toxic by a significant ratio, which is not as evident when examining the prediction parity across groups. Due to the low true-negative AAE samples, we did not conclude any significant results for the classifiers trained on *DWMW17* with regards to unfavorable outcomes.

Table 5.1: DWMW17 Results

Task Name [*]	Acc	F1	DI_{fav} [1]	DI_{unfav} [2]	FNR_{AAE} [3]	FNR_{SAE} [4]	FPR_{AAE} [5]	FPR_{SAE}
N-Gram	$0.947_{\pm 0.001}$	$0.968_{\pm 0.000}$	$0.112_{\pm 0.006}$	$1.203_{\pm 0.003}$	$0.003_{\pm 0.001}$	$0.027_{\pm 0.001}$	$0.308_{\pm 0.000}$	$0.206_{\pm 0.004}$
TF-IDF	$0.872_{\pm 0.000}$	$0.926_{\pm 0.000}$	$0.100_{\pm 0.009}$	$1.111_{\pm 0.001}$	$0.007_{\pm 0.001}$	$0.035_{\pm 0.001}$	$0.846_{\pm 0.000}$	$0.604_{\pm 0.003}$
GloVe	$0.886_{\pm 0.002}$	$0.933_{\pm 0.001}$	$0.137_{\pm 0.013}$	$1.177_{\pm 0.021}$	$0.011_{\pm 0.003}$	$0.062_{\pm 0.010}$	$0.523_{\pm 0.061}$	$0.421_{\pm 0.046}$
BERT	$0.965_{\pm 0.002}$	$0.979_{\pm 0.001}$	$0.098_{\pm 0.014}$	$1.231_{\pm 0.006}$	$0.002_{\pm 0.001}$	$0.024_{\pm 0.003}$	$0.308_{\pm 0.115}$	$0.108_{\pm 0.012}$
BERT+OC	$0.966_{\pm 0.001}$	$0.979_{\pm 0.001}$	$0.100_{\pm 0.005}$	$1.248_{\pm 0.008}$	$0.002_{\pm 0.001}$	$0.030_{\pm 0.003}$	$\mathbf{0.238_{\pm 0.044}}$	$\mathbf{0.081_{\pm 0.013}}$
BERT+SOC	$0.968_{\pm 0.001}$	$0.980_{\pm 0.001}$	$0.098_{\pm 0.008}$	$1.248_{\pm 0.005}$	$0.002_{\pm 0.001}$	$0.029_{\pm 0.002}$	$0.246_{\pm 0.049}$	$0.077_{\pm 0.009}$
BERT+MD	$0.887_{\pm 0.059}$	$0.936_{\pm 0.031}$	$0.053_{\pm 0.057}$	$1.108_{\pm 0.119}$	$0.003_{\pm 0.007}$	$0.016_{\pm 0.026}$	$0.746_{\pm 0.310}$	$0.601_{\pm 0.420}$
HxEnsemble	$0.964_{\pm 0.002}$	$0.978_{\pm 0.001}$	$\mathbf{0.114_{\pm 0.008}}$	$\mathbf{1.223_{\pm 0.004}}$	$\mathbf{0.004_{\pm 0.002}}$	$\mathbf{0.022_{\pm 0.001}}$	$0.262_{\pm 0.040}$	$0.121_{\pm 0.009}$

[*] Models that had the best classification accuracy are *italicized* and the best fairness indicators on the best performing models are **bolded** per column.
[1] Disparate Impact for favorable outcomes measures the prediction disparity for AAE and SAE authors being predicted as non-toxic.
[2] Disparate Impact for unfavorable outcomes measures the prediction disparity for AAE and SAE authors being predicted as toxic.
[3] FNR_{AAE}: the "best" fairness metric in this case is the highest FNR, since the annotations are biased to labeling AAE samples as toxic, an increase may be indicative of the model unlearning these biases.
[4] FNR_{SAE}: for this metric, we want the lowest score as we only care about classification performance of the SAE group.
[5] For both AAE and SAE group, a lower FPR is better.

5.1.2 FDCL18

Table 5.2 shows the results the classifiers achieved on *FDCL18*. Similar to *DWMW17*, there is a low true-negative count in the test set with 14.5% of 117 AAE samples being annotated as non-toxic. BERT, BERT+OC, BERT+SOC, and HxEnsemble achieved the best classifier performance. HxEnsemble achieved the best fairness results across favorable and unfavorable disparate impact scores and FPR_{AAE}. It's worth noting that HxEnsemble also has the highest FNR_{AAE} amongst this subset and the lowest disparity between FPR_{AAE} to FPR_{SAE}, providing some empirical evidence that increasing the language model's concept of the AAE dialect causes the model to disagree with the biased AAE annotations.

We noted that for all models the FNR_{AAE} is lower than FNR_{SAE}, while the FPR_{AAE} is larger than FPR_{SAE}. This means that all models are more likely to misclassify non-toxic AAE samples as toxic compared to SAE samples and less likely to misclassify toxic AAE samples as non-toxic compared to SAE samples. The high FPR_{AAE} and disparate impact scores show significant bias towards AAE authors for both favorable and unfavorable outcomes across all models that are trained on FDCL18. For this dataset, the bias remediation techniques helped reduce the FPR disparity but did not effectively mitigate the prediction disparity. These results demonstrated how underlying data bias to AAE authors propagate to the models even with bias-remediation techniques.

Table 5.2: FDCL18 Results

Task Name *	Acc	F1	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.935_{\pm 0.000}$	$0.878_{\pm 0.001}$	$0.194_{\pm 0.000}$	$3.398_{\pm 0.016}$	$0.040_{\pm 0.000}$	$0.147_{\pm 0.002}$	$0.235_{\pm 0.000}$	$0.036_{\pm 0.001}$
TF-IDF	$0.896_{\pm 0.001}$	$0.788_{\pm 0.001}$	$0.334_{\pm 0.005}$	$3.453_{\pm 0.021}$	$0.187_{\pm 0.005}$	$0.291_{\pm 0.002}$	$0.294_{\pm 0.000}$	$0.036_{\pm 0.000}$
GloVe	$0.900_{\pm 0.001}$	$0.810_{\pm 0.002}$	$0.209_{\pm 0.005}$	$3.417_{\pm 0.028}$	$0.064_{\pm 0.005}$	$0.222_{\pm 0.005}$	$0.294_{\pm 0.000}$	$0.056_{\pm 0.002}$
BERT	$0.943_{\pm 0.001}$	$0.896_{\pm 0.001}$	$0.155_{\pm 0.005}$	$3.287_{\pm 0.015}$	$0.001_{\pm 0.003}$	$\mathbf{0.098_{\pm 0.003}}$	$0.229_{\pm 0.019}$	$0.043_{\pm 0.001}$
BERT+OC	$0.943_{\pm 0.001}$	$0.895_{\pm 0.001}$	$0.164_{\pm 0.009}$	$3.325_{\pm 0.038}$	$0.006_{\pm 0.007}$	$0.108_{\pm 0.003}$	$0.206_{\pm 0.031}$	$\mathbf{0.040_{\pm 0.002}}$
BERT+SOC	$0.943_{\pm 0.001}$	$0.895_{\pm 0.002}$	$0.166_{\pm 0.016}$	$3.315_{\pm 0.038}$	$0.009_{\pm 0.009}$	$0.107_{\pm 0.005}$	$0.212_{\pm 0.041}$	$0.040_{\pm 0.003}$
BERT+MD	$0.824_{\pm 0.102}$	$0.429_{\pm 0.453}$	$0.616_{\pm 0.406}$	$1.465_{\pm 1.551}$	$0.526_{\pm 0.500}$	$0.555_{\pm 0.469}$	$0.082_{\pm 0.101}$	$0.035_{\pm 0.045}$
HxEnsemble	$0.943_{\pm 0.000}$	$0.896_{\pm 0.001}$	$\mathbf{0.180_{\pm 0.019}}$	$\mathbf{3.243_{\pm 0.051}}$	$\mathbf{0.016_{\pm 0.015}}$	$0.101_{\pm 0.004}$	$\mathbf{0.183_{\pm 0.035}}$	$0.045_{\pm 0.009}$

* Please refer to Table 5.1 for the explanation of the table results.

5.1.3 Toxic

In Table 5.3, the results for the classifiers on the aggregate *Toxic* dataset is shown, we note that for this aggregate dataset the true-negative count for AAE instances is 5.7%. BERT and HxEnsemble are the algorithms that had the best classification scores, while BERT with OC and SOC achieved slightly worse results. HxEnsemble achieved the lowest FPR_{AAE} across all algorithms, and the best disparate impact scores across all BERT-based models. MinDiff performs poorly on this dataset, even with more AAE samples available. This pattern is demonstrated across the low F1 scores MinDiff has on every dataset, excluding *DWMW17*. The same pattern of higher FPR and lower FNR for AAE samples exists in this aggregate but, compared to the standalone *FDCL18*, the disparate impact scores provide partial evidence that an effective strategy to mitigate biases is to aggregate a larger corpus. In comparison to Table 5.4, the F1 scores for the algorithms suggest that there is more label agreement for the definition of `toxic`, compared to the F1 scores on the more stringent definition of `hate`. This is indicative that in order to better deal with label inconsistency issues across toxic-language datasets, users must opt for a more general definition of toxic rather than hate.

Table 5.3: Toxic Results

Task Name *	Acc	F1	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.900_{\pm 0.000}$	$0.867_{\pm 0.001}$	$0.059_{\pm 0.001}$	$2.904_{\pm 0.015}$	$0.014_{\pm 0.001}$	$0.189_{\pm 0.003}$	$0.385_{\pm 0.000}$	$0.054_{\pm 0.001}$
TF-IDF	$0.844_{\pm 0.001}$	$0.780_{\pm 0.002}$	$0.136_{\pm 0.002}$	$3.121_{\pm 0.026}$	$0.077_{\pm 0.001}$	$0.321_{\pm 0.004}$	$0.473_{\pm 0.019}$	$0.065_{\pm 0.002}$
GloVe	$0.840_{\pm 0.000}$	$0.787_{\pm 0.000}$	$0.114_{\pm 0.003}$	$2.752_{\pm 0.019}$	$0.061_{\pm 0.002}$	$0.264_{\pm 0.002}$	$0.615_{\pm 0.000}$	$0.106_{\pm 0.002}$
BERT	$0.915_{\pm 0.000}$	$0.890_{\pm 0.001}$	$0.058_{\pm 0.003}$	$2.850_{\pm 0.012}$	$0.008_{\pm 0.001}$	$0.156_{\pm 0.002}$	$0.300_{\pm 0.040}$	$\mathbf{0.046_{\pm 0.001}}$
BERT+OC	$0.911_{\pm 0.001}$	$0.883_{\pm 0.001}$	$0.053_{\pm 0.004}$	$2.863_{\pm 0.031}$	$0.009_{\pm 0.001}$	$0.165_{\pm 0.006}$	$0.392_{\pm 0.065}$	$0.050_{\pm 0.003}$
BERT+SOC	$0.911_{\pm 0.001}$	$0.883_{\pm 0.001}$	$0.059_{\pm 0.003}$	$2.837_{\pm 0.025}$	$0.010_{\pm 0.001}$	$0.162_{\pm 0.005}$	$0.319_{\pm 0.036}$	$0.051_{\pm 0.003}$
BERT+MD	$0.632_{\pm 0.131}$	$0.227_{\pm 0.294}$	$0.700_{\pm 0.483}$	$0.410_{\pm 0.531}$	$0.637_{\pm 0.480}$	$0.629_{\pm 0.486}$	$0.362_{\pm 0.480}$	$0.343_{\pm 0.457}$
HxEnsemble	$0.914_{\pm 0.001}$	$0.887_{\pm 0.001}$	$\mathbf{0.072_{\pm 0.005}}$	$\mathbf{2.778_{\pm 0.029}}$	$\mathbf{0.016_{\pm 0.003}}$	$\mathbf{0.152_{\pm 0.004}}$	$\mathbf{0.277_{\pm 0.016}}$	$0.052_{\pm 0.003}$

* Please refer to Table 5.1 for the explanation of the table results.

5.1.4 Hate

As in the *Toxic* dataset, BERT and HxEnsemble perform best on the *Hate* aggregate dataset. However, across all algorithms the low F1-score suggested the aggregated datasets individually had a low agreement in their `hate` label. Although we attribute this to poor model performance, the favorable and unfavorable disparate impact scores achieved more fairness in their predictions than in other datasets. HxEnsemble introduces bias in favor of AAE as the high FNR and low FPR yielded a lower disparate impact score for the unfavorable outcome compared to vanilla-BERT. The use of this aggregate dataset is not very informative of biases against AAE authors as the learners themselves struggle to deal with the class-imbalance and inconsistent hate definitions across the datasets as seen by low classification scores.

Table 5.4: Hate Results

Task Name *	Acc	F1	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.936_{\pm 0.000}$	$0.368_{\pm 0.006}$	$0.990_{\pm 0.002}$	$1.316_{\pm 0.074}$	$0.736_{\pm 0.026}$	$0.735_{\pm 0.008}$	$0.022_{\pm 0.001}$	$0.013_{\pm 0.001}$
TF-IDF	$0.936_{\pm 0.000}$	$0.234_{\pm 0.002}$	$0.985_{\pm 0.001}$	$2.156_{\pm 0.071}$	$0.807_{\pm 0.008}$	$0.863_{\pm 0.002}$	$0.015_{\pm 0.000}$	$0.004_{\pm 0.000}$
GloVe	$0.928_{\pm 0.003}$	$0.181_{\pm 0.034}$	$1.001_{\pm 0.006}$	$0.947_{\pm 0.381}$	$0.959_{\pm 0.033}$	$0.881_{\pm 0.031}$	$0.015_{\pm 0.006}$	$0.011_{\pm 0.005}$
BERT	$0.943_{\pm 0.001}$	$\mathbf{0.522_{\pm 0.006}}$	$\mathbf{0.987_{\pm 0.005}}$	$\mathbf{1.231_{\pm 0.084}}$	$0.524_{\pm 0.027}$	$\mathbf{0.534_{\pm 0.008}}$	$0.035_{\pm 0.004}$	$0.024_{\pm 0.002}$
BERT+OC	$0.938_{\pm 0.002}$	$0.386_{\pm 0.031}$	$0.979_{\pm 0.016}$	$1.675_{\pm 0.499}$	$0.690_{\pm 0.090}$	$0.724_{\pm 0.035}$	$0.029_{\pm 0.013}$	$0.011_{\pm 0.004}$
BERT+SOC	$0.937_{\pm 0.002}$	$0.396_{\pm 0.019}$	$0.971_{\pm 0.029}$	$1.890_{\pm 0.954}$	$0.645_{\pm 0.111}$	$0.710_{\pm 0.028}$	$0.037_{\pm 0.023}$	$0.013_{\pm 0.003}$
BERT+MD	$0.930_{\pm 0.000}$	$0.079_{\pm 0.068}$	$0.400_{\pm 0.516}$	$0.600_{\pm 0.516}$	$0.400_{\pm 0.516}$	$0.400_{\pm 0.516}$	$0.600_{\pm 0.516}$	$0.600_{\pm 0.516}$
HxEnsemble	$0.942_{\pm 0.001}$	$0.516_{\pm 0.009}$	$1.030_{\pm 0.011}$	$0.445_{\pm 0.203}$	$\mathbf{0.795_{\pm 0.101}}$	$0.549_{\pm 0.011}$	$\mathbf{0.008_{\pm 0.004}}$	$\mathbf{0.021_{\pm 0.001}}$

* Please refer to Table 5.1 for the explanation of the table results.

5.2 Error Analysis and Challenges for Hate-Detection Algorithms in Classifying AAE Text

Table 5.5: AAE Tweets that were Misclassified by HxEnsemble

	Tweet	$\hat{y}$	y
1	RT : still stuntin on my ex bitch	1	0
2	RT : IF U SEE ME SMILING IM EITHER ABOUT TO KILL U OR REALLY FUCKED UP EITHER WAY I AINT IN THE RITE STATE OF MIND	0	1
3	RT : That ain't yo n**** if he let you go to sleep mad sis	1	0
4	ID SLAP YO BIPLAR ASS POLAR BITCH U AINT SHIT BUT A WANNA BE ASS BITCH WHO BETTER WATCH THE FUCK OUT YOU UGLY ASS HOE	1	0
5	RT : Whatchu know bout livin in the ghetto, eating cottage cheese 3 times a day. Lil salt Lil peppa lil A1, Lil barbeeequeee	0	1
6	Yea all these bad ass bey bey kids ride bikes over here	1	0
7	Y'all n****s annoying I'm bout to start calling y'all sis	1	0
8	Blatt trash too ... but twitter said he was good	0	1
9	YO IMMA MOTHERFUCKING BOYYYYYYYYYYYYYYYY	0	1
10	Fucc u primo aint nobody ask u 2 fav dink lies	0	1
11	RT : Who did y'all n****s give y'all hearts to for y'all to be so cold hearted	1	0
12	You gots da BET and da MSNBC wit dat colored reverend who only listen to coloreds. What else I missing?	0	1
13	He did go in i was actually surprised he when that hard RT : N the crib listen 2 my n** G-Eazy;don't diss my homie	0	1
14	Dumb Haitian fake black faggots. Go to Haiti and neck yourself.	0	1
15	#TripleTalaqArrest 2fgt dis social menace #TripleTalaq rgts shd b gvn 2d Muslim wmn2, den dis evil wl hv a natural death #TripleTalaqDebate	0	1
16	yall trash but cold	0	1
17	RT : Whatchu sayin hoe, YOU KNOW IM THE MAN HOE	1	0

In Table 5.5 we present all the misclassified AAE tweets from our test set from a run using our HxEnsemble. Below, we categorized these tweets into 7 categories in order to better understand the challenges in debiasing hate-detection classifiers as applied to AAE authors. We include the False-Negative (FN) and False-Positive (FP) prepended to our category names to clarify the type of challenge.

1. **FN: Mislabeled, Non-Toxic** HxEnsemble predicted examples 5, 9, 13, and 16 as non-toxic while annotators labeled them as `toxic`. For these examples we see that the HxEnsemble model overcame the annotators' bias and provided the prediction that matches the definition of non-toxic language provided to the annotators.

2. **FN: Mislabeled, Non-Targeted Threat** For example 2, while the text contains aggressive language it remains non-targeted and the author of the tweet is only talking about themselves. We disagreed with the annotators that label this as `toxic` and agree with the HxEnsemble prediction. However, this may be interpreted as a threat but, since it's not directed at a person or group, we dismiss it as such.

3. **FN: Missing Context, Unclear Toxicity** We observed that instances 8 and 10 are not clear in what constituted them as `toxic`. While these two examples have sentiments that may be slightly demeaning, we were unable to conclude whether we agreed with the annotators' labels.

4. **FN: Toxic and False-Positive AAE** Three examples of `toxic` instances were misclassified as AAE but the dialect model used the lower threshold of 0.6. However, had we used the threshold of 0.8 these would not have been classified as AAE. For 12 and 14, the general classifier predicted them as `toxic` and the specialized classifier predicted them as `non-toxic`. We hypothesize this is because these two examples are `toxic` to Black people, and the specialized classifier did not have many examples of that type of toxicity in the AAE training dataset. For instance 15, lexical variation may have been used as a common method to avoid moderation detection [67] as `toxic` by the general classifier.

5. **FP: Mislabeled, Toxic Tweets** We found only one false-positive example (instance 4) that was predicted by HxEnsemble to be `toxic` when the annotators did not label it as such. This tweet is directed at someone and harasses them based on a mental-health condition and calls them several derogatory curse words. We found the model

predicted this example correctly while annotation across all datasets included in the aggregate should have labeled it as `toxic` as well.

6. **FP: Use of Pejorative for Third-Parties** For instances 1, 3, 7, and 17, we found that these four examples the authors all referred to a third-party by either using the n-word or using the b-word to speak negatively about them. Instance 1 in particular is a retweeted tweet that is very similar to a popular rap lyric, "I'm just stuntin' on my ex-bitch." [68] This suggests that even in cases that may not be toxic, the model struggles to differentiate between toxicity and use of these words in negative sentiment when used to describe a third-party.

7. **FP: Use of Curse Words in Neutral Context** In examples 6 and 11, the false-positives are resulted from the HxEnsemble unable to understand the Ass Camouflage Construction (ACC) and the n-word to replace an equivalent "guys" commonly used in AAE without implying toxic connotation.

As seen in the above challenges and summarized in Table 5.6, we attributed 35.3% of the AAE misclassified instances by HxEnsemble to incorrect annotations in the datasets. We also observed that 11.8% of the misclassified AAE samples were due to non-AAE samples being misclassified by the dialect model, when the general classifier correctly predicted the texts as toxic. While the subjectivity of some challenge types may mean that there is higher disagreement with the labels provided in the datasets. These findings demonstrated how the efficacy of bias-mitigation strategies in addressing annotation biases may be under-reported using standard classification and fairness metrics. We also note that while the HxEnsemble method was able to effectively mitigate some of the biases towards AAE, it still struggled at times with some AAE characteristics such as ACC and the usage of the n-word. Lastly, we note that effects of incorrect annotations of AAE texts on the performance of the specialized AAE classifier and its ability to remediate the biases more effectively.

Table 5.6: Breakdown of challenges in AAE Hate-Detection

Challenge Type	(n =)	%
FN: Mislabeled, Non-Toxic	4	23.5
FP: Use of Pejorative for Third-Parties	4	23.5
FN: Toxic and False-Positive AAE	3	17.6
FP: Curse Words in Neutral Contexts	2	11.8
FN: Missing Context, Unclear Toxicity	2	11.8
FN: Mislabeled, Non-Targeted Threat	1	5.9
FP: Mislabeled, Toxic	1	5.9

CHAPTER 6

DISCUSSION, LIMITATIONS, AND FUTURE WORK

In our investigation, we evaluated various hate-detection algorithms with a focus on examining the effectiveness of our proposed hierarchical ensemble model and two bias-mitigation algorithms that addressed the racial biases present in toxic language datasets. In our experiments, we observed that our hierarchical ensemble model consistently achieved state-of-the-art classification results while improving upon the fairness metrics we evaluated on. The explanation regularization technique was also able to reduce the biases against AAE authors better than the vanilla-BERT model without a dramatic decrease in classification performance. Across all datasets, the MinDiff regularization framework consistently performed worse than other classifiers, and while it had a closer false-positive error rate balance, this came at the cost of classification accuracy to the SAE authored tweets, rather than an improvement to AAE authored tweets. Since our HxEnsemble proposed framework is model agnostic, combining it with other bias-mitigation techniques to better address the complex underlying reasons for model biases can be investigated in future work.

Table 5.6 suggests that the hierarchical ensemble model increased FNR is strongly correlated with annotations that we deemed as misclassified as `toxic` by annotators. This further strengthens the importance of utilizing additional metrics when evaluating debiasing methods. In fact, when investigating the effects of annotation bias, understanding the prediction disparities and classification inaccuracies provides valuable insight into the fairness of the underlying black-box models. We noted that at minimum, 35.3% of misclassified AAE tweets by our HxEnsemble are mislabeled by original annotators. As a result, the effectiveness of the bias-mitigation strategies is likely a lower-bound estimate and provided empirical evidence that the increased FNR of our hierarchical ensemble model successfully mitigates some of the annotation biases towards AAE authors in toxic-language datasets.

For *DWMW17*, *FDCL18*, and *Toxic* datasets we observed how disparate impact of non-toxic predictions strongly favors SAE, while for toxic prediction AAE are more likely than SAE. Additionally, almost all classifiers have FNR_{SAE} computed larger than FNR_{AAE}, while the FPR_{SAE} is smaller than FPR_{AAE}. While the biases of the dataset are present and perpetuated, HxEnsemble and the explanation regularization methods were able to decrease the disparities in these fairness metrics with minimal impact to classification metrics. This shows that on their own, bias-mitigation strategies are not enough to correct the underlying biased data and to fully address this issue in future work, efforts to relabel the dataset or create a new dataset that is less biased to AAE authors is essential.

We saw that using the aggregated dataset *Toxic* provided mixed results for addressing biases. A decrease in the unfavorable disparate impact metric across all the classifiers is observed in the *Toxic* dataset compared to the *FDCL18*, the largest dataset in the aggregate. However, a slight increase in the FPR_{AAE} and a decrease in the disparate impact of a favorable outcome indicates that using a larger corpus is an insufficient bias mitigation strategy for this issue. Additionally, label inconsistencies in the `hate` annotation resulted in classifiers that had very poor performance and struggled to overcome the class imbalance in *Hate*. Comparatively, using a less stringent definition of `toxic` in the *Toxic* aggregate allowed the classifiers to achieve strong performance. If a supplementary dataset of AAE authors for hate-detection is created to address the racial biases present, ensuring that annotation consistency with other datasets will be a challenging issue to address.

The biases against AAE dialect may be correlated to other historically oppressed groups' dialects, such as LGBTQIA+ dialects [69]. Similarly to AAE, LGBTQIA+ dialects reclaim oppressive identity terms which in-group members use in a neutral or positive context. We also note that the study of biases towards AAE authors may be present in other NLP domains, such as sentiment analysis which is commonly used during hate and toxic language dataset creation. We hypothesize that our proposed HxEnsemble framework can be extended to these domains as we show its effectiveness in ensuring models are better able to

represent the protected groups' target population.

We also want to note the importance of maintaining high-accuracy in the models during the bias-mitigation strategies that are applied to the toxic and hateful language detection domain. In particular, we care about ensuring the models have high-accuracy as the domain context is aimed at protecting not just one protected class. In our case we are studying the issue of racial bias towards African American authors in toxic language datasets. Within the datasets, samples of AAE represent 22% of samples in DWMW17, 1.4% in FDCL18, 4% in Toxic aggregate dataset, and 4.9% in the Hate aggregate. While previous work shows annotation biases towards the AAE group, it's important to note that not all samples in the dataset are mislabeled and that the metrics of model performance are still important when comparing and benchmarking models. These models are able to protect other protected groups that are targeted with online bigotry and harassment when participating in online communities and dialogues.

The choice between general accuracy and improved fairness metrics for a protect group is an ethical trade-off. Bias mitigation algorithms often come at the cost of accuracy to the generalized population. In most literature regarding bias-mitigation algorithms, benchmarking their performance is done against the baseline model and the fairness criteria the algorithm was attempting to improve on. As one of the goals of our HxEnsemble algorithm is to reduce the propagation of these annotation biases while minimizing the impact to model performance, those were the two dimensions we compared when comparing the benchmarked models on. Choosing between the optimization of both accuracy and bias mitigation begins to explore the ethics of equity in the performance of algorithms. For example, is it fair that a social media company will use a worse-performing algorithm that misclassifies more people's post as toxic language so that the error rate is equal to the protected classes' error rate in the algorithm?

Lastly, we echo Blodgett et al.'s [19] sentiment on the progress in bias mitigation within natural language processing to have a more unified and systematic approach to fairness.

Assumptions and definitions of what entails the biases and fairness metrics should be emphasized during the research so that the reproducibility, explainability, and cross-domain applications are able to benefit tremendously.

CHAPTER 7
CONCLUSIONS

In this work, we first explored the current state of research regarding AI biases, hate speech, and biases in hate speech detection and some of the related mitigation strategies. We proposed and evaluated an ensemble framework that leveraged a general toxic language classifier, dialect estimation model, and a specialized AAE classifier to reduce the racial biases in hate and toxicity detection datasets. We evaluated the HxEnsemble, two bias-mitigation algorithms, and common machine-learning classifiers using several fairness metrics and datasets that provided insights into how these models learn and propagate the annotation biases in the underlying datasets. Experiments conducted revealed that across all datasets, classifiers had higher FPR and lower FNR for AAE instances than the SAE instances. Additionally, both favorable and unfavorable prediction biases exist against AAE authors, where the disparate impact score for non-toxic predictions is heavily biased against AAE authors, and predictions for toxic is heavily biased towards AAE authors. Although the data biases are propagated to the models, both our HxEnsemble and the explanation regularization bias-remediation techniques were able to mitigate some of the racial biases with minimal impact on classifier performance.

Using a thorough error analysis, we noted that the challenges in debiasing these datasets resulted from a large portion of misclassified AAE samples. We also presented characteristics of AAE samples that the HxEnsemble framework struggled with, which can further motivate future research on debiasing hate-detection on AAE texts. Future usage of our proposed framework is extensible to other low-resource and biased domains, where it can be combined with other bias-mitigation techniques. Lastly, we demonstrated the effects of label consistency issues on classifier performances with two thresholds of dataset aggregation. We call for future work to create a new toxic language dataset that has AAE samples

labeled by in-group annotators, has cultural training materials available, and/or adds racial priming or utilizes CASM to urge annotators to consider the cultural context of the tweet.

Appendices

APPENDIX A

RESULTS WITH A STRICTER AAE PREDICTION THRESHOLD

In the following tables we show the fairness metrics for the classifiers on each dataset using the threshold of Pr(AAE ≥ 0.8) for texts belonging to the AAE dialect using Blodgett et al.'s model [50]. Please see Section 3.2 for more information.

A score of 0 is used to indicate a missing metric or a division error. For example, in DWMW17 there are no mis-classified non-toxic AAE instances in the dataset, hence a FPR of 0.

Table A.1: DWMW17 Fairness Metrics for Pr(AAE ≥ 0.8) [$n_{\text{AAE}} = 74, n_{\text{AAE}} = 2396$]

Task Name	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.000_{\pm 0.000}$	$1.187_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.022_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.209_{\pm 0.004}$
TF-IDF	$0.220_{\pm 0.076}$	$1.079_{\pm 0.007}$	$0.020_{\pm 0.007}$	$0.029_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.612_{\pm 0.003}$
GloVe	$0.000_{\pm 0.000}$	$1.169_{\pm 0.020}$	$0.000_{\pm 0.000}$	$0.052_{\pm 0.009}$	$0.000_{\pm 0.000}$	$0.424_{\pm 0.045}$
BERT	$0.055_{\pm 0.038}$	$1.196_{\pm 0.010}$	$0.009_{\pm 0.007}$	$0.019_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.114_{\pm 0.014}$
BERT+OC	$0.000_{\pm 0.000}$	$1.222_{\pm 0.007}$	$0.000_{\pm 0.000}$	$0.024_{\pm 0.003}$	$0.000_{\pm 0.000}$	$0.086_{\pm 0.012}$
BERT+SOC	$0.000_{\pm 0.000}$	$1.222_{\pm 0.004}$	$0.000_{\pm 0.000}$	$0.023_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.082_{\pm 0.009}$
BERT+MD	$0.026_{\pm 0.061}$	$1.090_{\pm 0.098}$	$0.005_{\pm 0.013}$	$0.013_{\pm 0.022}$	$0.000_{\pm 0.000}$	$0.606_{\pm 0.416}$
HxEnsemble	$0.056_{\pm 0.039}$	$1.192_{\pm 0.008}$	$0.009_{\pm 0.007}$	$0.018_{\pm 0.001}$	$0.000_{\pm 0.000}$	$0.129_{\pm 0.008}$

Table A.2: FDCL18 Fairness Metrics for Pr(AAE ≥ 0.8) [$n_{\text{AAE}} = 7, n_{\text{SAE}} = 9079$]

Task Name	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.193_{\pm 0.000}$	$3.313_{\pm 0.015}$	$0.000_{\pm 0.000}$	$0.143_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.036_{\pm 0.001}$
TF-IDF	$0.549_{\pm 0.001}$	$2.597_{\pm 0.010}$	$0.333_{\pm 0.000}$	$0.286_{\pm 0.002}$	$0.000_{\pm 0.000}$	$0.036_{\pm 0.000}$
GloVe	$0.191_{\pm 0.001}$	$3.377_{\pm 0.034}$	$0.000_{\pm 0.000}$	$0.217_{\pm 0.005}$	$0.000_{\pm 0.000}$	$0.057_{\pm 0.002}$
BERT	$0.198_{\pm 0.000}$	$3.090_{\pm 0.016}$	$0.000_{\pm 0.000}$	$0.094_{\pm 0.003}$	$0.000_{\pm 0.000}$	$0.044_{\pm 0.001}$
BERT+OC	$0.196_{\pm 0.001}$	$3.152_{\pm 0.023}$	$0.000_{\pm 0.000}$	$0.104_{\pm 0.003}$	$0.000_{\pm 0.000}$	$0.040_{\pm 0.002}$
BERT+SOC	$0.176_{\pm 0.062}$	$3.201_{\pm 0.132}$	$0.000_{\pm 0.000}$	$0.103_{\pm 0.005}$	$0.100_{\pm 0.316}$	$0.040_{\pm 0.003}$
BERT+MD	$0.582_{\pm 0.445}$	$1.531_{\pm 1.647}$	$0.500_{\pm 0.527}$	$0.554_{\pm 0.470}$	$0.100_{\pm 0.316}$	$0.035_{\pm 0.045}$
HxEnsemble	$0.197_{\pm 0.000}$	$3.119_{\pm 0.019}$	$0.000_{\pm 0.000}$	$0.098_{\pm 0.004}$	$0.000_{\pm 0.000}$	$0.042_{\pm 0.001}$

Table A.3: Toxic Fairness Metrics for Pr(AAE ≥ 0.8) [$n_{\text{AAE}} = 70, n_{\text{SAE}} = 13376$]

Task Name	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.022_{\pm 0.000}$	$2.774_{\pm 0.013}$	$0.000_{\pm 0.000}$	$0.172_{\pm 0.002}$	$0.500_{\pm 0.000}$	$0.055_{\pm 0.001}$
TF-IDF	$0.144_{\pm 0.006}$	$2.874_{\pm 0.024}$	$0.101_{\pm 0.005}$	$0.297_{\pm 0.004}$	$1.000_{\pm 0.000}$	$0.066_{\pm 0.002}$
GloVe	$0.056_{\pm 0.012}$	$2.687_{\pm 0.025}$	$0.037_{\pm 0.008}$	$0.245_{\pm 0.002}$	$1.000_{\pm 0.000}$	$0.107_{\pm 0.002}$
BERT	$0.043_{\pm 0.007}$	$2.688_{\pm 0.013}$	$0.000_{\pm 0.000}$	$0.142_{\pm 0.001}$	$0.050_{\pm 0.158}$	$0.046_{\pm 0.001}$
BERT+OC	$0.022_{\pm 0.015}$	$2.726_{\pm 0.041}$	$0.000_{\pm 0.000}$	$0.150_{\pm 0.005}$	$0.500_{\pm 0.333}$	$0.051_{\pm 0.003}$
BERT+SOC	$0.034_{\pm 0.012}$	$2.695_{\pm 0.023}$	$0.000_{\pm 0.000}$	$0.148_{\pm 0.004}$	$0.250_{\pm 0.264}$	$0.052_{\pm 0.003}$
BERT+MD	$0.722_{\pm 0.503}$	$0.397_{\pm 0.515}$	$0.644_{\pm 0.477}$	$0.629_{\pm 0.486}$	$0.300_{\pm 0.483}$	$0.343_{\pm 0.457}$
HxEnsemble	$0.056_{\pm 0.012}$	$2.627_{\pm 0.032}$	$0.007_{\pm 0.008}$	$0.138_{\pm 0.003}$	$0.000_{\pm 0.000}$	$0.053_{\pm 0.003}$

Table A.4: Hate Fairness Metrics for Pr(AAE ≥ 0.8) [$n_{\text{AAE}} = 66, n_{\text{SAE}} = 12323$]

Task Name	DI_{fav}	DI_{unfav}	FNR_{AAE}	FNR_{SAE}	FPR_{AAE}	FPR_{SAE}
N-Gram	$0.980_{\pm 0.007}$	$1.631_{\pm 0.228}$	$0.800_{\pm 0.000}$	$0.735_{\pm 0.008}$	$0.038_{\pm 0.008}$	$0.013_{\pm 0.001}$
TF-IDF	$0.968_{\pm 0.000}$	$3.330_{\pm 0.081}$	$0.800_{\pm 0.000}$	$0.860_{\pm 0.002}$	$0.033_{\pm 0.000}$	$0.004_{\pm 0.000}$
GloVe	$1.019_{\pm 0.008}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.884_{\pm 0.031}$	$0.000_{\pm 0.000}$	$0.011_{\pm 0.005}$
BERT	$0.991_{\pm 0.015}$	$1.154_{\pm 0.258}$	$0.680_{\pm 0.103}$	$0.533_{\pm 0.008}$	$0.043_{\pm 0.011}$	$0.024_{\pm 0.001}$
BERT+OC	$0.969_{\pm 0.022}$	$2.002_{\pm 0.745}$	$0.860_{\pm 0.165}$	$0.722_{\pm 0.037}$	$0.054_{\pm 0.019}$	$0.011_{\pm 0.004}$
BERT+SOC	$0.956_{\pm 0.039}$	$2.317_{\pm 1.316}$	$0.720_{\pm 0.193}$	$0.707_{\pm 0.027}$	$0.059_{\pm 0.029}$	$0.014_{\pm 0.003}$
BERT+MD	$0.400_{\pm 0.516}$	$0.600_{\pm 0.516}$	$0.400_{\pm 0.516}$	$0.400_{\pm 0.516}$	$0.600_{\pm 0.516}$	$0.600_{\pm 0.516}$
HxEnsemble	$1.042_{\pm 0.011}$	$0.210_{\pm 0.204}$	$0.960_{\pm 0.126}$	$0.551_{\pm 0.009}$	$0.008_{\pm 0.009}$	$0.021_{\pm 0.001}$

REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, Jul. 2018.

[2] Anti-Defamation League, "Online hate and harassment: The american experience 2021," Tech. Rep., Mar. 2021.

[3] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678.

[4] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2799–2804.

[5] M. Wich, J. Bauer, and G. Groh, "Impact of politically biased data on hate speech classification," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online: Association for Computational Linguistics, Nov. 2020, pp. 54–64.

[6] B. Kennedy, X. Jin, A. Mostafazadeh Davani, M. Dehghani, and X. Ren, "Contextualizing hate speech classifiers with post-hoc explanation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 5435–5442.

[7] F. Prost, H. Qian, E. H. Chi, J. Chen, and A. Beutel, "Toward a better trade-off between performance and fairness with kernel-based distribution matching," 2019.

[8] M. Halevy, C. Harris, A. Bruckman, D. Yang, and A. Howard, "Mitigating racial biases in toxic language detection with an equity-based ensemble framework," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO '21, –, NY, USA: Association for Computing Machinery, 2021, ISBN: 9781450385534.

[9] I. Bogost, "'artificial intelligence' has become meaningless," *The Atlantic*, Mar. 2017.

[10] A. Howard and J. Borenstein, "The ugly truth about ourselves and our robot creations: The problem of bias and social inequity," *Science and Engineering Ethics*, vol. 24, no. 5, pp. 1521–1536, Sep. 2017.

[11] D. Kumar *et al.*, "Designing toxic content classification for a diversity of perspectives," *arXiv preprint arXiv:2106.04511*, 2021.

[12] D. Chavalarias and J. P. A. Ioannidis, "Science mapping analysis characterizes 235 biases in biomedical research.," *Journal of clinical epidemiology*, vol. 63 11, pp. 1205–15, 2010.

[13] Z. Waseem, "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*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 138–142.

[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Montréal, Québec, Canada: Association for Computing Machinery, 2016, pp. 145–153, ISBN: 9781450341431.

[15] D. U. Patton, W. R. Frey, K. A. McGregor, F.-T. Lee, K. McKeown, and E. Moss, "Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20, New York, NY, USA: Association for Computing Machinery, 2020, pp. 337–342, ISBN: 9781450371100.

[16] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18, Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7, ISBN: 9781450357463.

[17] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, New York, NY, USA: PMLR, Feb. 2018, pp. 77–91.

[18] *Amazon scrapped 'sexist ai' tool*, Oct. 2018.

[19] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 5454–5476.

[20] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "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*, ser. NIPS'16, Barcelona, Spain: Curran Associates Inc., 2016, pp. 4356–4364, ISBN: 9781510838819.

[21] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," *CoRR*, vol. abs/1807.11714, 2018. arXiv: 1807.11714.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[23] A. Guterres. Jun. 2018.

[24] A. Founta *et al.*, "Large scale crowdsourcing and characterization of twitter abusive behavior," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, Jun. 2018.

[25] *Hate speech*, Oct. 2021.

[26] *Twitter's policy on hateful conduct — twitter help*.

[27] *Hate speech policy - youtube help*.

[28] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.

[29] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93.

[30] J. Golbeck *et al.*, "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on Web Science Conference*, ser. WebSci '17, Troy, New York, USA: Association for Computing Machinery, 2017, pp. 229–233, ISBN: 9781450348966.

[31] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the Second Workshop on Language in Social Media*, Montréal, Canada: Association for Computational Linguistics, Jun. 2012, pp. 19–26.

[32] M. L. Williams and P. Burnap, "Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data," *The British Journal of Criminology*, vol. 56, no. 2, pp. 211–238, Jun. 2015. eprint: https://academic.oup.com/bjc/article-pdf/56/2/211/7451319/azv059.pdf.

[33] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," *ArXiv*, vol. abs/1603.07709, 2016.

[34] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "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*, ser. CIKM '12, Maui, Hawaii, USA: Association for Computing Machinery, 2012, pp. 1980–1984, ISBN: 9781450311564.

[35] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," *ArXiv*, vol. abs/1701.08118, 2017.

[36] B. Kennedy *et al.*, *The gab hate corpus: A collection of 27k posts annotated for hate speech*, 2018.

[37] O. de Gibert, N. Perez, A. Garcia-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20.

[38] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, (Proceedings of the First . . .), 2017, pp. 86–95. eprint: http://ceur-ws.org/Vol-1816/paper-09.pdf.

[39] H. Zhong *et al.*, "Content-driven detection of cyberbullying on the instagram social network," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16, New York, New York, USA, 2016, pp. 3952–3958, ISBN: 9781577357704.

[40] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399, ISBN: 9781450349130.

[41] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," *arXiv preprint arXiv:2012.10289*, 2020.

[42] M. ElSherief *et al.*, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 345–363.

[43] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World*

Wide Web Companion, ser. WWW '17 Companion, Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 759–760, ISBN: 9781450349147.

[44] J. Qian, M. ElSherief, E. Belding, and W. Y. Wang, "Leveraging intra-user and inter-user representation learning for automated hate speech detection," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 118–123.

[45] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, pp. 925–945, 2019.

[46] S. S. Mufwene, J. R. Rickford, G. Bailey, and J. Baugh, *African-American English structure, history, and use*. Routledge, Taylor & Francis, 2022.

[47] L. J. Green, *African American English: a linguistic introduction*. Cambridge Univ. Press, 2009.

[48] *Do you speak american . for educators . curriculum . college . aae*, 2005.

[49] A. Jørgensen, D. Hovy, and A. Søgaard, "Learning a POS tagger for AAVE-like language," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1115–1120.

[50] S. L. Blodgett, L. Green, and B. O'Connor, "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*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1119–1130.

[51] D. Preoţiuc-Pietro and L. Ungar, "User-level race and ethnicity predictors from Twitter text," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1534–1545.

[52] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, pp. 67–73, ISBN: 9781450360128.

[53] N. Zueva, M. Kabirova, and P. Kalaidin, "Reducing unintended identity bias in Russian hate speech detection," in *Proceedings of the Fourth Workshop on Online Abuse*

and Harms, Online: Association for Computational Linguistics, Nov. 2020, pp. 65–69.

[54] M. Awal, R. Cao, R. K. Lee, and S. Mitrovic, "On analyzing annotation consistency in online abusive behavior datasets," *CoRR*, vol. abs/2006.13507, 2020. arXiv: 2006.13507.

[55] M. Wich, H. Al Kuwatly, and G. Groh, "Investigating annotator bias with a graph-based approach," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online: Association for Computational Linguistics, Nov. 2020, pp. 191–199.

[56] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–35.

[57] M. Xia, A. Field, and Y. Tsvetkov, "Demoting racial bias in hate speech detection," in *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7–14.

[58] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. Smith, "Challenges in automated debiasing for toxic language detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 3143–3155.

[59] F. Klubicka and R. Fernández, "Examining a hate speech corpus for hate speech detection and popularity prediction," *CoRR*, vol. abs/1805.04661, 2018. arXiv: 1805.04661.

[60] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10.

[61] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[62] L. J. Green, *African American English: A Linguistic Introduction*. Cambridge University Press, 2002.

[63] C. Collins, S. Moody, and P. M. Postal, "An aae camouflage construction," *Language*, vol. 84, no. 1, pp. 29–68, 2008.

[64] J. Sweetland, "Unexpected but authentic use of an ethnically–marked dialect," *Journal of Sociolinguistics*, vol. 6, no. 4, pp. 514–538, 2002.

[65] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017, pp. 1–7.

[66] D. Biddle, "Adverse impact and test validation : A practitioner's guide to valid and defensible employment testing," 2006, "A Gower Book"–Cover.

[67] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury, "#Thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ser. CSCW '16, San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1201–1213, ISBN: 9781450335928.

[68] 2. Savage and M. Boomin, *X*. Jul. 2016.

[69] J. Calder, "Language and sexuality: Language and lgbtq+ communities," in *The International Encyclopedia of Linguistic Anthropology*. American Cancer Society, 2020, pp. 1–7, ISBN: 9781118786093.