AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, Feng Luo
School of Computing, Clemson University, Clemson, USA
{eokpala, lcheng2, nmbwamb, luofeng}@clemson.edu

***Abstract*—Hate speech datasets contain bias which machine learning models propagate. When these models classify tweets written in African American English (AAE), they predict AAE tweets as hate/abusive at a higher rate than tweets written in Standard American English (SAE). This paper assesses bias in language models fine-tuned for hate speech detection and the effectiveness of adversarial learning in reducing such bias. We introduce AAEBERT, a pre-trained language model for African American English obtained by re-training BERT-base on AAE tweets. AAEBERT is used to extract the representation of each tweet in the various hate speech datasets and to classify tweets into two classes - AAE dialect and non-AAE dialect. A three-layer feedforward neural network that takes the representation from AAEBERT and a dialect label as input is used as the adversarial network for debiasing. We evaluate bias in language models fine-tuned for hate speech detection. Then assess the effectiveness of adversarial debiasing in these models by comparing results before and after adversarial debiasing is applied. Analysis reveals that the fine-tuned models are biased towards AAE, and adversarial debiasing is effective in reducing bias.**

***Index Terms*—hate speech detection, language model, BERT, adversarial learning**

I. Introduction

Social media enables instantaneous access to trending topics and news, provides a medium to keep in touch with friends, and means to connect with people outside our network. Social media platforms have made sharing, viewing, and subscribing to content relatively easy. While this is beneficial, some of the contents are hateful, and offensive [1], [2]. Even worse, when such contents are directed to specific groups, it can lead to social unrest [3].

Due to the rise of social media, researchers and social media platforms use automatic systems based on machine learning and deep neural networks to tackle the problem of hate speech detection. Detection methods based on traditional machine learning require feature engineering and do not generalize well to new datasets. Neural network-based word embeddings can automatically learn word representations reducing the amount of feature engineering needed. Most recently, pre-trained language models like the Bidirectional Encoder Representations from Transformers (BERT) [4] have been used to improve the performance of hate detection systems. Despite these successes, these models have a bias problem.

Hate speech detection models learn from annotated datasets and can assimilate the bias in these datasets. Training traditional machine learning classifiers on these datasets results in classifiers that are racially biased towards AAE [5]. These datasets contain texts written in standard English, which differ in terms of syntax, phonology, and lexicon when compared to the AAE variety [6]. BERT [4] based language models fine-tuned on hate speech datasets for hate speech detection also propagate this racial bias [7].

To reduce racial bias, researchers in [8] used adversarial training [9] to demote the racial bias learned by a bidirectional long short term with attention (BLSTMA) model in a two step training procedure. Bias reduction have also been studied in a pre-trained BERT-base model by reweighting hate speech datasets and using the reweighted scores and dataset as model input during fine-tuning [7]. The current approach have focused on recurrent neural network based models [8] and BERT-based pre-trained model [7] without considering other pre-trained models used in hate speech detection.

Pre-training language models with domain-specific corpus have been used to address the problem of models trained on general knowledge corpus [10], [11]. Due to new hateful terms introduced during the pandemic, [12] developed COVID-HateBERT to detect general hate speech and COVID-19 related hate speech. HateBERT [13] was introduced for abusive language detection. AlBERTo [14] and BERTweet [15], are pre-trained language models for Italian and English tweets respectively.

Researchers have argued that the high rate of assigning AAE tweets to negative classes is due to the presence of AAE in the datasets and the high use of words like ”n***a” in the AAE tweets [5]. Given the presence of AAE in hate speech datasets, we introduce AAEBERT, a pre-trained language model trained using AAE tweets. We study its effects and the effectiveness of adversarial learning in reducing racial bias in pre-trained models fine-tuned on hate speech datasets. To accomplish the goal of this work:

1) We fine-tune popular pre-trained language models (BERT, BERTweet, and HateBERT) commonly used in hate speech detection. On eight hate speech detection datasets and assess bias in the models obtained from fine-tuning each pre-trained language model on each dataset.
2) We introduce AAEBERT, a pre-trained language model based on BERT-base. AAEBERT is re-trained using 1.2M African-American English tweets and is a language model for African-American English.
3) We develop a three-layer feedforward neural network as the adversary network, trained during the fine-tuning

process. Finally, we analyze the effect of adversary debiasing in the models obtained by comparing the models before and after applying adversarial debiasing.

II. Related Work

Social media enables communication, instant news, and socialization. Users of social media platforms like Twitter generate an enormous amount of content, some of which are hateful and cannot be efficiently moderated manually. Researchers have explored automatic methods based on machine learning [16] and deep learning techniques [1], [17], [18] to solve this problem.

Fine-tuning transformer-based models such as BERT [4] on downstream tasks have achieved impressive performance. However, they do not perform so well on specialized domains. For example using BERT which was pre-trained on general corpus in biomedical text mining produces undesired results due to the difference in word distribution in both the general corpus and biomedical corpus. To solve this problem, pre-training language models on datasets from specialized domains have been employed. COVID-HateBERT [12] was introduced for detecting Covid-19 related hate speech, ALBERTo-HS [19] for hate speech detection in Italian tweets. BERTweet [15] and ALBERTo [14] are both pre-trained language models for English and Italian tweets. Caselli et. al developed HateBERT [13], a model skewed towards social media and offensive, abusive, and hate related task.

Hate speech detection datasets contain systematic racial bias due to annotation as demonstrated by [5]. Models trained on these datasets automatically inherit the bias. There have been works to mitigate such bias, to mitigate the bias propagated from biased datasets in trained models, researchers in [7] used different fine-tuning strategies to develop a BERT based hate speech detection model. They mitigate bias in the fine-tuned model by employing a re-weighting mechanism that re-weight the training and validation datasets.

Adversarial training championed by [9] have been used to train multiple networks with one network fooling the other to achieve its objective. Xia et. al [8] used adversarial training to mitigate bias towards AAE texts. They train a classifier that learns to detect hate speech while using an adversarial network to prevent the classifier from learning a representation predictive of AAE attribute. They use a bidirectional long short term memory (BiLSTM) with attention as a feature extractor and multilayer perceptrons (MLPs) as classifer and adversary. Model was trained using a two-phase training procedure.

Our work differs from [7] in the following ways: (1) we do not use a re-weighting mechanism on the training and validation sets but use a simple adversarial network to mitigate bias. (2) we introduce AAEBERT a retrained BERT model skewed towards AAE to extract useful represenation used in bias mitigation. (2) we assess and mitigate bias in other pre-trained language models (BERTweet and HateBERT) fine-tuned for hate speech detection, and extend our analysis to more than three datasets. In addition, (1) we do not apply a two-phase training procedure, rather we employ a simple debiasing procedure performed during model fine-tuning. (2) we use AAEBERT to infer AAE dialect instead of using the model introduced by [6].

III. Methodology

This section introduces our solution for debiasing language models fine-tuned on hate speech detection datasets. By exploring the representation of tweets from existing pre-trained language models and the representation of tweets from the AAEBERT model, we introduce a debiasing network with a fusion mechanism capable of reducing bias propagated by pre-trained language models fine-tuned for hate speech detection datasets. The general description of the architecture is shown in Fig. 1. The input is a sequence tweet, the tweet's label, the tweet's AAE dialect label, and a sequence of the tweet's representation from AAEBERT. Specifically, the classifier takes the sequence of a tweet and the tweet's label as input and learns to predict the label. The adversary input is the fusion (concatenation) of a tweet's representation from AAEBERT with the tweet's representation from the classifier and the dialect label (AAE dialect or not) assigned to the tweet by AAEBERT. The adversary learns to predict the dialect label. The network is trained end-to-end and can be divided into three components, the classifier, fusion, and the adversary. We introduce each of these components below.

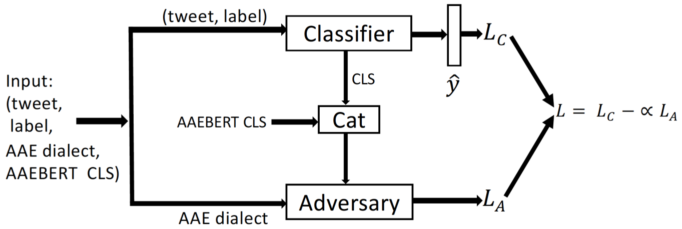

Fig. 1. Illustration of the proposed debiasing architecture

A. AAEBERT

We use the huggingface library [20] to retrain BERT$_{base}$ using the race dataset described below. AAEBERT is optimized with Adam [21], trained for 100 epochs on 1 V100 GPU[1] using a batch size of 64 and a learning rate of 5e-5. AAEBERT is trained using the black-aligned corpus from the race dataset containing 1,288,525 tweets. Masked language modeling was used as the training objective.

B. Classifier

We assess bias by fine-tuning three popular pre-trained language models to detect hate or abusive speech. The models are fine-tuned on datasets annotated for hate speech or abusive language. The following models are considered; First is BERT$_{base}$ [4], pre-trained on BookCorpus and English Wikipedia corpus, and the second is BERTweet [15], pre-trained on a large corpus of tweets. Finally, HateBERT [13]

[1] The V100 GPU is a shared resource with a wall time of 72 hours. It took nine days to complete the training.

pre-trained on Reddit comments from communities banned for writing hateful or offensive comments. The input to the classifier is a tweet sequence and its corresponding label. The first token in the sequence is a special classification token ([CLS]). The [CLS] token in the final hidden layer is used as the representation of the entire sequence and passed to an output layer for classification. The classifier takes as input a sequence of a hateful or offensive tweet and its label and learns to predict the label well. The classifier is trained using the cross-entropy loss.

C. Race dataset

We use the African American English (AAE) dataset developed by [6]. Blodgett et al. [6] created the AAE dataset by collecting tweets from the Gardenhose/Decahose Twitter archive and mapped tweet authors to the demography of the location they lived in using the authors' tweet geolocation. The mapping is done by looking up the US Census block group geographic area from which the message originated and using the race information associated with the block group. They defined four covariates - the percentage of non-Hispanic whites, non-Hispanic blacks, Hispanics, and Asians. They utilized a mixed membership demographic-language probabilistic model, which learns a demographic-aligned language model for each demographic category. The model calculates the posterior proportion of language from each category in each tweet. The dataset contains 59.2 million tweets generated by 2.8 million users.

Following [5] and [7], we create a black-aligned corpus by filtering tweets with an average posterior proportion $>$ 0.80 for the non-Hispanic black category and $<$ 0.10 when the Hispanic and Asian categories are combined—enabling us to obtain tweets written by users who use AAE. Similarly, we create a white-aligned corpus obtained by filtering tweets with an average posterior proportion $>$ 0.80 for the non-Hispanic white category and $<$ 0.10 when the Hispanic and Asian categories are combined. After extraction, we obtained 1,314,176 and 16,077,312 tweets written by non-Hispanic blacks and non-Hispanic whites, respectively. From the black-aligned and white-aligned corpus, we sampled 1000 tweets (999 tweets after preprocessing) used to assess bias in our fine-tuned models. After excluding the sampled tweets and preprocessing, the black-aligned corpus used to train AAEBERT had 1,288,525 tweets.

D. Fusion

The 768 dimensional CLS representation from the last layer of the classifier is concatenated with the CLS representation from AAEBERT, producing a 1536 dimensional vector representation. This representation serves as the input to the adversary. We experimented with different representations as input to the adversary, using the classifier output logits and classifier [CLS] representation separately as input. The fusion between the classifier [CLS] and AAEBERT [CLS] representation performed the best, and only those results are presented because of space constraints.

TABLE I
DATASETS USED IN OUR WORK

Dataset	Count
Waseem and Hovy [24]	10338
Waseem [25]	5988
Davidson et. al [16]	24773
Golbeck et. al [26]	20305
Founta et al. [27]	45549
OffensEval 2019 [28]	14100
AbusEval [29]	14100
HatEval [30]	11991

E. Adversary

The adversary is a three-layer feedforward neural network with a Leaky ReLU activation function. The first and second layer contains 256 and 100 neurons, respectively, and the output layer contains two neurons with softmax function. The adversary learns to predict the AAE/Non-AAE dialect of a tweet and uses the cross-entropy loss. It takes as input the fused representations from the classifier and AAEBERT and a label indicating if a tweet is AAE or Non-AAE dialect. Similar to the setup in [22] and [23], the adversary optimizes the equation $L = L_C - \alpha L_A$ as its objective. The variable L_C is the classifier loss, L_A is the adversary loss, and α is a hyperparameter that controls the rate at which the adversary is maximized, and the classifier minimized. The goal is for the classifier to learn to predict hate speech well and for the adversary to not perform well in predicting AAE dialect from the fused representations.

IV. Experiments

1) Datasets: We use English Twitter datasets for fine-tuning our models and briefly describe each in this section.

Waseem and Hovy [24] collected 136,052 and labeled 16,914 tweets into three categories - racism, sexism, and neither. After preprocessing, 10,338 tweets are obtained.

Waseem [25] investigated the effect of using datasets annotated for hate speech by experts (feminist and anti-racism experts) and amateurs (recruited from CrowdFlower) on classification models. The dataset contains 5,988 tweets after preprocessing and has 4 classes, racism, sexism, racism and sexism, and neither.

Davidson et. al [16] studied the distinction between hate speech and offensive language by extracting 24,802 tweets labeled into three classes, hate, offensive, and none. After preprocessing, the dataset contained a total of 24,773 tweets.

Golbeck et. al [26] developed a hand-labeled dataset of online harassment containing 35,000 tweets with 20,305 tweets remaining after preprocessing.

Founta et. al. [27] sort out to solve the challenge of having different but related labels (hate, offensive, cyberbullying, and aggressive) in hate speech and developed a dataset containing 80K tweets. After rehydrating and preprocessing, we obtained 45,549 tweets labeled as abusive, hateful, spam, or normal. In our experiments, we do not consider the spam class.

OffensEval 2019 [28] uses the Offensive Language Identification Dataset (OLID) [31] which is the main dataset used in the SemEval 2019 Task 6 (OffensEval[2]) competition. The dataset contains 14,100 tweets after preprocessing.

AbusEval, [29] created AbuseEval v1.0 dataset, which is the same dataset created by [31] but with the introduction of new labels (implicit and explicit abuse).

The HatEval dataset is the primary dataset used in the SemEval 2019 Task 5, focusing on detecting hate towards women and immigrants on Twitter [30].

2) Data preprocessing: Before training AAEBERT and fine-tuning our models, we preprocessed our datasets by removing duplicate tweets and tweets with two or fewer words. The dataset was normalized by replacing user mentions with @USER, URLs with URL, and emojis with text representation using the Python emoji package. Additional processing of the dataset included removing emojis and hashtags, converting tweets to lowercase, replacing extra blank spaces with a single space, and removing additional empty newlines.

3) Model fine-tuning: The fine-tuning uses the train and test splits provided in the datasets above. For the datasets without train and test splits, we randomly split the entire corpus into two sets; train - 80% and test - 20%. The hate datasets do not have an AAE dialect label for adversarial training. To provide this label, AAEBERT is utilized to classify hateful tweets into two classes - AAE dialect and non-AAE dialect. We fine-tune each of the classifiers described in Section III-A on each dataset for each fine-tuning configuration. Evaluate the performance of the fine-tuned models and access bias without applying adversarial debiasing (as shown in Fig. 1). Then the process is repeated with adversarial debiasing (as shown in Fig. 1). The bias rate is compared as defined in Section IV-4 with and without adversarial debiasing. To assess bias, we reinitialize our network and we pass the sampled (999 tweets) black-aligned and white-aligned tweets through the classifier (see the top section of Fig. 1). We use $\alpha = 1$, experimented with different values of alpha and discuss its effects in Section V-A.

4) Measuring bias: To assess the rate at which fine-tuned models assign black and white aligned tweets to negative classes and the effects of adversarial debiasing in reducing bias, we calculate the percentage of tweets from each racial group assigned to a negative class i. Let 1 represent a tweet that belongs to a negative class c_i and 0 otherwise. Also, let $P(c_i = 1|black) = P(c_i = 1|white)$ denote the null hypothesis that the probability of a tweet belonging to a negative class is independent of the race of the tweet's author. The null hypothesis is rejected in favor of the second hypothesis that black aligned tweets are assigned to a negative class at a higher rate than white aligned tweet or vice versa. The second hypothesis is defined as: $P(c_i = 1|black) > P(c_i = 1|white)$ or $P(c_i = 1|black) < P(c_i = 1|white)$.

We assess the effectiveness of AAEBERT and adversarial debiasing using the black-aligned and white-aligned tweets randomly sampled from the race dataset. Each tweet is passed through each fine-tuned model to predict its probability of belonging to each class. The assessment is repeated with and without applying adversarial debiasing. For each fine-tuned model, a vector of dimension n (the number of tweets in each race dataset) is created containing the probability p_i of belonging to each class i. The proportion of tweets belonging to class i for the black and white groups are given by $\widehat{p_{iblack}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ and $\widehat{p_{iwhite}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ respectively. If the ratio $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ is greater than 1, then black-aligned tweets are assigned to class i at a higher rate than white-aligned tweets. A t-test between $\widehat{p_{iblack}} = \widehat{p_{iwhite}}$ is conducted. P values < 0.001 is indicated with stars (***) and no stars indicate p values > 0.05 in the result tables in Section V.

TABLE II
EVALUATION RESULTS OF FINE-TUNED MODELS ON EACH HATE SPEECH DATASET WITHOUT APPLYING ADVERSARIAL DEBIASING

Dataset	Model	F1	Precision	Recall
Waseem	BERT	0.403	0.406	0.404
	BERTweet	0.399	0.407	0.402
	HateBERT	0.413	0.415	0.414
	AAEBERT	**0.462**	0.425	0.429
Waseem and Hovy	BERT	0.557	0.566	0.561
	BERTweet	0.551	0.563	0.556
	HateBERT	**0.566**	0.555	0.559
	AAEBERT	**0.566**	0.557	0.561
Davidson et al.	BERT	0.794	0.745	0.763
	BERTweet	**0.803**	0.775	0.787
	HateBERT	0.797	0.745	0.764
	AAEBERT	0.785	0.731	0.749
Golbeck et al.	BERT	0.689	0.618	0.628
	BERTweet	**0.707**	0.628	0.640
	HateBERT	**0.707**	0.626	0.638
	AAEBERT	0.699	0.610	0.618
Founta et al.	BERT	0.761	0.712	0.726
	BERTweet	0.766	0.721	0.733
	HateBERT	**0.776**	0.712	0.727
	AAEBERT	0.775	0.707	0.723
OffensEVal 2019	BERT	0.828	0.810	0.817
	BERTweet	0.813	0.798	0.804
	HateBERT	**0.832**	0.800	0.813
	AAEBERT	0.815	0.777	0.792
AbusEval	BERT	0.571	0.528	0.533
	BERTweet	0.528	0.542	0.533
	HateBERT	**0.688**	0.533	0.537
	AAEBERT	0.600	0.521	0.528
HatEval	BERT	0.682	0.574	0.457
	BERTweet	**0.702**	0.612	0.522
	HateBERT	0.684	0.585	0.480
	AAEBERT	0.687	0.581	0.470

V. RESULTS

This section discusses the results of our experiments. The goal is to study racial bias in fine-tuned language models, the effectiveness of adversarial training, and AAEBERT representation in mitigating racial bias in these models.

We evaluate the performance of models fine-tuned for hate speech detection without adversarial training, assess bias in these models, and evaluate the effectiveness of adversarial training in bias mitigation. Table II shows the performance evaluation result of the models without adversarial debiasing in terms of macro averaged F1 score, precision, and recall. From Table II, we observe that the models fine-tuned on the Waseem [25] dataset had the least performance compared to models fine-tuned on other datasets in terms of F1 score. The

[2] https://competitions.codalab.org/competitions/20011

TABLE III

RACIAL BIAS ANALYSIS OF FINE-TUNED BERT MODELS. SHOWING RESULT WITH AND WITHOUT ADVERSARIAL DEBIASING

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.074	0.056	2.099	0.0360	1.328	0.120	0.098	4.482	***	1.224
	Implicit	0.035	0.025	3.61	***	1.398	0.044	0.039	4.94	***	1.145
OffensEval	Offensive	0.346	0.204	8.707	***	1.692	0.358	0.259	8.854	***	1.381
HatEval	Hate	0.147	0.089	5.619	***	1.649	0.282	0.196	12.131	***	1.439
Davidson et al.	Hate	0.018	0.014	1.733		1.313	0.083	0.082	1.657		1.016
	Offensive	0.357	0.139	13.423	***	2.572	0.498	0.316	16.618	***	1.576
Founta et al.	Hate	0.070	0.033	5.192	***	2.144	0.051	0.044	11.701	***	1.153
	Abuse	0.257	0.125	8.651	***	2.055	0.235	0.154	9.638	***	1.522
Waseem and Hovy	Racism	0.004	0.004	-0.749		**0.971**	0.004	0.004	2.1	0.0360	1.014
	Sexism	0.209	0.065	11.987	***	3.234	0.155	0.090	10.24	***	1.734
Waseem	Racism	0.017	0.004	15.69	***	4.069	0.013	0.011	14.79	***	1.164
	Sexism	0.111	0.016	14.284	***	6.763	0.068	0.045	14.542	***	1.511
	Racism and Sexism	0.008	0.003	14.365	***	2.989	0.007	0.006	14.726	***	1.211
Golbeck	Harassment	0.093	0.068	4.45	***	1.365	0.226	0.203	11.432	***	1.114

TABLE IV

RACIAL BIAS ANALYSIS OF FINE-TUNED BERTWEET MODELS. SHOWING RESULT WITH AND WITHOUT ADVERSARIAL DEBIASING

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.083	0.058	2.747	0.006	1.426	0.133	0.127	3.344	***	1.047
	Implicit	0.015	0.016	-0.677		**0.951**	0.050	0.049	3.411	***	1.019
OffensEval	Offensive	0.358	0.208	9.676	***	1.724	0.354	0.295	9.322	***	1.200
HatEval	Hate	0.107	0.053	6.352	***	2.022	0.265	0.214	8.93	***	1.238
Davidson et al.	Hate	0.020	0.015	1.928		1.321	0.059	0.063	-21.651	***	**0.929**
	Offensive	0.476	0.179	17.681	***	2.653	0.744	0.682	22.506	***	1.091
Founta et al.	Hate	0.096	0.038	7.857	***	2.527	0.044	0.043	16.447	***	1.036
	Abuse	0.302	0.148	10.13	***	2.045	0.215	0.199	16.111	***	1.079
Waseem and Hovy	Racism	0.004	0.004	3.581	***	1.071	0.011	0.011	5.958	***	1.028
	Sexism	0.080	0.040	5.34	***	1.98	0.174	0.161	5.25	***	1.078
Waseem	Racism	0.009	0.006	7.389	***	1.538	0.016	0.015	5.827	***	1.043
	Sexism	0.046	0.019	6.483	***	2.492	0.060	0.056	5.148	***	1.074
	Racism and Sexism	0.006	0.004	6.943	***	1.436	0.013	0.013	5.427	***	1.038
Golbeck	Harassment	0.091	0.079	2.227	0.026	1.144	0.238	0.235	3.831	***	1.014

AAEBERT model had the best F1 score. The low performance of the models fine-tuned on the Waseem [24] dataset is as expected given the dataset size. The AAEBERT and HateBERT models performed the best on the Waseem and Hovy [24] dataset. BERTweet obtained the best F1 score in the Davidson et al. [16], Founta et al. [27], and HatEval [30] datasets and tied with HateBERT on the Golbeck et al. [26] dataset. HateBERT outperformed other models in the OffensEval 2019 [28] and AbusEval [29] datasets and is as competitive as AAEBERT on the Waseem and Hovy [25] dataset.

Tables III, IV, and V show the results of racial bias in these models. From the "without adversarial debiasing" column in Table III, we observe that the $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of the negative classes of the datasets we fine-tuned on is > 1, indicating that the models are biased as they assigned black-aligned tweets to negative classes at a higher rate than white-aligned tweets. The exception to this is in the racism class of the Waseem and Hovy [24] dataset with a $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of 0.971. Indicating the opposite, the model assigned white-aligned tweets to the racism class at a higher rate than black-aligned tweets. When adversarial debiasing is applied as seen in the "with adversarial debiasing" column, bias is reduced to some extent. The highest bias reduction occurred in the racism, sexism, and racism and sexism classes of the Waseem dataset [25] with $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values of 2.9%, 5.2%, and 1.7% respectively. The $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ value of the racism class of Waseem and Hovy [24] became slightly greater than 1 which could indicate that adversarial debiasing is trying to equalize $\widehat{p_{iblack}}$ and $\widehat{p_{iwhite}}$.

Table IV shows the results of fine-tuning BERTweet on different datasets. Similar to the fine-tuned BERT models in Table III, without adversarial debiasing, $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values of the classes in the datasets are greater than 1 indicating that fine-tuned BERTweet models are biased towards black-aligned tweets. Except for the implicit class of the AbusEval dataset [29]. With adversarial debiasing applied, we observe reduction in $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ values. The $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$ are also slightly greater than 1 indicating that adversarial debiasing is effective in reducing bias and in achieving equality between $\widehat{p_{iblack}}$ and $\widehat{p_{iwhite}}$ equal. With the exception of the models fine-tuned on the OffensEval [28] and HatEval [30] datasets though bias is reduced from 1.7 to 1.2 and from 2.022 to 1.238 respectively. After adversarial debiasing, the model obtained from fine-tuning BERTweet on the Davidson et al. [16] dataset obtained $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} < 1$ in the hate class. Indicating that the model became

TABLE V

RACIAL BIAS ANALYSIS OF FINE-TUNED HATEBERT MODELS. SHOWING RESULT WITH AND WITHOUT ADVERSARIAL DEBIASING

		Without adversarial debiasing					With adversarial debiasing				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.079	0.063	1.830		1.254	0.119	0.103	2.809	0.005	1.154
	Implicit	0.027	0.025	0.949		1.078	0.045	0.042	4.498	***	1.068
OffensEval	Offensive	0.346	0.218	8.439	***	1.587	0.370	0.305	7.157	***	1.215
HatEval	Hate	0.155	0.091	6.665	***	1.690	0.305	0.236	10.017	***	1.290
Davidson et al.	Hate	0.016	0.014	1.2		1.162	0.074	0.074	-0.592		**0.995**
	Offensive	0.369	0.139	14.37	***	2.651	0.527	0.339	17.679	***	1.555
Founta et al.	Hate	0.070	0.029	6.171	***	2.468	0.054	0.046	13.054	***	1.163
	Abuse	0.274	0.134	9.229	***	2.051	0.262	0.17	10.483	***	1.542
Waseem and Hovy	Racism	0.004	0.003	5.776	***	1.202	0.004	0.003	7.618	***	1.089
	Sexism	0.145	0.055	9.415	***	2.636	0.147	0.102	8.491	***	1.433
Waseem	Racism	0.019	0.005	11.523	***	3.708	0.012	0.011	11.295	***	1.125
	Sexism	0.052	0.014	9.461	***	3.652	0.058	0.046	10.701	***	1.255
	Racism and Sexism	0.009	0.003	10.498	***	2.533	0.008	0.007	10.701	***	1.124
Golbeck	Harassment	0.084	0.070	2.629	0.009	1.203	0.251	0.249	5.139	***	1.007

more biased towards white-aligned tweets than black-aligned tweets. The top 3 reduction in bias occurred in the offensive class of Davidson et al. [16], hate class of Founta et al. [27], and sexism class of Waseem [24] datasets with a $\frac{\widehat{p_{iblack}}}{p_{iwhite}}$ value reduction of 1.5%, 1.49%, and 1.41% respectively.

Results of fine-tuning the HateBERT model is shown in Table V. In all classes, $\frac{\widehat{p_{iblack}}}{p_{iwhite}} > 1$ indicating that fine-tuned HateBERT models are biased. After adversarial debiasing is introduced, bias is reduced in all classes except for the Hate class of the Davidson et al. [16] dataset with a $\frac{\widehat{p_{iblack}}}{p_{iwhite}}$ value of 0.995. Indicating that the fine-tuned HateBERT model on the Davidson et al. [16] dataset assigned white-aligned tweets to the hate class at a higher rate than black-aligned tweets. The top three highest reduction of $\frac{\widehat{p_{iblack}}}{p_{iwhite}}$ values occurred in the Waseem [24] dataset with 2.5%, 2.3%, and 1.4% reduction in the racism, sexism, and racism and sexism classes respectively.

A. *Effect of α value*

The α value determines the rate at which bias is reduced. This section investigates the effects of α on bias reduction and performance when adversarial debiasing is applied. For each of the pre-trained language models, we evaluate the models obtained from fine-tuning when adversarial debiasing is applied using α values of 0.01, 0.05, 0.09, 0.1, 0.3, 0.5, 0.8, and 1. We do not show the results of these experiments due to space. The following are the conclusions from the experiments. First, when $\alpha < 0.5$, performance is as competitive as when the pre-trained models are fine-tuned without adversarial debiasing. There is a reduction in performance when $\alpha \geq 0.5$. Having $\alpha \geq 0.8$ leads to a better reduction in bias. We use $\alpha = 1$ in fine-tuning pre-trained models when adversarial debiasing is applied because it showed a better reduction in bias, and its performance is as competitive as $\alpha = 0.8$. Second, bias reduction is slow and starts improving from $\alpha \geq 0.8$.

VI. CONCLUSION

This paper presents an adversarial debiasing network for debiasing BERT-based hate speech detection models. We introduced AAEBERT, a pre-trained language model based on BERT-base for African-American English, and assessed bias in three pre-trained language models used in hate speech detection. We assessed the effect of adversarial debiasing in reducing bias by utilizing tweet representations from AAEBERT and fine-tuned pre-trained language models. Bias assessment of fine-tuned models without adversarial debiasing indicates that fine-tuned models are more biased towards AAE than Standard American English (SAE). Analyses of fine-tuning with and without adversarial debiasing, show that adversarial debiasing is effective in reducing bias by achieving $\frac{\widehat{p_{iblack}}}{p_{iwhite}} \approx 1$ in some models. Equalization is observed in fine-tuned pre-trained models when α approaches 1.

ACKNOWLEDGMENT

This work is supported in part by National Science Foundation (NSF) under the Grant No. 2228616, 2114920, and 2031002.

REFERENCES

[1] 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*, 2017, pp. 759–760.

[2] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & internet*, vol. 7, no. 2, pp. 223–242, 2015.

[3] M. Fahim and S. S. Gokhale, "Detecting offensive content on twitter during proud boys riots," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021, pp. 1582–1587.

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

[5] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," *arXiv preprint arXiv:1905.12516*, 2019.

[6] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case

study of african-american english," *arXiv preprint arXiv:1608.08868*, 2016.

[7] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on bert model," *PloS one*, vol. 15, no. 8, p. e0237861, 2020.

[8] M. Xia, A. Field, and Y. Tsvetkov, "Demoting racial bias in hate speech detection," *arXiv preprint arXiv:2005.12246*, 2020.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[11] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[12] M. Li, S. Liao, E. Okpala, M. Tong, M. Costello, L. Cheng, H. Hu, and F. Luo, "Covid-hatebert: a pretrained language model for covid-19 related hate speech detection," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021, pp. 233–238.

[13] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," *arXiv preprint arXiv:2010.12472*, 2020.

[14] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, and V. Basile, "Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets," in *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, vol. 2481. CEUR, 2019, pp. 1–6.

[15] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[16] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.

[17] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European conference on information retrieval*. Springer, 2018, pp. 141–153.

[18] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European semantic web conference*. Springer, 2018, pp. 745–760.

[19] M. Polignano, P. Basile, M. De Gemmis, and G. Semeraro, "Hate speech detection through alberto italian language understanding model." in *NL4AI@ AI* IA*, 2019, pp. 1–13.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, 2018.

[23] X. Han, T. Baldwin, and T. Cohn, "Diverse adversaries for mitigating bias in training," *arXiv preprint arXiv:2101.10001*, 2021.

[24] 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*, 2016, pp. 88–93.

[25] 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*, 2016, pp. 138–142.

[26] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, R. K. Gnanasekaran, R. R. Gunasekaran *et al.*, "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on web science conference*, 2017, pp. 229–233.

[27] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[28] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," *arXiv preprint arXiv:1903.08983*, 2019.

[29] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in *Proceedings of the 12th language resources and evaluation conference*, 2020, pp. 6193–6202.

[30] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti *et al.*, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019, pp. 54–63.

[31] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," in *Proceedings of NAACL*, 2019.