The Risk of Racial Bias in Hate Speech Detection

Maarten Sap[◇] **Dallas Card**[♣] **Saadia Gabriel**[◇] **Yejin Choi**[◇♡] **Noah A. Smith**[◇♡]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[♣]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
[♡]Allen Institute for Artificial Intelligence, Seattle, USA
msap@cs.washington.edu

Abstract

We investigate how annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. We first uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases, such that AAE tweets and tweets by self-identified African Americans are up to two times more likely to be labelled as offensive compared to others. Finally, we propose dialect and race priming as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

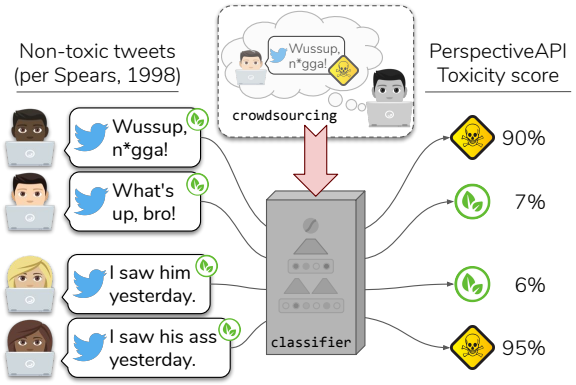


Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

1 Introduction

Toxic language (e.g., hate speech, abusive speech, or other offensive speech) primarily targets members of minority groups and can catalyze real-life violence towards them (O'Keeffe et al., 2011; Cleland, 2014; Mozur, 2018). Social media platforms are under increasing pressure to respond (Trindade, 2018), but automated removal of such content risks further suppressing already-marginalized voices (Yasin, 2018; Dixon et al., 2018). Thus, great care is needed when developing automatic toxic language identification tools.

The task is especially challenging because what is considered toxic inherently depends on social context (e.g., speaker's identity or dialect). Indeed, terms previously used to disparage communities (e.g., "n*gga", "queer") have been reclaimed by those communities while remaining offensive when used by outsiders (Rahman, 2012). Figure 1 illustrates how phrases in the African American English dialect (AAE) are labelled by a publicly available toxicity detection tool as much more toxic than general American English equivalents, despite their being understood as non-toxic by AAE speakers (Spears, 1998, see §2).

In this work, we first empirically characterize the racial bias present in several widely used Twitter corpora annotated for toxic content, and quantify the propagation of this bias through models trained on them (§3). We establish strong associations between AAE markers (e.g., "n*ggas", "ass") and toxicity annotations, and show that models acquire and replicate this bias: in other corpora, tweets inferred to be in AAE and tweets from self-identifying African American users are more likely to be classified as offensive.

Second, through an annotation study, we introduce a way of mitigating annotator bias through *dialect* and *race priming*. Specifically, by designing tasks that explicitly highlight the inferred dialect of a tweet or likely racial background of its author, we show that annotators are significantly less likely to label an AAE tweet as offensive than when not shown this information (§4).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678
Florence, Italy, July 28 - August 2, 2019.

Our findings show that existing approaches to toxic language detection have racial biases, and that text alone does not determine offensiveness. Therefore, we encourage paying greater attention to the confounding effects of dialect and a speaker's social identity (e.g., race) so as to avoid unintended negative impacts.

2 Race and Dialect on Social Media

Since previous research has exposed the potential for other identity-based biases in offensive language detection (e.g., gender bias; Park et al., 2018), here we investigate *racial* bias against speech by African Americans, focusing on Twitter as it is a particularly important space for Black activism (Williams and Domoszlai, 2013; Freelon et al., 2016; Anderson et al., 2018). Race is a complex, multi-faceted social construct (Sen and Wasow, 2016) that has correlations with geography, status, dialect, and more. As Twitter accounts typically do not have self-reported race information, researchers rely on various correlates of race as proxies. We use the African American English *dialect* (AAE) as a proxy for race. AAE is a widely used dialect of English that is common among, but not unique to, those who identify as African American,[1] and is often used in written form on social media to signal a cultural identity (Green, 2002; Edwards, 2004; Florini, 2014).

Dialect estimation In this work, we infer dialect using a lexical detector of words associated with AAE or white-aligned English. We use the topic model from Blodgett et al. (2016), which was trained on 60M geolocated tweets and relies on US census race/ethnicity data as topics. The model yields probabilities of a tweet being AAE (p_{AAE}) or White-aligned English (p_{white}).[2]

3 Biases in Toxic Language Datasets

To understand the racial and dialectic bias in toxic language detection, we focus our analyses on two corpora of tweets (Davidson et al., 2017; Founta et al., 2018) that are widely used in hate speech detection (Park et al., 2018; van Aken et al., 2018; Kapoor et al., 2018; Alorainy et al., 2018; Lee et al., 2018; Waseem et al., 2018).[3] Different protocols were used to collect the tweets in these corpora, but both were annotated by Figure-Eight[4] crowdworkers for various types of toxic language, shown in Table 1.

	category	count	AAE corr.
DWMW17	hate speech	1,430	−0.057
	offensive	19,190	0.420
	none	4,163	−0.414
	total	**24,783**	
FDCL18	hateful	4,965	0.141
	abusive	27,150	0.355
	spam	14,030	−0.102
	none	53,851	−0.307
	total	**99,996**	

Table 1: Number of tweets in each category, and correlation with AAE (Pearson r, $p \ll 0.001$). We assign tweets to categories based on the label for FDCL18, and majority class for DWMW17. Correlations are colored for interpretability.

DWMW17 (Davidson et al., 2017) includes annotations of 25K tweets as *hate speech*, *offensive* (but not hate speech), or *none*. The authors collected data from Twitter, starting with 1,000 terms from HateBase (an online database of hate speech terms) as seeds, and crowdsourced at least three annotations per tweet.

FDCL18 (Founta et al., 2018) collects 100K tweets annotated with four labels: *hateful*, *abusive*, *spam* or *none*. Authors used a bootstrapping approach to sampling tweets, which were then labelled by five crowdsource workers.

3.1 Data Bias

To quantify the racial bias that can arise during the annotation process, we investigate the correlation between toxicity annotations and dialect probabilities given by Blodgett et al. (2016).

Table 1 shows the Pearson r correlation between p_{AAE} and each toxicity category. For both datasets, we uncover strong associations between

[1] Of course, many African Americans might not use AAE in every context, or at all. For further discussion of AAE, please refer to Blodgett et al. (2016).

[2] The model yields AAE, Hispanic, Asian/Other and White-aligned dialect probabilities, but for the purpose of our study we only focus on AAE and White-aligned dialects.

[3] Our findings also hold for the widely used data from Waseem and Hovy (2016). However, because of severe limitations of that dataset (see Schmidt and Wiegand, 2017; Klubika and Fernandez, 2018), we relegate those analyses to supplementary (§A.3).

[4] www.figure-eight.com

Within dataset proportions

DWMW17

Group	Acc.	% false identification: None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

FDCL18

Group	Acc.	% false identification: None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

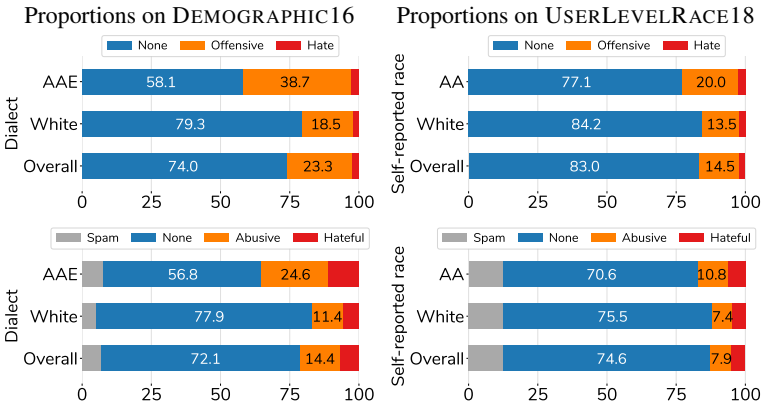


Figure 2: *Left*: classification accuracy and per-class rates of false positives (FP) on test data for models trained on DWMW17 and FDCL18, where the group with highest rate of FP is bolded. *Middle and right*: average probability mass of toxicity classes in DEMOGRAPHIC16 and USERLEVELRACE18, respectively, as given by classifiers trained on DWMW17 (top) and FDCL18 (bottom). Proportions are shown for AAE, White-aligned English, and overall (all tweets) for DEMOGRAPHIC16, and for self-identified White authors, African American authors (AA), and overall for USERLEVELRACE18.

inferred AAE dialect and various hate speech categories, specifically the "offensive" label from DWMW17 ($r = 0.42$) and the "abusive" label from FDCL18 ($r = 0.35$), providing evidence that dialect-based bias is present in these corpora. As additional analyses, we examine the interaction between unigrams indicative of dialect and hate speech categories, shown in §A.1.

3.2 Bias Propagation through Models

To further quantify the impact of racial biases in hate speech detection, we investigate how these biases are acquired by predictive models. First, we report differences in rates of false positives (FP) between AAE and White-aligned dialect groups for models trained on DWMW17 or FDCL18. Then, we apply these models to two reference Twitter corpora, described below, and compute average rates of reported toxicity, showing how these biases generalize to other data.[5]

DEMOGRAPHIC16 (Blodgett et al., 2016) contains 56M tweets (2.8M users) with dialect estimated using a *demographic-aware topic model* that leverages census race/ethnicity data and geocoordinates of the user profile. As recommended, we assign dialect labels to tweets with dialect probabilities greater than 80%.

USERLEVELRACE18 (Preoţiuc-Pietro and Ungar, 2018) is a corpus of 5.4M tweets, collected from 4,132 survey participants (3,184 White, 374 AA) who reported their race/ethnicity and Twitter user handle. For this dataset, we compare differences in toxicity predictions by *self-reported race*, instead of inferring message-level dialect.[6]

For each of the two toxic language corpora, we train a classifier to predict the toxicity label of a tweet. Using a basic neural attention architecture (Wang et al., 2016; Yang et al., 2016), we train a classifier initialized with GloVe vectors (Pennington et al., 2014) to minimize the cross-entropy of the annotated class conditional on text, $\boldsymbol{x}$:

$$p(\text{class} \mid \boldsymbol{x}) \propto \exp(\boldsymbol{W}_o \boldsymbol{h} + \boldsymbol{b}_o), \qquad (1)$$

with $\boldsymbol{h} = f(\boldsymbol{x})$, where f is a BiLSTM with attention, followed by a projection layer to encode the tweets into an H-dimensional vector.[7] We refer the reader to the appendix for experimental details and hyperparameters (§A.2).

Results Figure 2 (left) shows that while both models achieve high accuracy, the false positive rates (FPR) differ across groups for several toxicity labels. The DWMW17 classifier predicts almost 50% of non-offensive AAE tweets as being offensive, and FDCL18 classifier shows higher FPR for

[5] We assume *a priori* that the average tweet is not inherently more toxic in a particular dialect. Assessing the veracity of this assumption requires a deep understanding of sociocultural norms of profane and toxic speech.

[6] Note that lexical dialect inferences of AAE (p_{AAE}) significantly correlate with both the AAE group from DEMOGRAPHIC16 (Pearson $r = 0.61$, $p \ll 0.001$) and self-reported AA race from USERLEVELRACE18 (Pearson $r = 0.21$, $p \ll 0.001$).

[7] In preliminary experiments, our findings held regardless of our choice of classifier.

the "Abusive" and "Hateful" categories for AAE tweets. Additionally, both classifiers show strong tendencies to label White tweets as "none". These discrepancies in FPR across groups violate the *equality of opportunity* criterion, indicating discriminatory impact (Hardt et al., 2016).

We further quantify this potential discrimination in our two reference Twitter corpora. Figure 2 (middle and right) shows that the proportions of tweets classified as toxic also differ by group in these corpora. Specifically, in DEMOGRAPHIC16, AAE tweets are more than twice as likely to be labelled as "offensive" or "abusive" (by classifiers trained on DWMW17 and FDCL18, respectively). We show similar effects on USERLEVELRACE18, where tweets by African American authors are 1.5 times more likely to be labelled "offensive". Our findings corroborate the existence of racial bias in the toxic language datasets and confirm that models propagate this bias when trained on them.[8]

4 Effect of Dialect

To study the effect of dialect information on ratings of offensiveness, we run a small controlled experiment on Amazon Mechanical Turk where we prime annotators to consider the dialect and race of Twitter users. We ask workers to determine whether a tweet (a) is offensive *to them*, and (b) could be seen as offensive *to anyone*. In the *dialect priming* condition, we explicitly include the tweet's dialect as measured by Blodgett et al. (2016), as well as extra instructions priming workers to think of tweet dialect as a proxy for the author's race. In the *race priming* condition, we encourage workers to consider the likely racial background of a tweet's author, based on its inferred dialect (e.g., an AAE tweet is likely authored by an African American Twitter user; see §A.5 for the task instructions). For all tasks, we ask annotators to optionally report gender, age, race, and political leaning.[9]

With a distinct set of workers for each condition, we gather five annotations apiece for a sample of 1,351 tweets stratified by dialect, toxicity category, and dataset (DWMW17 and FDCL18).[10]

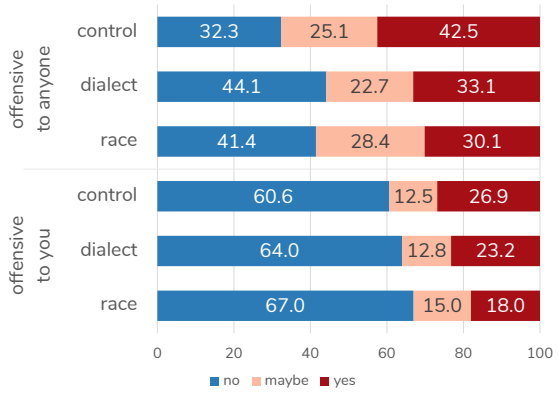


Figure 3: Proportion (in %) of offensiveness annotations of AAE tweets in control, dialect, and race priming conditions. Results show that dialect and race priming significantly reduces an AAE tweet's likelihood of being labelled offensive ($p \ll 0.001$).

Despite the inherent subjectivity of these questions, workers frequently agreed about a tweet being offensive to anyone (76% pairwise agreement, $\kappa = 0.48$) or to themselves (74% p.a., $\kappa = 0.30$).

Results Figure 3 shows that priming workers to think about dialect and race makes them significantly less likely to label an AAE tweet as (potentially) offensive to anyone. Additionally, race priming makes workers less likely to find AAE tweets offensive to them.

To confirm these effects, we compare the means of the control condition and treatment conditions,[11] and test significance with a t test. When rating offensiveness to anyone, the mean for control condition ($M_c = 0.55$) differs from dialect ($M_d = 0.44$) and race ($M_r = 0.44$) conditions significantly ($p \ll 0.001$). For ratings of offensiveness to workers, only the difference in means for control ($M_c = 0.33$) and race ($M_d = 0.25$) conditions is significant ($p \ll 0.001$).

Additionally, we find that overall, annotators are substantially more likely to rate a tweet as being offensive to *someone*, than to rate it as offensive to *themselves*, suggesting that people recognize the subjectivity of offensive language.

Our experiment provide insight into racial bias in annotations and shows the potential for reducing it, but several limitations apply, including the skewed demographics of our worker pool (75% self-reported White). Additionally, research suggests that motivations to not seem prejudiced

[8] As noted by Chung (2019), the PerspectiveAPI displays similar racial biases shown in the appendix (§A.4).

[9] This study was approved by the Institutional Review Board (IRB) at the University of Washington.

[10] Annotations in the control setting agreed moderately with toxicity labels in DWMW17 and FDCL18 (Pearson r = 0.592 and r = 0.331, respectively; $p \ll 0.001$).

[11] We convert the offensiveness labels to real numbers (0: "no", 0.5: "maybe", 1: "yes").

could buffer stereotype use, which could in turn influence annotator responses (Plant and Devine, 1998; Moskowitz and Li, 2011).

5 Related Work

A robust body of work has emerged trying to address the problem of hate speech and abusive language on social media (Schmidt and Wiegand, 2017). Many datasets have been created, but most are either small-scale pilots (∼100 instances; Kwok and Wang, 2013; Burnap and Williams, 2015; Zhang et al., 2018), or focus on other domains (e.g., Wikipedia edits; Wulczyn et al., 2017). In addition to DWMW17 and FDCL18, published Twitter corpora include Golbeck et al. (2017), which uses a somewhat restrictive definition of abuse, and Ribeiro et al. (2018), which is focused on network features, rather than text.

Past work on bias in hate speech datasets has exclusively focused on finding and removing bias against explicit identity mentions (e.g., woman, atheist, queer; Park and Fung, 2017; Dixon et al., 2018). In contrast, our work shows how insensitivity to dialect can lead to discrimination against minorities, even without explicit identity mentions.

6 Conclusion

We analyze racial bias in widely-used corpora of annotated toxic language, establishing correlations between annotations of offensiveness and the African American English (AAE) dialect. We show that models trained on these corpora propagate these biases, as AAE tweets are twice as likely to be labelled offensive compared to others. Finally, we introduce *dialect* and *race priming*, two ways to reduce annotator bias by highlighting the dialect of a tweet in the data annotation, and show that it significantly decreases the likelihood of AAE tweets being labelled as offensive. We find strong evidence that extra attention should be paid to the confounding effects of dialect so as to avoid unintended racial biases in hate speech detection.

Acknowledgments

The authors thank Dan Jurafsky, Emily Bender, Emily Gade, Tal August, Wesley McClean, Victor Zhong, and Laura Vianna, as well as anonymous reviewers, for helpful feedback. This work was in part supported by NSF grant IIS-1714566.

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew Williams. 2018. Cyber hate classification: 'othering' language and paragraph embedding. *CoRR*, abs/1801.07495.

Monica Anderson, Skye Toor, Lee Rainie, and Aaron Smith. 2018. Activism in the social media ages. http://www.pewinternet.org/2018/07/11/activism-in-the-social-media-age/. Accessed: 2019-03-01.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American english. In *EMNLP*.

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7:223–242.

Anna Chung. 2019. How automated tools discriminate against black language. https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db. Accessed: 2019-03-02.

Jamie Cleland. 2014. Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football. *J. Sport Soc. Issues*, 38(5):415–431.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of Conference on AI, Ethics, and Society*.

Walter F. Edwards. 2004. African American Vernacular English: phonology. In *A Handbook of Varieties of English: Morphology and Syntax*.

Sarah Florini. 2014. Tweets, tweeps, and signifyin': Communication and cultural performance on "Black Twitter". *Television & New Media*, 15(3):223–237.

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 *ICWSM*.

Deen Freelon, Charlton D. McIlwain, and Meredith D. Clark. 2016. Beyond the hashtags. http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf. Accessed: 2019-03-01.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *WebSci*, pages 229–233. ACM.

Lisa Green. 2002. *African American English: A Linguistic Introduction*, 8.3.2002 edition edition. Cambridge University Press.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2018. Mind your language: Abuse and offense detection for code-switched languages. *CoRR*, abs/1809.08652.

Filip Klubika and Raquel Fernandez. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *LREC*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. *CoRR*, abs/1808.10245.

Gordon B. Moskowitz and Peizhong Li. 2011. Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *J. Exp. Soc. Psychol.*, 47(1):103–116.

Paul Mozur. 2018. A genocide incited on Facebook, with posts from Myanmar's military. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html. Accessed: 2018-12-6.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, and Council on Communications and Media. 2011. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the Workshop on Abusive Language Online*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*.

E. Ashby Plant and Patricia G. Devine. 1998. Internal and external motivation to respond without prejudice. *J. Pers. Soc. Psychol.*, 75(3):811–832.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *COLING*.

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

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Workshop on NLP for Social Media*.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19.

Arthur K Spears. 1998. African-American language use: Ideology and so-called obscenity. In Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh, editors, *African-American English: Structure, History and Use*, pages 226–250. Routledge New York.

Luiz Valério P Trindade. 2018. On the frontline: The rise of hate speech and racism on social media. https://discoversociety.org/2018/09/04/on-the-frontline-the-rise-of-hate-speech-and-racism-on-social-media/. Accessed: 2018-12-6.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55. Springer International Publishing, Cham.

Apryl Williams and Doris Domoszlai. 2013. Black-Twitter: a networked cultural identity. *Harmony Institute*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *WWW*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Danyaal Yasin. 2018. Black and banned: Who is free speech for? https://www.indexoncensorship.org/2018/09/black-and-banned-who-is-free-speech-for/. Accessed: 2018-12-6.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of ESWC*.

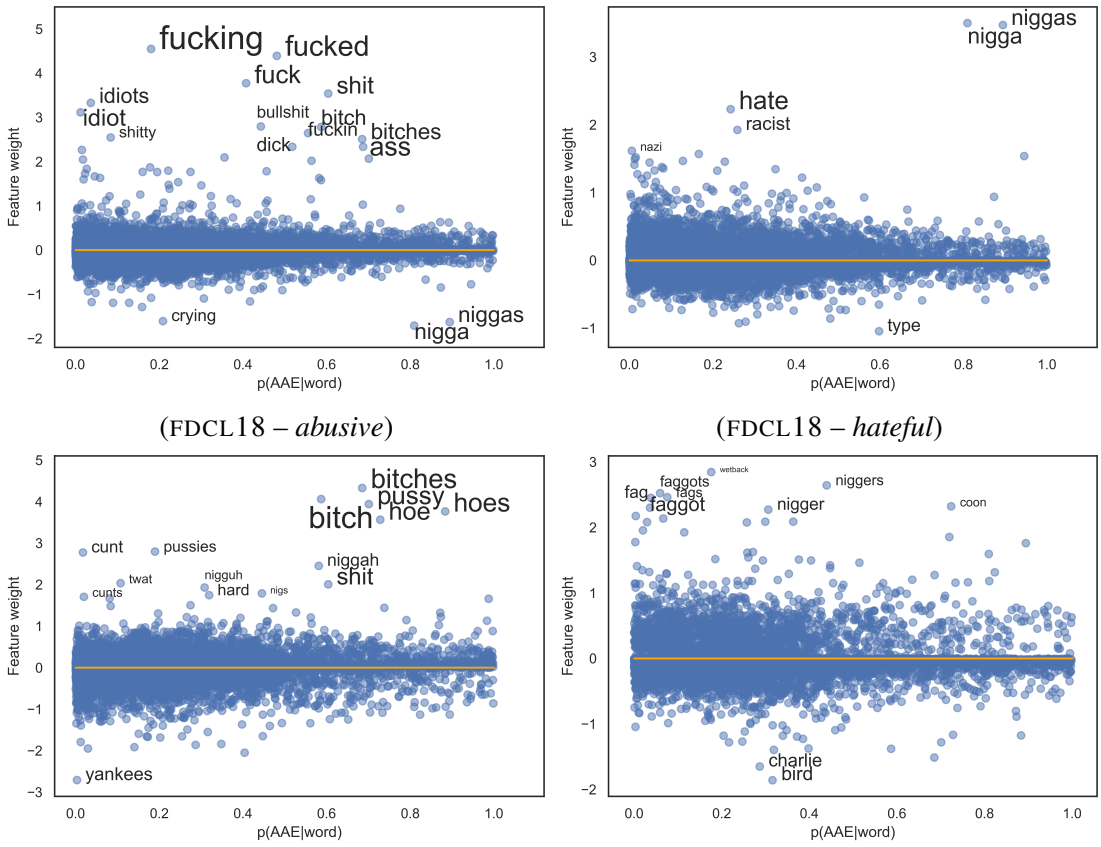


Figure 4: Feature weights learned by l_2-regularized multiclass logistic regression models with unigram features, plotted against p_{AAE} for each term, based on Blodgett et al. (2016). Top: weights for predicting *abusive* (left) and *hateful* (right) from a model trained on FDCL18. Bottom: weights for predicting *offensive* (left) and *hate speech* (right) from a model trained on DWMW17. Labels are shown for the most heavily-weighted terms, with label size proportional to the log count of the term in validation data. *Note*: "c*nt", "n*gger," "f*ggot," and their variations are considered sexist, racist, and homophobic slurs, respectively, and are predictive of hate speech DWMW17.

A Appendix

We present further evidence of racial bias in hate speech detection in this appendix.

***Disclaimer*:** due to the nature of this research, figures and tables contain potentially offensive or upsetting terms (e.g. racist, sexist, or homophobic slurs). We do not censor these terms, as they are illustrative of important features in the datasets.

A.1 Lexical Exploration of Data Bias

To better understand the correlations between inferred dialect and the annotated hate speech categories (abusive, offensive, etc.) we use simple linear models to look for influential terms. Specifically, we train l_2-regularized multiclass logistic regression classifiers operating on unigram features for each of DWMW17 and FDCL18 (tuning the regularization strength on validation data). We then use the Blodgett et al. (2016) model to infer p_{AAE} for each individual vocabulary term in isolation. While this does not completely explain the correlations observed in section §3.1, it does allow us to identify individual words that are both strongly associated with AAE, and highly predictive of particular categories.

Figure 4 shows the feature weights and p_{AAE} for each word in the models for FDCL18 (top) and DWMW17 (bottom), with the most highly weighted terms identified on the plots. The size of words indicates how common they are (proportional to the log of the number of times they appear in the corpus).

These results reveal important limitations of these datasets, and illustrate the potential for discriminatory impact of any simple models trained on this data. First, and most obviously, the most highly weighted unigrams for predicting "hateful" in FDCL18 are "n*gga" and "n*ggas", which are

WH16	% false identification			
Group	Acc.	Racism	Sexism	None
AAE	83.8	0.9	2.8	32.5
White	83.5	3.2	2.7	34.6
Overall	84.1	2.7	3.0	35.9

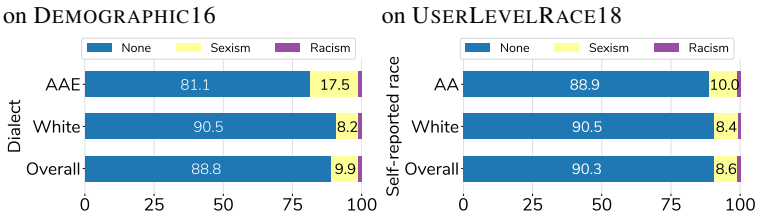


Figure 5: *Left*: classification accuracy and per-class rates of false positives (FP) on test data for the model trained on WH16. *Middle and right*: average probability mass of toxicity classes in DEMOGRAPHIC16 and USERLEVELRACE18, respectively, as given by the WH16 classifier. As in Figure 2, proportions are shown for AAE, White-aligned English, and overall (all tweets) for DEMOGRAPHIC16, and for self-identified White authors, African American authors (AA), and overall for USERLEVELRACE18.

strongly associated with AAE (and their offensiveness depends on speaker and context; Spears, 1998). Because these terms are both frequent and highly weighted, any simple model trained on this data would indiscriminately label large numbers of tweets containing either of these terms as "hateful".

By contrast, the terms that are highly predictive of "hate speech" in DWMW17 (i.e., slurs) partly reflect the HateBase lexicon used in constructing this dataset, and the resulting emphasis is different. (We also see artefacts of the dataset construction in the negative weights placed on "charlie", "bird", and "yankees" — terms which occur in HateBase, but have harmless primary meanings.)

To verify that no single term is responsible for the correlations reported in section §3.1, we consider each word in the vocabulary in turn, and compute correlations excluding tweets containing that term. The results of this analysis (not shown) find that almost all of the correlations we observe are robust. For example, the correlation between p_{AAE} and "abusive" in FDCL18 increases the most if we drop tweets containing "fucking" (highly positively weighted, but non-AAE aligned), and decreases slightly if we drop terms like "ass" or "bitch". The one exception is the correlation between "hateful" and p_{AAE} in FDCL18: if we exclude tweets which contain "n*gga" or "n*ggas", the correlation drops to r=0.047. However, this also causes the correlation between p_{AAE} and "abusive" to increase to r=0.376.

A.2 Experimental Details for Classification

For each dataset, we randomly split the data into train/dev./test sets (73/12/15%), and perform early stopping when classification accuracy on dev. data stops increasing. For DWMW17, which has multiple annotations per instance, we use the majority class as the label, dropping instances that are tied. For both datasets, we preprocess the text using an adapted version of the script for Twitter GloVe vectors.[12] In our experiments, we set $H = 64$, and use a vocabulary size of $|V| = 19$k and $|V| = 74$k for DWMW17 and FDCL18, respectively, and initialize the embedding layer with 300-dimensional GloVe vectors trained on 840 billion tokens. We experimented with using ELMo embeddings, but found that they did not boost performance for this task. We optimize these models using Adam with a learning rate of 0.001, and a batch size of 64.

category	count	AAE corr.
racism	1,976	−0.117
sexism	3,430	0.168
none	11,501	−0.064
total	**16,907**	

Table 2: Data statistics in WH16, as well as the Pearson r correlations with the labels and inferred AAE dialect. All correlations are $p \ll 0.001$.

A.3 Bias in Waseem and Hovy (2016)

We replicate our analyses in §3 on the widely used dataset by Waseem and Hovy (2016, henceforth, WH16), which categorizes tweets in three hate speech categories: *racist*, *sexist*, or *none*, shown in Table 2, along with their correlations with AAE. This dataset suffers from severe sampling bias that limit the conclusions to be drawn from this data: 70% of sexist tweets were written by two users, and 99% of racist tweets were written by a single user (Schmidt and Wiegand, 2017; Klubika and Fernandez, 2018).

[12] https://nlp.stanford.edu/projects/glove/preprocess-twitter.rb

A note on race/ethnicity of the tweet author

We also provide an estimate of the *tweet dialect*, as determined by an AI system. Previous research has showed that dialects of English are strongly associated to a speaker's racial or ethnic identity. Additionally, certain words are usually less toxic when used by a minority (e.g., the word "n*gga" or the suffix "-ass" are considered harmless in African American English), therefore it's useful to know the dialect a tweet is in before labelling it for toxic content. Our AI system detects the following dialects:

- General American English (gen Eng): associated with generic newscaster English.
- African-American English (Afr-Am Eng): dialect spoken usually by African-American or Black folks.
- Latino American English (Lat Eng): dialect spoken usually by Latino/a folks both in New York and California, Texas, Chicago, etc.

(*dialect priming*)

A note on race/ethnicity of the tweet author

We also provide an estimate of the Twitter user's race or ethnicity, as inferred by our AI system. Note that certain words are usually less toxic when used by a minority (e.g., the word "n*gga" or the suffix "-ass" are considered harmless when spoken by Black folks), therefore it's useful to know the identity of a Tweeter before labelling it for toxic content.

(*race priming*)

Figure 6: Additional instructions shown to workers in the *dialect* and *race* priming. In the dialect condition, we provide links to the dialect Wikipedia pages.

In Figure 5 (left), we show how models trained on this dataset have slightly higher false positive rates of sexism on AAE tweets, and of the "none" label for White tweets compared to AAE tweets. When predicting on our reference corpora (Figure 5, middle and right), we see AAE tweets (or tweets by African Americans) are labelled as sexist more than White-aligned tweets or tweets by White users. Again, due to the sampling issues, these results should be interpreted cautiously.

A.4 Toxicity and AAE in the PerspectiveAPI

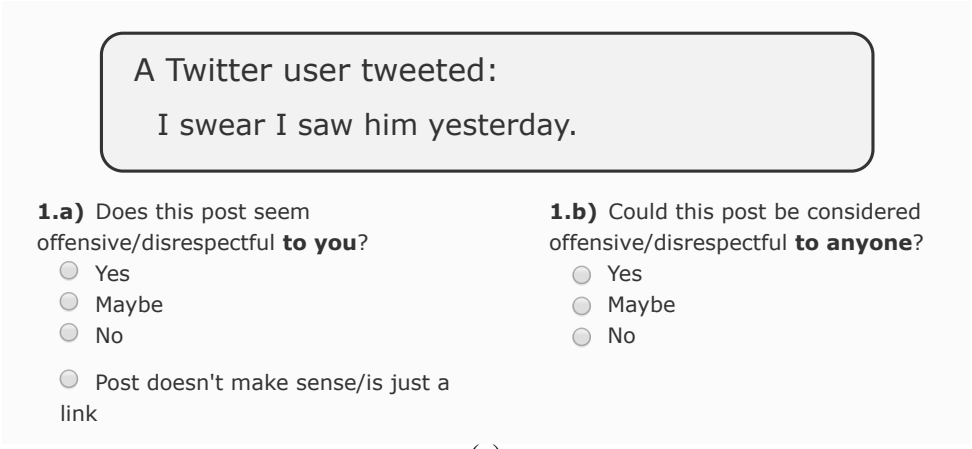
We compare the toxicity detection rates from our trained models to those of PerspectiveAPI,[13] a Jigsaw/Alphabet initiative to detect hate speech online. We obtain TOXICITY scores for all tweets in DWMW17 and FDCL18, as well as for 100K random tweets from DEMOGRAPHIC16 and USERLEVELRACE18. We show correlations (Pearson r) between dialects/groups in our datasets and the Perspective TOXICITY scores. All correlations are significant ($p \ll 0.001$, Holm-corrected for multiple comparisons) which indicates potential racial bias for all datasets.

dataset	dialect/group	TOXICITY corr.
DWMW17	White	−0.320
	AAE	0.310
FDCL18	White	−0.340
	AAE	0.453
DEMOGRAPHIC16	White	−0.096
	AAE	0.056
USERLEVELRACE18	White	−0.046
	AA	0.042

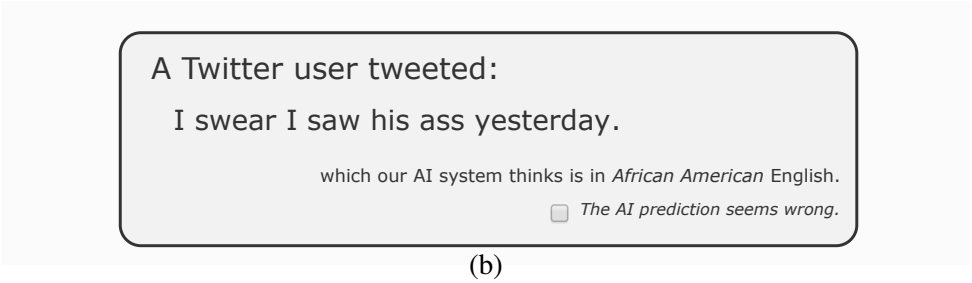
[13]The API (http://perspectiveapi.com) was accessed in December 2018

A.5 Dialect Priming Experimental Details

We collected annotations from 110 (76% White), 143 (77% White), and 81 (72% White) workers in the control, dialect, and race priming conditions, respectively. Figure 6 shows the instruction snippet related to dialect and race shown to workers in the two treatment conditions. Additionally, Figure 7 shows the annotation interface, with (a) and without priming (b,c).



(a)



(b)

(c)

Figure 7: Interface for the controlled experiment. (a) shows the control condition along with the offensiveness questions. (b) and (c) show the changes to the treatment interface in the *dialect* and *race priming* conditions.