Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification

Camille Harris
Georgia Institute of Technology
Atlanta, GA, USA
charris320@gatech.edu

Matan Halevy
Georgia Institute of Technology
Atlanta, GA, USA
matan@gatech.edu

Ayanna Howard
The Ohio State University
OH, USA

Amy Bruckman
Georgia Institute of Technology
Atlanta, GA, USA

Diyi Yang
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

Language usage on social media varies widely even within the context of American English. Despite this, the majority of natural language processing systems are trained only on "Standard American English," or SAE, the construction of English most prominent among white Americans. For hate speech classification, prior work has shown that African American English (AAE) is more likely to be misclassified as hate speech. This has harmful implications for Black social media users as it reinforces and exacerbates existing notions of anti-Black racism. While past work has highlighted the relationship between AAE and hate speech classification, no work has explored the linguistic characteristics of AAE that lead to misclassification. Our work uses Twitter datasets for AAE dialect and hate speech classifiers to explore the fine-grained relationship between specific characteristics of AAE such as word choice and grammatical features and hate speech predictions. We further investigate these biases by removing profanity and examining the influence of four aspects of AAE grammar that are distinct from SAE. Results show that removing profanity accounts for a roughly 20 to 30% reduction in the percentage of samples classified as 'hate' 'abusive' or 'offensive,' and that similar classification patterns are observed regardless of grammar categories.

CCS CONCEPTS

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

KEYWORDS

Natural Language Processing, Linguistics, Fairness, African American English, Hate Speech, Social Media

ACM Reference Format:
Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the Role of Grammar and Word Choice in Bias Toward African American English (AAE) in Hate Speech Classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533144

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533144

1 INTRODUCTION

Language usage on social media differs widely even within the context of American English speakers. Dialect and word choice varies widely by location, age, ethnicity, and online community. However, the vast majority of natural language processing systems rely mainly on "Standard American English" (SAE), the dialect most commonly used among white Americans [32]. Prior works have explored how this impacts performance of these systems on African American English (AAE) text. In particular, prior works provide evidence that hate speech classification systems and sentiment analysis tend to classify AAE text as more likely to be hate speech or another form of toxic speech [8, 17, 22, 35]. This issue is particularly dangerous as it exacerbates existing notions of anti-Black racism that frame Black people as aggressive and inherently dangerous.

The majority of prior works in this domain focus on identifying these biases or proposing various strategies to mitigate bias, the majority of which, rely on educating data annotators on AAE and generally improving data by including more AAE samples [8, 17, 22, 35]. Identifying and providing evidence of these biases and proposing mitigation strategies are both important steps in reducing the harm towards Black people perpetuated by these systems. Our work takes a unique approach that adds depth to the existing work on this topic. We expand on exploring the hate speech classifications of AAE text by focusing on the specific aspects of grammar and word choice. Few prior efforts explore how specific characteristics of AAE text may correlate with higher or lower rates of these types of misclassification. We hope by exploring how word choice and grammar patterns may impact the classification, this work can inform future bias-mitigation strategies, especially those that do not rely on data collection and annotation alone.

In this work, we focus on investigating the relationship between AAE and hate speech classification. First we seek to verify the relationship between hate speech, offensive speech, and abusive speech classifications for AAE. Then we explore several linguistic aspects of AAE tweet text including word choice and grammar features to understand which aspects of the dialect most strongly relate to the way it is classified. Concretely, we frame our work around two primary research questions:

(1) How strongly does use of swear words or "offensive language" impact the hate speech classification of AAE text?
(2) How do grammatical patterns of AAE tweets impact the hate speech classification of AAE text?

We set up our experiments by training two hate speech classification models and applying them to a dataset of AAE text. Then we answer RQ1 by first searching our AAE dataset for samples containing offensive language and analysing the predictions on this corpus. Next, we replace swear words with a semantically equivalent non-offensive word, screen our samples for similar meaning, and then apply the same classification system to our censored text. To answer RQ2, we develop an AAE-Like Grammar Dictionary that captures words and phrases that are common within specific grammar sub-categories of AAE: auxiliary verbs, aspectual markers, preverbal markers and syntactic and morphosyntactic properties. We give some background information on the AAE dialect in Section 4, and explain the AAE-Like Grammar Dictionary in full in Section 5.4. Our findings suggest that controlling for swear words has a significant impact on reducing the bias towards AAE, and that the four grammar categories we examined appear to have little impact. The implications of these findings are that future hate speech classification systems should rely less strongly on individual words and perhaps consider the identity of the speaker or context of the text in classifications. Further, we discuss the importance of reducing this disparity, not only due to the racist, anti-Black implications of classifying benign AAE text as hate speech, but also that these misclassifications may obscure true in-group hate speech directed towards Black people with additional marginalized identities.

2 DEFINITIONS OF HATE SPEECH TERMS

Our work relies on several different terms from hate speech classification literature. These terms are defined differently across social media platforms and across literature that seeks to label and regulate online speech. Here we will discuss the definitions and uses of these terms as well as how we use them in our work.

"Hate speech" differs in exact definition across the literature but there are some common components of how it is defined. Generally, hate speech is defined as any speech that includes violence, emotional harm, or that is derogatory or humiliating towards others [9, 13]. Hate speech definitions place an emphasis on speech that is harmful to a group on individual on the basis of marginalized identities such as racial group, religion, ethnicity, nationality, sexual orientation, disability, or gender [9, 13, 23]. While hate speech is a large concern for researchers and social media sites alike, there are other harmful forms of speech online that fall into other categories of toxic language. In our work, "hate speech" is the most severe form of toxic language.

Founta et al. propose a definition of "abusive language" as a category of online language that is harmful but distinct from and less severe than hate speech [13]. The definition can be summarized as any language that is hurtful or rude, especially language that includes profanity and language used to degrade others. The most obvious distinction from hate speech is that abusive language is generally derogatory or rude, while hate speech often targets a specific marginalized group or identity and often includes violence. For the purposes of our work, the abusive speech classification is for language that is less severe than hate speech, but still potentially harmful. Founta et al. also explains that abusive language is rare, with it making up .1 to 3% of data, and that is highly correlated with offensive language by human annotators even when clear distinct definitions are provided [13].

Davidson et al. [9] propose a definition of "offensive language" as a category of online language that is distinct from hate speech and much less severe. Offensive language is any speech that includes profanity or other offensive terminology. Like abusive language, offensive language is distinct from hate speech because it does not necessarily include violence, threats, or attacks on a specific protected category, and is considered less severe than "hate speech" [9, 13]. The definitions of abusive language and offensive language are similar across the literature, however the key difference is offensive language includes any text regardless of harmfulness rudeness that includes offensive words (such as any profanity).

One complication with defining "offensive language" for natural language systems is that what terms are considered offensive depends heavily on context. Words that are considered very offensive within one community may be considered much less offensive or normal in another. This issue presents with the relationship of AAE text and offensive speech classification. While the "n slur" is considered hate speech when used by non-Black individuals, within the Black community the term is often reclaimed and used in a non-derogatory way, especially within inter-community spaces such as within "Black Twitter" [34]. Similarly, individuals of many different marginalized identities reclaim the slurs and negative epithets used against them. This particular issue was a motivation for Davidson et al. [9] to propose the definition for offensive speech: to distinguish offensive language that is not necessarily hateful such as use of these reclaimed words, from hate speech. Generally speaking, when natural language systems rely simply on word choice rather than dialect, community, and context, marginalized communities can be punished for the practice of reclaiming these words. We discuss in detail in Section 6.

Toxic language is an umbrella term that includes identity-based attacks, bullying, trolling, threats of violence, and sexual harassment. Hate speech along with any other online harassment is included under toxic speech [23]. In our work, we use "toxic speech" as an umbrella term that encompasses "hate speech", "abusive speech", and "offensive speech."

3 RELATED WORK

Our work focuses on understanding the relationship between hate speech classification and AAE as an under-served and under-studied dialect of American English. In general, accurately classifying English text as hate speech is an important topic within natural language processing research. More general solutions to this include Hate Base [39] which relies on a corpus of words heavily associated with hate speech to identify hate speech. Another common approach is to leverage block lists of specific accounts such as with BlockTogether [18] to preemptively block accounts that have tweeted hate speech or other toxic language. However, these solutions do not offer much nuance. As explored by Jhaver et al., while blocklists can be helpful for those who experience continual

harassment, they do not reduce harassment entirely. Furthermore many users placed on block lists feel that they were added unfairly and have little opportunity to be removed [19].

For further nuance and complexity, scholars rely on machine-learning-based methods to classify hate speech. For example, Waseem and Hovy (2016) create a model that relies on a collection of predictive features of hate speech [42]. Davidson et al. [9] proposes a model which detects offensive speech and hate speech as two distinct categories (per definitions in section 2). These works are part of a large body of work that addresses hate speech classification with classical machine learning models (such as reinforcement learning, support vector machines, etc). As explained by Schmidt et al., [36], classical machine-learning models for this task tend to utilize a specific set of features. These include, simple surface features (such as word and character n-grams) [4, 5, 9, 27, 40, 42], word generalization [11, 27], sentiment analysis [40], lexical resources [4, 27], linguistic features (such as word dependencies or part of speech information) [4, 5, 9, 27], and meta information (such as user information or post history) [42].

More recent work has focused on utilizing deep learning rather than classic machine-learning methods for this task. One such example is the model proposed by Founta et al. which leverages features that have been successful for classical methods including user and text data for a Recurrent Neural Network [12]. Siddiqua et al. [37] combine several deep learning techniques and relies on three pre-trained models for feature encoding, BERT, DeepMoji, and InferSent. These and other deep learning models for this task mostly rely on a few popular approaches including Convolutional Neural Nets [1], Recurrent Neural Nets [30], Bi-LSTM [33], and word embeddings (ie BERT, Word2Vec, and TF-IDF Vectorizor), many of which combine these techniques [1].

Recently scholars have begun devoting more attention to identifying and attempting to address bias in natural language systems. Prior work has noted biases towards AAE and towards African American names compared to European American names in sentiment analysis [16, 22] and in English language identification applications [2]. Similar biases were found towards AAE and towards Black twitter users (not necessarily using AAE) in hate speech classification systems [8, 35]. Sap et al. also propose the approach of racial priming, or educating data annotators on racial and dialectical differences before annotating training data, and show that this approach reduces bias [35]. This is consistent with prior work that notes the difference annotation expertise makes in model performance [41]. Other bias mitigation strategies have been proposed in the literature. For example, predicting users' demographic information [31] to inform other predictions such as offensive speech. Zhou et al. [44] shows that common debiasing techniques are not as effective at mitigating bias towards AAE as de-biasing the training data. Halevy et al. proposes an framework that utilizes a specialized classifier trained on AAE text to mitigate bias towards AAE [17].

While many of these works show evidence of the bias towards AAE in various text classification systems and propose mitigation strategies, none of these works explore the specific attributes of AAE that contribute to these biases. Our work expands on these works by exploring specific and granular linguistic aspects of AAE text and how they contribute to the biases. Our work makes a novel contribution by exploring how the use of profanity and use of specific grammar properties unique to the dialect affect the classifications of AAE text.

4 BACKGROUND ON AFRICAN AMERICAN ENGLISH

African American English, also referred to as African American Vernacular English (AAVE) or Ebonics, is an English variation created and used by Black people in the United States. For the sake of simplicity we refer to AAE as a dialect in this work, however, we acknowledge that debates exist among linguists about weather AAE is better classified as a dialect, sociolect, or separate language in it of itself. Nonetheless, due to a long history and culture of anti-Black racism in the United States, African American English has long been criticized as an improper, incorrect, or negative form of English[43]. For many years, it was not recognized as a legitimate dialect, and it took predominately white academic institutions studying AAE to recognize it as a legitimate dialect with its own structure and grammar rules [21, 43]. Furthermore African American English is often categorized with the same racist stereotypes that are applied to Black people as a whole, such as angriness, danger, and aggression. Many of the bias issues surrounding AAE and natural language processing systems derive from and directly support these notions and stereotypes of anti-Black racism. For example, prior works on sentiment analysis have found that use of African American names can also cause a more negative sentiment compared to traditionally European American names [22]. Further, one study found that AAE in general fares more negatively in sentiment analysis, and is more likely to be identified as angry or fearful sentiment than a Standard English equivalent sentence [16]. Understanding this historical and cultural context is crucial to understanding the issue of hate speech misclassification of AAE. While in this work we search for specific aspects of AAE that may contribute more or less strongly to this bias issue, it is important to note that historically AAE as a whole, as with most things heavily associated with Black people, has been attributed to many negative notions and stereotypes.

One hypothesis of this work is that normalizing for swear words in hate speech and toxic language detection can significantly reduce bias towards AAE tweets. This is informed by an analysis of the dialect itself and exploration of the Twitter AAE dataset. Most commonly, misclassifications of AAE arise with swear words that are more common in this dialect than other English dialects. In particular, the n-word is widely understood as a reclaimed slur and therefore derogatory when used by non-Black people, but non-derogatory when used by Black people [34].

Furthermore AAE makes use of a grammatical construction known as "Ass Camouflage Construction" or ACC [6], in which the word "ass" is commonly used to dramatize and place emphasis in sentences. Where this may be considered a more harsh swear word within white American English contexts it is a relatively common practice across regional variations of AAE. It is likely that with AAE being underrepresented in social media data, the n-word is categorized as derogatory regardless of who uses it. Similarly while ACC grammar convention is common within Black communities, within a white American English context it could be interpreted in a much more derogatory sense. These instances as well as other swear words and word choice conventions that are more common

in AAE than SAE likely contribute largely to the misclassification [6].

5 METHODS

5.1 Datasets

We rely on three publicly available datasets of Twitter data for this project. We use two toxic language datasets DWMW17 [8] created by Davidson et al. and FDCL18 by Founta et al.[13]. Both are twitter hate speech datasets that categorize tweets by the various hate speech terms discussed in section three. FDCL18 categorizes tweets as "abusive" "hate speech" "normal" or "spam" while DWMW17 categorizes tweets by "offensive" "hate speech" or "normal." We use these two datasets to train separate BERT language classification models which we discuss further in the next section. We also used Blodgett et al.'s AAE twitter dataset [3] as our AAE data.

For the publicly available versions of all three datasets, tweets are stored by tweet ID. So, we used the Twitter API to collect the tweet text. Many of the tweets in the original versions of each dataset were deleted, and couldn't be accessed using the Twitter API. Consequently, for each of these we used a subsample of the original dataset from the tweets that could be collected. We use 24783 tweets from DWMW17 to train one BERT based model, 50487 tweets from FDCL18 to train another BERT based model. For the purposes of our analysis, we also further limited the dataset from Blodgett et al. to tweets with a high likelihood of AAE dialect as predicted by their model (.9 or above prediction). This left us with 50,000 tweets from the dataset for our analysis. We summarize the details of each dataset in Table 1.

5.2 Models

To evaluate the relationship between AAE and hate speech, we create BERT-based hate speech classification models and a BERT-based dialect classification model. We do this by training and testing BERT models on the DWMW17 and FDCL18 hate speech datasets. BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model which uses a transformer to learn contextual relationships between words in a text and is considered state of the art for the majority of natural language applications [10]. For the hate speech classification models we use the labels of "offensive speech," "hate speech," and "normal" for the DWMW17 model, and "abusive speech," "hate speech," "spam," and "normal" for the FDCL18 model following the definitions provided in Section 3. To calculate correlation for a sample of tweets, we use the AAE classifier from Blodgett et al. to get the estimated dialect of the tweet, then use the DWMW17 and FDCL18 Hate speech models to classify the hate speech category.

5.3 Word Choice Analysis

We answer RQ1 by testing for the relationship between hate speech labels, AAE, and swear words. We do this by first training word2vec [26] on Blodgett AAE Twitter dataset [3]. We also use the Linguistic Inquiry and Word Count (LIWC2007) dictionary [29]. LIWC2007 is a dictionary of words and word stems categorized by sub-categories of text. We use the "Swear" category of LIWC2007, as well as a list of swear words sourced from sample tweets, to create a dictionary of swear words and replacement words with similar meaning. We then use cosine similarity substitute swear words in the AAE tweet samples with the words with similar meaning. This yielded roughly 6,000 tweets that were successfully reworded. We use the models in Section 5.2 to classify the original and reworded tweets.

5.4 Grammar Analysis

To answer RQ2, our work proposes an AAE-like Grammar Dictionary Similar in structure structure of the LIWC dictionary, our dictionary focuses on 4 key categories of grammatical patterns in AAE. We choose to focus on these categories as they are commonly used across regional variations of AAE. Within each category, we create a list of words and phrases common to each grammar category. We limit this list to words and phrases that are uncommon, grammatically incorrect, or nonexistent in Standard American English. It is important to note that we refer to this dictionary as "AAE-like" because it is not all encompassing of African American English as a dialect. There are many grammar patterns within AAE that are not represented in this dictionary. This is because as an initial effort, we limit our dictionary to simple words and phrases, which significantly limits our ability to capture more nuanced grammar rules. Furthermore, AAE differs drastically by region and generation [20, 21]. Our dictionary focuses on grammar conventions that are widely applicable to different regional variations of the dialect; it also focuses on aspects of the dialect that are more common among African American Generation X, Millennials, and Generation Z as they are more within the demographics of Twitter users. Additionally, we note that AAE is traditionally a spoken dialect, and the word choice and spellings may differ slightly in an online context. We provide this dictionary in Table 2.

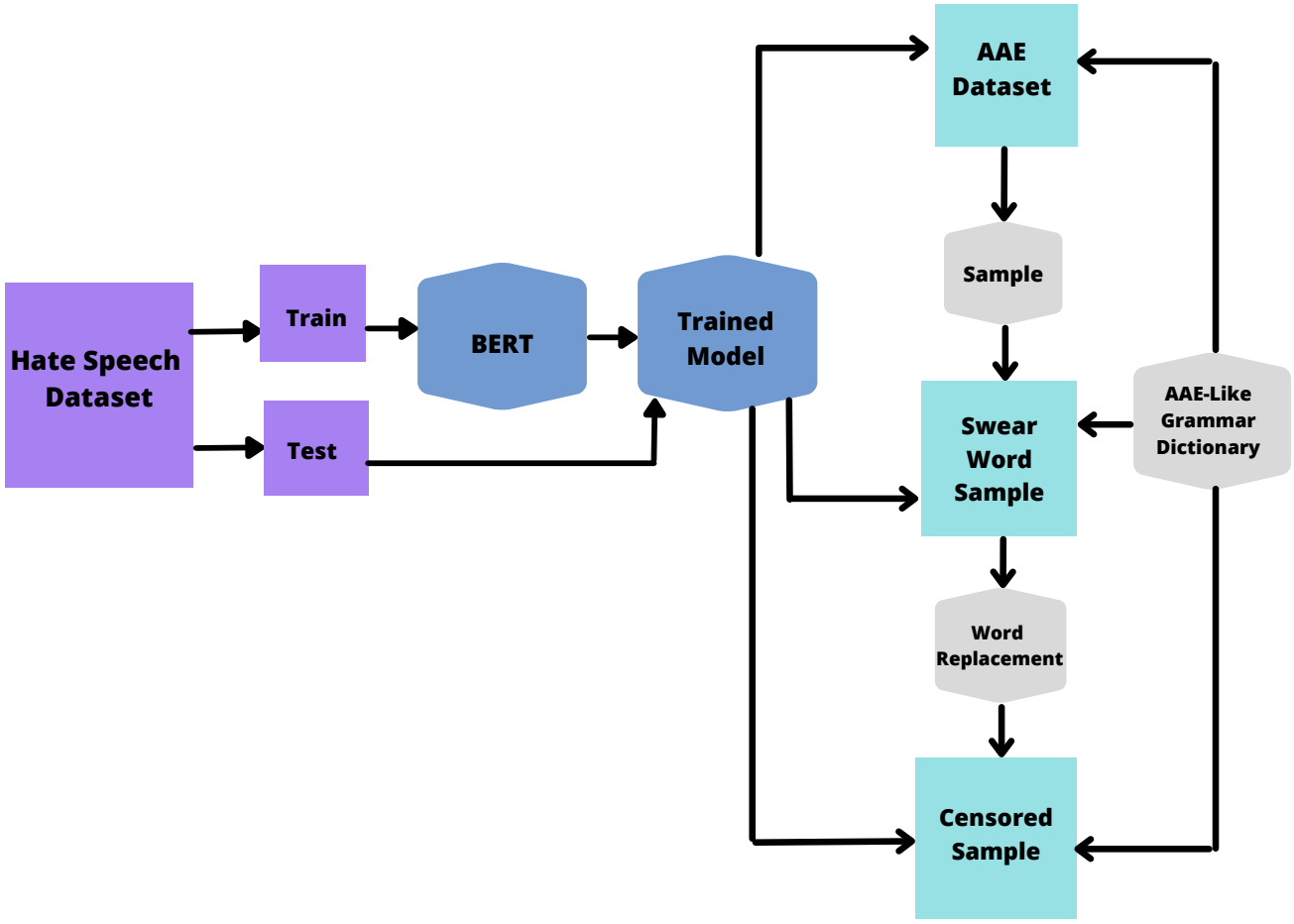
The four categories we focus on are *auxiliary verbs, aspectual markers, preverbal markers*, and *syntactic and morphorosyntactic properties*:

- Auxiliary verbs are words that form the tenses and voices of other verbs, such as the words "be" "have" and "do" and their various forms. AAE has unique uses of standard auxiliary verbs and some that are unique to the dialect [25]. An example of an auxiliary verb that is unique to AAE is the word "finna" which originated as a contraction of the words "fixing to," and indicates that something will happen in the immediate future [24]
- Aspectual markers are grammatical and lexical means of expressing aspect in a sentence. The word "be" used in a habitual sense, such as in the sentence "he be working out" is an aspectual marker indicating something that happens regularly in AAE [7]. The habitual form of the word "be" is common in AAE. Preverbal markers are words that mark a verb with tense, aspect, or modality, of the proceeding verb.
- Preverbal markers commonly vary from Standard English within African influenced dialects of English such as Jamaican Creole and AAE [15]. A commonly known preverbal marker in AAE is the word "ain't," a contraction of "am not" that has become common within some communities outside of the African American English dialect.
- Syntactic and Morphosyntatic properties are properties of AAE that influence word order and word co-occurrence in

Dataset	Total Number of Tweets	Number of Hate Speech Samples	Number of Offensive /Abusive Samples
DWMW17	24783	1430	19190
FDCL18	50487	1887	4563
Blodgett et al.	50,000	n/a	n/a

Table 1: Dataset Details

Figure 1: Methodology Pipeline Diagram: This is the pipeline followed for both the DWMW17 and the FDCL18 hate speech datasets. First we split the dataset appropriately and train a BERT classifier on the data. Then we test the model and use the best performing model. Next we use the classifier to classify tweets from the AAE dataset. Then we sample our dataset for tweets containing swear words and use word replacement to create a censored version of the sample. We use the model to classify both the censored and uncensored tweets. We also apply the AAE-like grammar dictionary to all sets of data to see how classifications vary across grammar characteristics of AAE.



sentences [38]. An example of a syntactic property in AAE is the use of multiple negatives within a negative sentence.

Table 3 details multiple aspects of these four categories of grammar properties with examples. We use these properties to inform the words and phrases that make up our AAE-Like Grammar Dictionary. For each tweet in our sub-sample, we count how many dictionary words/phrases for each category are in the tweet.

Table 2: AAE-Like Grammar Dictionary

Auxiliary	Aspectual	Preverbal	Syntactic
we was	I be	aint	cant nobody
they was	he be	ain't	can't nobody
finna	they be	steady	he don't
tryna	she be	stay	she don't
imma	I been	he done	don't never
i'mma	he been	she done	he dont
bitches was	she been	they done	she dont
niggas was	they been	yall done	dont never
yall was	it be	y'all done	yall don't
y'all was	niggas be	you done	y'all don't
you was	bitches be	u done	aint nothing
wanna	yall be		ain't nothing
gonna	y'all be		aint nobody
ima	you be		ain't nobody
ion	u be		yall dont
u was			she done
iont			he done
			they done
			yall done

Table 3: African American English Grammar Properties with AAE and SAE examples

	African American English Example	Standard English Example Equivalent
Auxiliary Verbs		
Single verb used for singular and plural subjects	We was eating	We were eating
verbs contracted, reduced, or removed	They walking too fast	They are walking too fast
verbs used for if-clause	We asked did he want to go with us	We asked if he wanted to go with us
Semi- verbs- (-ing verbs contracted)	I'm tryna sleep	I'm trying to sleep
Aspectual Markers		
Habitual "Be"	I be at the office by 7:30.	Usually I am at the office by 7:30
Remote Past "Been"	I been knowing that.	I have known that for a long time
Preverbal Markers		
"Ain't" (am not contracted)	I ain't interested	I am not interested
"Stay" or "steady" to mark consistency	They stay doing they own thing	They always do their own thing
Syntactic and Morphosyntactic Properties		
Multiple negators in a single negative sentece	He dont want no teacher yelling at him	He doesnt want a teacher yelling at him
Existential "It" to indicate existance	It's some coffee in the kitchen	There's some coffee in the kitchen
Unmarked posessive	Somebody car was parked out there	Somebody's car was parked out there

We use this dictionary by using string comparison to find any tweets in our sample that contain at least one word or phrase from each category of the grammar dictionary. We consider a tweet to contain elements of a grammar category if the tweet has at least 1 match to a element for that grammar category in our AAE-Like Grammar Dictionary. We do this on the full dataset, as well as the sample of swear word tweets and the censored version of the sample. We map out our overall pipeline of our methodology used for each hate speech dataset in Figure 1.

6 RESULTS

As a baseline, we present the classification results for each BERT classifier on the overall data which we present in Table 4 . We find that for the AAE dataset, 28.462% of samples were classified as offensive by the DWMW17 model, and 0.888% were classified as hate speech. Furthermore for the FDCL18 model, 19.886% of tweets were classified as abusive and 5.422% as hate speech. These results suggest consistency with prior work [35]. In the following sections we further discuss our results and answer our research questions.

Table 4: Percentage of Tweets Classified as Each Hate Speech Label for Overall Dataset

	DWMW17-Hate	FDCL18-Hate	DWMW17-Offensive	FDCL18-Abusive
Full Data	0.888	5.422	28.462	19.886

6.1 Impact of Swear Words or Offensive Language on AAE Hate Speech Classification

To answer our RQ1, we use these models to classify the subsample of swear word tweets and the censored version of the sample. On the subsample, the original tweets were classified by the DWMW17 96.18% were offensive and 2.46% as hate. For FDCL18, 67.77% of tweets were classified as abusive and 27.89% were classified as hate. We find that by removing swear words (and substituting it with a semantically similar non-swear words) from the subsample of AAE tweets, the DWMW17 model classifies 64.07% as offensive and 1.93% of tweets as hate. The FDCL18 model gives 51.04% classified as abusive speech and 5.496% hate speech. We summarize these results in Table 5. In a small sample of 50 tweets, through manual inspection we observed a 78% success rate in removing all swear words and having identical meaning to the original tweet. These results are shown in Table 5. However we see a much less significant reduction in the hate, offensive, and abusive classifications in the edited text.

In some rare instances rewording tweets did not result in a less harsh classification. For the DWMW17 based model, 57 tweets changed from 'normal' to 'offensive' with rewording, 1 from 'normal' to 'hate' speech, and 108 from 'offensive' to 'hate.' For the FDCL18 based model 133 from 'normal' to 'abusive', 12 from 'normal' to 'hate', and 217 samples from 'abusive' to 'hate'. For a very small number of samples we notice this phenomenon with both classifiers at the same time. There were 2 samples where both FDCL18 went from 'abusive' to 'hate' and DWMW17 went from 'normal' to 'offensive' and 5 samples where FDCL18 went from abusive to hate and DWMW17 went from 'offensive' to 'hate'. However for the majority of tweets, we see this rewording create a less harsh classification, such as 'hate' to 'abusive' or 'offensive.' Table 7 shows examples of original and reworded tweets and their classifications by both models.

The overall findings for answer RQ1 are shown in Table 5. Our findings suggest that censoring tweets for swear words has a significant impact on reducing hate speech, offensive speech, and abusive speech classifications. In Table 5 we see that within the sample, removing swear words from the original tweet we significantly reduce the classifications for hate for the FDCL18 classifier from 27.89% to 5.40%, and that we see a moderate reduction in the DWMW17 offensive and FDCL18 abusive categories, from 96.18 to 64.07% and 67.77 to 51.04% respectively. We see only a slight reduction in the DWMW17 hate classification, from 2.46 to 1.93%. Censoring the tweets gives results more similar to the overall dataset but does not yield a lower percentage for any classification category.

Table 5: Percentage of Tweets Classified as Each Hate Speech Label for Swear Word Sample, Edited Sample, and the Difference

	DWMW17-Hate	FDCL18-Hate	DWMW17-Offensive	FDCL18-Abusive
Original Sample	2.46	27.89	96.18	67.77
Edited Sample	1.93	5.496	64.07	51.04
Difference	-0.53	-22.394	-32.11	-16.73

6.2 Grammatical Patterns of AAE and Hate Speech Classification

To answer RQ2, we use our AAE-Like Grammar Dictionary to categorize tweets from the full dataset, sample, and censored sample into the four grammar categories.

For the full dataset of 50,000 tweet samples, we identified 4169 tweets containing Auxiliary Verbs, 1211 with Aspectual Markers 1893 tweets containing Preverbal markers and 138 with Syntactic and Morphosyntactic properties. Across the swear word sample and the censored sample the grammar categories were not impacted, so the grammar categories are the same across the original and reworded sample. For both versions of the sample, we identified 640 samples with Auxiliary Verbs, 239 samples with Aspectual Markers, 403 samples with Preverbal Markers and 33 samples with Syntactic and Morphosyntactic Properties. We include our results for Syntactic and Morphosyntactic Properties in this instance, but note that with so few samples further work is needed for a meaningful analysis of this category. We provide these results compared to the overall data in Table 6 for the percentages of tweets classified as hate, abusive, and offensive speech.

We find similar patterns for each of the classifiers across all four of the grammar categories we explore. Tweets within the four grammar categories from each source show results consistent with the source for each the full data, the sample, and the censored version of the sample. Comparing the results for the original sample versus censored sample across all grammar categories, we see the most significant reductions in the DWMW17 classification of 'offensive' with Syntactic and Morphosyntactic Properties, and the most significant reduction in the FDCL18 classification of 'abusive' for Auxiliary Verbs. In all cases, the number of tweets classified by DWMW17 as 'hate' is below 20, with such low number of samples we are unable to make meaningful conclusions about the relationship between these grammar categories and this classification. For the FDCL18 classification of 'hate' we see similar results for all categories except Syntactic and Morphosyntactic Properties which sees a smaller reduction in 'hate' classifications, again this category is the smallest so we consider these results interesting but inconclusive. We can conclude from these results that hate speech classification is not dependent on the four grammar categories we explored individually. Instead it is possible that a more complex combination of these or other grammar categories that we did not explore may have a strong association with hate speech classification, or that grammar patterns do not significantly impact classification.

7 DISCUSSION

The bias issues of AAE and hate speech classification support anti-Black racism. The existence of these biases should be unsurprising for technologies that emerge in a culture in which anti-Black racism is heavily embedded. Under anti-Black racism, Blackness as a whole is attributed to many negative attributes and as such, AAE is as well. In this study, we are unable to identify a singular aspect of the AAE dialect that cause these issues of misclassification, but note that reducing swear words significantly impacts the classification. We observe that the four grammar categories we explored have minimal impact on the classification. There are likely no specific attributes of AAE that when controlled can completely mitigate these biases due to how sentiment against Black English as a whole is embedded in our culture and appear in our algorithms via training data that embodies racism.

Preventing unnecessarily harsh classifications of AAE text is only one aspect of the importance of reducing these biases. Another is that true examples of AAE hate speech and toxic speech could be obscured by false positives. In-group harassment towards specific subsets of the Black community such as Black women, Black disabled people, Black LGBT people, etc. is a hate speech issue as well [14]. These groups, or individuals apart of them, can face harassment from within and outside of the Black community. The biases towards AAE in hate speech classification systems can obscure the true instances of AAE hate speech that likely impact individuals or groups within the Black community with additional marginalized identities.

Prior works suggest training annotators on understanding AAE [35], or debiasing training data [44] as strategies to mitigate bias. As we discuss in our limitations, a dataset of AAE text with ground-truth labels for dialect and hate speech classification would be a valuable contribution to expanding bias mitigation of AAE in hate speech classification.

8 LIMITATIONS AND FUTURE WORK

One major limitation of this work is that while the Blodgett et al. (2016) dataset is the best available source of tweets labeled as AAE, it does not provide ground-truth labels for racial background, nor does it provide ground-truth labels for hate speech. The missing ground truth of racial background is heavily complicated by the issue of cultural appropriation. Cultural appropriation of AAE in an online context has become a growing issue of concern in the Black community, particularly on social media. It is possible that non-Black individuals appropriating this dialect may be present in this dataset.This appropriative use of AAE is often an incorrect representation of the dialect, both culturally and grammatically. This phenomenon may contribute to misrepresentations of AAE

Table 6: Percentage of Tweets Classified as each Hate Speech Label by Grammar Category

	DWMW17-Hate	FDCL18-Hate	DWMW17 Offensive	FDCL18 Abusive
All Data				
Original Text (full data)	0.888	5.422	28.462	19.886
Original Text (sample)	2.46	27.89	96.18	67.77
Edited Text (sample, censored)	1.93	5.496	64.07	51.04
Auxiliary Verbs Only				
Original Text (full data)	0.96	5.56	28.30	20.74
Original Text (sample)	2.50	27.19	96.25	69.21
Edited Text (sample, censored)	2.03	5.47	66.72	49.69
Aspectual Markers Only				
Original Text (full data)	0.66	5.28	27.91	21.47
Original Text (sample)	1.67	29.29	97.07	67.78
Edited Text (sample, censored)	2.51	6.28	63.18	50.21
Preverbal Markers Only				
Original Text (full data)	0.63	5.49	28.68	21.71
Original Text (sample)	3.47	27.30	96.28	68.24
Edited Text (sample, censored)	2.98	5.46	65.76	46.40
Syntactic and Morphosyntactic Properties Only				
Original Text (full data)	2.17	2.17	25.36	20.29
Original Text (sample)	3.03	33.34	96.97	60.61
Edited Text (sample, censored)	6.06	15.15	54.55	42.42

Table 7: Swear Word and Censored Tweet Samples with Classification

Text	DWMW17	FDCL18
Ya friend telling all ya business to ya nigga and #UONEENKNOWIT	hate	abusive
Ya friend telling all ya business to ya man and #UONEENKNOWIT	offensive	normal
@[name removed] check!!@ already know!! These bitches don't mean us NO good	offensive	hate
@[name removed] check!!@ already know!! These people don't mean us NO good	normal	normal
Oomf is straight out here dawg lol I just met a complete stranger that knew about her ass	offensive	normal
Oomf is straight out here dawg lol I just met a complete stranger that knew about her butt	normal	normal
Brittany dry ass	offensive	abusive
Brittany dry butt	offensive	abusive
@[name removed] ain't none wrong with yo ass you just wanna go off on me	hate	abusive
@[name removed] ain't none wrong with yo butt you just wanna go off on me	normal	abusive
Too many bitches got relationship or nigga problems! #GetYallShitTogetha	offensive	abusive
Too many people got relationship or man problems! #GetYallShitTogetha	hate	normal
@[name removed] Nah you my nigga bruh	hate	abusive
@[name removed] Nah you my man bruh	offensive	abusive
Why don't kane beat the hell out of cm punk he need it @[name removed]	offensive	hate
Why don't kane beat the h*** out of cm punk he need it @[name removed]	normal	normal

in this dataset and in turn in our analysis. The ground truth of hate speech labels is less concerning; considering that this dataset is sampled from Twitter which is a moderated platform with few true instances of hate speech, and that our data collection used tweet id's to collect only the tweets that were still present on the platform at the time of analysis, the vast majority of posts are very unlikely true hate speech. A current gap in the study of AAE and hate speech classification is a dataset of text sampled from Black individuals who use AAE and ground-truth hate speech labels. A valuable future work would be the creation of such a dataset.

Similarly, another limitation of this work was in utilizing the LIWC2007 dictionary and supplementing it with our own words. The LIWC2007 dictionary was created with English text samples sourced from a variety of studies from the United States, England, Canada, New Zealand, and Australia [28]. However, this system

does not provide a good representation of African American English. As discussed in Section 4, there are many swear words that are common in AAE that are less common in Standard American English and similarly less common in the other Standard English variations used to build this tool. Likewise, the swear words in the LIWC2007 "Swear" category included many words that are not commonly found in the AAE dialect, and missing many words that are. To successfully sample enough tweets with swear words, we needed to heavily supplement the list of words in the LIWC2007 "Swear" category, primarily using words that were found in the AAE data. While we were able to identify many more tweets with this supplementation, the list we used is not all encompassing of swear words used in AAE. Creation of a language tool similar to LIWC2007 using samples of AAE text would be valuable to future work exploring biases in natural language systems that affect Black people, and in creating inclusive natural language systems. Furthermore, to expand the reach of such a tool, dialects of African-descended people in each of the countries used in the original LIWC system would add relevance for other countries outside of the United States.

Another limitation of this study is the black-box nature of Twitter's hate speech identification process. Despite the success of many different architectures and techniques of hate speech classification proposed in the NLP literature, in practice, Twitter and other social media websites do not necessarily use these systems. Twitter and other platforms do not disclose their exact approach to identifying hate speech, or their evaluation process. It is known that Twitter uses content moderators and allows users to report content that violates policy where it will be reviewed. Whether or not tweets are reviewed by an automated process, human moderators, or a mix of both is unknown. Replicating this study using a BERT model trained on Twitter's real classifications of tweets rather than on the DWMW17 and FDCL18 would be most relevant to understanding the real world implications of this study. But this is impossible without internal access to this information from Twitter.

Finally, another area for future work is an expansion on the grammar analysis we performed. Future work should include grammar analysis that directly compares the grammatical variations we studied with the equivalent Standard American English grammar. This would give a stronger understanding of how these grammar patterns impact the classifications.

9 CONCLUSION

Our study explores the impact of more granular aspects of AAE text such as word choice and grammar patterns on the hate speech text classification. We explore word choice by taking a sample of the AAE data that uses swear words and creating an identical sample with swear words replaced with a non-offensive word with similar meaning. We find that reducing the swear words does significantly reduce classifications of "hate" "abusive" and "offensive" speech. The censored version of the sample received similar classifications for these three categories and the overall dataset.

We explore grammar categories by creating the AAE-Like Grammar Dictionary which we hope will be a valuable tool for future researchers interested in exploring AAE. We use this dictionary to the full data, the swear word sample, and the censored sample to extract the tweets with evidence of these grammar categories from each. We consistently see similar patterns in classification within each grammar category to the overall data. This suggests that the four grammar categories we explored have little relationship to the hate speech classification label. However due to low samples especially within the Syntactic and Morphosyntactic Properties grammar category, more research could be done to explore this further. Overall this research contributes to our understanding of why AAE is sometimes unfairly censored online, and suggests avenues for mitigating this bias in the future.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for the helpful feedback. This work was supported in part by a grant from Cisco Systems, Inc.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*. 759–760.

[2] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[3] Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1415–1425. https://doi.org/10.18653/v1/P18-1131

[4] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science* 5 (2016), 1–15.

[5] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 71–80.

[6] Chris Collins, Simanique Moody, and Paul M. Postal. 2008. An AAE Camouflage Construction. *Language* 84, 1 (2008), 29–68. http://www.jstor.org/stable/40071011

[7] Patricia Cukor-Avila and Guy Bailey. 1995. Grammaticalization in AAVE. In *Annual Meeting of the Berkeley Linguistics Society*, Vol. 21. 401–413.

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

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

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.

[12] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*. 105–114.

[13] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[14] Kishonna L Gray. 2018. Gaming out online: Black lesbian identity development and community building in Xbox Live. *Journal of lesbian studies* 22, 3 (2018), 282–296.

[15] Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

[16] Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. *arXiv preprint arXiv:2010.02510* (2020).

[17] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[18] Jacob Hoffman-Andrews. 2015. *BlockTogether*. https://blocktogether.org/

[19] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[20] Taylor Jones. 2015. Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter". *American Speech* 90, 4 (2015), 403–440.

[21] Tiffany Marquise' Jones. 2008. " You Done Lost Yo'Mind Ain't No Such Thang as AAVE": Exploring African American Resistance to AAVE. (2008).

[22] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[23] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. *arXiv preprint arXiv:2106.04511* (2021).

[24] Austin Lane. 2014. "You Tryna Grammaticalize?" An Analysis of "Tryna" as a Grammaticalized Semi-Auxiliary. *Eagle Feather* 11, 2014 (2014).

[25] Stefan Martin, Walt Wolfram, et al. 1998. The sentence in African-American vernacular English. *African American English: structure, history, and use* (1998), 11–36.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[27] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[28] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[29] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[30] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence* 48, 12 (2018), 4730–4742.

[31] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1534–1545. https://www.aclweb.org/anthology/C18-1130

[32] Thomas Purnell, William Idsardi, and John Baugh. 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of language and social psychology* 18, 1 (1999), 10–30.

[33] Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. *arXiv preprint arXiv:1804.03124* (2018).

[34] Jacquelyn Rahman. 2012. The N word: Its history and use in the African American community. *Journal of English Linguistics* 40, 2 (2012), 137–171.

[35] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[36] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[37] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. KDEHatEval at SemEval-2019 Task 5: A Neural Network Model for Detecting Hate Speech in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 365–370. https://doi.org/10.18653/v1/S19-2064

[38] Deanna Thompson. 2016. The Morpho-syntax of Aspectual Stay in AAVE. (2016).

[39] Christopher Tuckwood. 2017. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org* (2017).

[40] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *International conference recent advances in natural language processing (RANLP)*. 672–680.

[41] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[42] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[43] Robert L Williams. 1997. The ebonics controversy. *Journal of Black Psychology* 23, 3 (1997), 208–214.

[44] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. *arXiv preprint arXiv:2102.00086* (2021).