Surfacing Racial Stereotypes through Identity Portrayal

Gauri Kambhatla*
gkambhat@utexas.edu
University of Texas at Austin
Austin, TX, USA

Ian Stewart
ianbstew@umich.edu
University of Michigan
Ann Arbor, MI, USA

Rada Mihalcea
mihalcea@umich.edu
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

Content warning: this paper discusses and contains content that may be offensive or upsetting.

People express racial stereotypes through conversations with others, increasingly in a digital format; as a result, the ability to *computationally* identify racial stereotypes could be beneficial to help mitigate some of the harmful effects of stereotyping. In this work, we seek to better understand how we can computationally surface racial stereotypes in text by identifying linguistic features associated with differences in racial identity portrayal, focused on two races (Black and White). We collect novel data of individuals' self-presentation via crowdsourcing, where each crowdworker answers a set of prompts from their own perspective (real identity), and from the perspective of another racial identity (portrayed identity), keeping the gender constant. We use these responses as a dataset to identify stereotypes. Through a series of experiments based on classifications between real and portrayed identities, we show that generalizations and stereotypes appear to be more prevalent amongst white participants than black participants. Through analyses of predictive words and word usage patterns, we find that some of the most predictive features of an author portraying a different racial identity are known stereotypes, and reveal how people of different identities see themselves and others.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Social and professional topics** → *Race and ethnicity*; • **Human-centered computing** → *User studies.*

KEYWORDS

datasets, racial bias, stereotypes

ACM Reference Format:
Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. 2022. Surfacing Racial Stereotypes through Identity Portrayal. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533217

1 INTRODUCTION

People of color in America encounter both explicit and implicit racial stereotypes and discrimination every day. Racial stereotyping does not occur in isolation, but rather when people communicate; we see this everywhere from policing – where White community members have been shown to be much more likely to hear a more respectful utterance (e.g., expressing apology or gratitude) and Black members have been shown to be more likely to hear a disrespectful utterance (e.g., "hands on the wheel") [51] – all the way to sports commentary, where White players have been shown to be mostly praised for their personality or intelligence (e.g., "spectacular", "cool", "calm", "smart"), while Black players have been shown to be mostly praised for athleticism (e.g., "speed", "versatile", "beast") [36]. While the speaker may mean no harm, such language, resulting from implicit biases, can lead to serious consequences such as wrongful imprisonment [23] or microaggressions, which can be detrimental to physical and mental health [4, 10, 21].

The extent of these communication-driven incidents, as well as the depth of their consequences has been made increasingly clear in recent years, exposing the high degree of institutionalized racism and calling for more action against it. While many of these issues require interdisciplinary long-term solutions, it is important to understand racial bias in language itself to attempt to prevent harm at the *individual* level as well as the *social* level. In today's world, although much of our communication is digital, racial bias is no less prevalent [13, 16, 38, 53]. Given the vast number of digital interactions, it would be beneficial to *computationally* identify racial biases and stereotypes to help mitigate some of the consequences described above; in particular, computational methods can help identify bias at a *large* scale, which would be next to impossible for humans to identify manually. More specifically, understanding racial stereotyping in language could help provide training for occupations that require high levels of interaction with minority populations and are highly influential on these populations' lives, such as reporters, teachers, social workers, and police officers.

A single definition of "stereotype" has been elusive in psychology research due to inconsistencies, such as whether stereotypes are inherently bad, whether they are individual or consensual beliefs, and whether stereotypes are necessarily inaccurate descriptions [29]. In light of these inconsistencies, in this paper we define stereotypes broadly as cognitive generalizations (such as beliefs or expectations) about the characteristics of a group or social category, based on the American Psychological Association's definition.[1]

In this work we seek to better understand how we can computationally surface racial stereotypes in text by identifying linguistic features associated with differences in racial identity portrayal. Here, we focus on two racial identities (Black and White), since

[*] Work done while the author was a student at the University of Michigan

[1] https://dictionary.apa.org/stereotype

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533217

racial tensions have intensified between these two groups in recent years, apparent through the Black Lives Matter movement [30, 35]. To limit cultural bias, we focus on individuals currently residing in the United States. We address the following research questions in this paper:

- RQ 1. Given text written by people of different racial identities, can we distinguish between an author writing as themselves, and the same author writing from the perspective of a different racial identity? With this question, we seek to explore what stereotypes and generalizations people of specific racial groups make (e.g., what stereotypes do White men make? Black women?).
- RQ 2. Given a dataset of text of real and portrayed racial identities, can we distinguish between two authors of different race writing from the perspective of the same racial identity? With this research question, we aim to identify the stereotypes that individuals have about a specific racial group; e.g., given the social group of Black women, a classification task between writing of real Black women and individuals "pretending" to be Black women might help us identify stereotypes these individuals have about Black women.
- RQ 3. How well do the differences in language use between people of different racial identities reflect known stereotypes? We compare the language between different racial groups, and whether we can identify stereotypes based on the words people use.

Our main contributions in this paper are as follows:

- A novel dataset of implicit stereotypes compiled via crowdsourcing, grouped by racial and gender identities
- An experimental methodology for surfacing stereotypes from text
- An analysis of linguistic features associated with stereotypes about various racial identities

While this study uses a controlled experimental setting to identify stereotypes, the implications can extend to a variety of other domains. The method for eliciting stereotypes can be readily extended to real-world decision-making settings, particularly teaching and law where racial stereotypes can impact judgments about a student's performance or a defendant's intentions [44]. The stereotype-detection models developed in this work can help identify bad-faith actors in online discussions, such as bots who spread misinformation online by acting as a member of a minority community [24]. Finally, our work confirms the presence of known stereotypes (e.g. "basketball" used by White men to describe Black men), as well as the possibility of more subtle or complicated social stereotypes that reflect different racial attitudes than expected.

2 RELATED WORK

Racial bias in language has been studied extensively. Prior research specifically in natural language processing (NLP) has considered race in different ways, including: addressing racial bias in datasets and data labels, in model architectures, in model outputs, and in social analysis through NLP models [20]. Here, we first discuss how racial bias has been shown to be reflected in language, before focusing on prior work in racial bias and stereotype detection. Significant efforts have been made in detecting bias in text representations (embeddings) and text generation. Other work has focused on racial bias detection in abusive language or hate speech detection. In addition, some work concentrates around detecting bias in social contexts such as social media.

2.1 Racial Bias in Language

Previous work in sociolinguistics and linguistic anthropology have shown how racial bias is not only reflected, but also propagated, through language. For instance, Purnell discusses how attitudes about race and racial prejudices are reflected in language, as well the fact that language can reinforce or sustain such biases; these can be seen in writing as seemingly innocuous as newspaper advertisements (e.g., "the clean new look is milk white.") to writing guides such as dictionaries or thesauri, which can list mostly positive synonyms for "whiteness", and mostly negative synonyms for "blackness" [42]. In addition, racial biases can be propagated in a variety of ways, such as in the medical domain, where Goddu and colleagues show that stigmatizing language in patient reports can affect physicians' attitudes towards patients [39], or in legal language, where Rice and collaborators show that African American names are more likely to be associated with negative concepts and White names are more likely to be associated with positive concepts in US state and federal court opinions [45].

2.2 Racial Bias in NLP

Word embeddings trained on large corpora can learn implicit racial bias by associating race with harmful stereotypes, leading to work that studies bias detection and debiasing embeddings [6, 9, 28, 40]. While most of this work focuses on gender bias, some also include race. In particular, Manzini et al. extend the word embedding debiasing methodology of Bolukbasi and colleagues [6] to a multi-class setting, where they focus on three classes of bias: gender, race, and religion [34]. Other works analyze bias at the intersection of race and gender in embeddings [27, 31]. Lepori examines race and gender bias in embeddings through the lens of intersectional theory [14], and shows that the theory (that the "multiple different aspects of a person's identity often combine to create unique modes of discrimination") holds true in contextualized and non-contextualized word embeddings [31]. Jiang and Fellbaum also highlight the importance of studying racial and gender bias together in NLP by showing the interdependencies of the individual demographics in PCA projections of BERT embeddings [27].

In addition to investigating bias in latent word representations, NLP researchers have also investigated the impact of racial bias on downstream applications, such as abusive language detection. Sap et al. and Waseem both discuss how annotators are more likely to incorrectly label African American English (AAE) as toxic compared to Standard American English (SAE), which can lead to biased corpora [48, 52]. This bias is propagated through models trained on these datasets, which leads to AAE language to be much more likely to be predicted abusive than SAE [17, 48]. Similarly, Zhang and colleagues found that language which uses certain demographic terms, such as "black,""Mexican," or "gay", in any manner is more likely to be classified as offensive simply for including these words because most abusive language detection datasets contain toxic language with these terms [54].

2.3 Stereotype Detection

More closely related to the data collection methodology of this paper, other works look specifically at detecting stereotypes in language. Cryan et al. look at detecting gender stereotypes in a comparison of lexicon-based and end-to-end approaches, creating a dataset through online articles which describe men or women, and are either consistent with, or contradictory to traditional gender stereotypes [15]. While this method provides valuable insights into approaches to detect stereotypes or anti-stereotypes about gender, it relies on crowdworkers to identify articles which contain stereotypes, rather than direct model predictions. In a different angle, Carpenter et al. center on quantitatively differentiating between true generalizations and inaccurate stereotypes about various demographic categories, including gender, age, education, and political orientation [11]. Participants guessed a tweet's author's group membership for a collection of tweets, and inaccurate stereotypes were determined by the correlation of terms with the proportion of participants who falsely believed an author to be of a group. This method surfaced interesting beliefs about demographic groups, but did not discuss race. Similar to our work, Ali and colleagues concentrate on the detection of racial bias in language, although the focus is specifically on microaggressions [1]. The authors use classic machine learning approaches to identify racial microaggressions from non-racial microaggressions in a binary classification task, where the authors label samples in a corpus of microaggressions as "racial" if they contain particular terms. Though focused specifically on extracting racial bias, this study provides little analysis beyond classification results.

Our work builds upon prior work in the space of race in NLP with a focus on discovering racial stereotypes in naturally occurring text, rather than detecting bias in NLP systems or tasks (such as in word embeddings or abusive language detection). In comparison to work that aims to detect biases more broadly, we seek to identify authors' implicit biases with respect to race.

3 DATA

As we seek written language from individuals posing as themselves and as a false identity, we gather a corpus of written samples from crowd-workers on the Prolific[2] platform. We targeted four identities: 1. Black woman, 2. Black man, 3. White woman, 4. White man. We decided to focus on Black and White racial identities within the United States because of prominent racial tensions between these two groups. Within these racial groups, we focus on the two most common gender identities to gather the most participants. Participants were selected if they were English speakers and self-identified as one of the four above identities, and were given a survey of the *other* racial identity and *corresponding* gender identity (e.g., an individual who self-identifies as a White woman is given a survey which asks her to imagine herself as a Black woman, in addition to writing as a White woman).

We created four surveys using the Qualtrics software, where each survey corresponds to a specific identity (one of the four listed above). In each survey, participants are asked to write responses to the following prompts, which were pilot-tested on other Prolific workers to ensure participants understood the task and responded appropriately. The surveys were given to all participants with the same question order.

(1) *Please describe yourself. Write a full paragraph of 5-6 sentences or more.*
(2) *Please describe your typical evening on a workday, after a day at work or school. Write a full paragraph of 5-6 sentences or more.*
(3) *Imagine you are a **White/Black man/woman**, the same age as you. Please describe yourself. Write a full paragraph of 5-6 sentences or more. Please write from the first-person perspective of a **White/Black man/woman**. Others will read what you wrote; your goal is to convince them it was written from the perspective of a **White/Black man/woman** without saying so explicitly. For example, do not write a sentence like "I am a **White/Black man/woman**" as this is an explicit statement.*
(4) *Imagine you are a **White/Black man/woman**, the same age as you. Please describe your typical evening on a workday, after a day at work or school. Write a full paragraph of 5-6 sentences or more. Please write from the first-person perspective of a **White/Black man/woman**. Others will read your writing; your goal is to convince them it was written from the perspective of a **White/Black man/woman** without saying so explicitly. For example, do not write a sentence like "I am a **White/Black man/woman**" as this is an explicit statement.*

Prompt (3) is designed to be more open ended to allow participants some creativity. We hypothesize a prompt like this may surface assumptions or stereotypes relating to identity, personality, and intelligence. Prompt (4), while also open-ended, is designed to be more specific than the first. This prompt might lead to stereotypes relating to work, home-life, family, friends, and/or leisure activities. The responses to both of these prompts are combined for each individual in our final dataset. We choose this method for our data collection as a way to (1) surface *implicit* stereotypes, i.e.,beliefs that participants do not realize are considered "wrong" by society, and (2) elicit stereotypes participants might not admit to in normal circumstances; our method allows participants to write down their thoughts privately and anonymously, creating an environment where they are more likely to share their thoughts freely [32, 47].

After filtering the data collected to remove inadequate or incomplete responses, we use the responses from 615 individuals. We remove a total of 265 responses from individuals who started the survey but did not finish, or who completed the survey, but did not answer answer the prompts appropriately. Two (partial) sample responses can be seen in Table 1. The class distribution of racial and gender identities, as well as the average word counts (with standard deviations) of each identity are shown in Table 2.

In addition to the prompt responses, we also collect age, location, and education statistics about participants. The average age of participants (shown in Table 2) is around 30 years old, although individuals' ages spanned from 18 to over 72. The percentage of participants from each identity group with a higher education degree (bachelor's, master's, and/or doctoral) can also be seen in Table 2.

[2] https://prolific.co/

The state distribution of participants' location of residence can be seen in Figure 1. As expected, the largest number of individuals come from the most populous states, including California, New York, Texas, and Florida.

4 DESIGN AND MODELS

4.1 Data Processing

We apply basic preprocessing to our text, including tokenization and lowercasing. We decided not to remove stopwords, as they have been shown to be psychologically meaningful [12]. In addition, to make sure our results are indeed due to racial identity and not deception (i.e., from participants pretending to be a different racial identity, not just from pretending to be a different *person*), we first remove highly deceptive features from our data. We train the same model we use in our experiments (described below) using term frequency, inverse document frequency (TF-IDF) features on a previously introduced deception data set of age and gender deception [43] to find features most associated with identity deception. In this task, we perform a classification between "real" and "portrayed" identities, and obtain a ranking of the top predictive TF-IDF features. We remove the top 30 predictive features from our racial identity data vocabulary and train our models on this new vocabulary.

4.2 Models and Features

Our classification experiments are performed using the models and features described below.

Models. The baseline we compare our models against is the majority class, i.e., the performance of a model that always predicts the majority class of the data. We use a Linear SVM as our main classifier, with no regularization or class weights, squared hinge loss, and L2 norm. We choose to use a very simple classifier for interpretability; while we also experimented with using BERT [18] (a common deep neural net language model) for our classification tasks, we found that performance was comparable to the SVM classifier while using the best features. However, BERT (and other transformer-based LMs) are not interpretable, which is an essential criterion for our study. Each experiment is conducted using leave-one-out cross validation, and the results in our tables are the average of each iteration.

Features. The following features are extracted from our data and are used to create our SVM classifiers. We hypothesize that each of these features will be useful for a model to distinguish different identities in each of our classification tasks.

Ngrams. Unigrams and bigrams extracted from a bag of words representation of the text data.

Ngrams + POS tags. Unigrams and bigrams extracted from a bag of words representation of each token in the dataset appended to its corresponding part of speech, where the POS tags were extracted using the Stanford CoreNLP parser.

LIWC. The semantic categories provided by the 2007 LIWC lexicon [41]. The feature vector for an individual participant's response is the number of words in each semantic category, normalized by the length of the response.

Word2Vec. 300 dimensional word2vec embeddings trained on our data, where the vector for a participant's response is the mean of the vectors of each word in the response.

Lexical Diversity. These features include: 1) type/token ratio; 2) mean word frequency; 3) Yule's I index (a metric which is proportionate to the number of types weighted by their frequency), i.e., the larger the value, the smaller the lexical diversity; and 4) Yule's K index (a metric based on the reciprocal of Yule's I index), i.e., the larger the value, the larger the lexical diversity. The feature vector for an individual response consists of these four values. We included this feature as lexical diversity has been shown to be a perceived measure of status [7], which might indicate that participants who write less diversely might do so as a method of stereotyping.

Readability. These features include: 1) Flesch-Kincaid, 2) Flesch Reading Ease, 3) Gunning Fog, and 4) Automatic Readability Index (ARI). As with the lexical diversity metrics, these four values were used to create a feature vector for each participant's response.

TF-IDF. The term frequency, inverse document frequency (TF-IDF) features for each response, where a "document" is one response, and the vocabulary consists of unigrams.

5 RQ 1: STEREOTYPES *FROM* RACIAL IDENTITIES

Our first research question is concerned with the degree to which we can identify stereotypes from particular racial groups; concretely, we address this by distinguishing between an author's "real" (participants answering prompts as themselves) and "portrayed" (participants answering prompts from a different racial identity perspective) identities. To answer this question, we perform classification tasks between "real" and "portrayed" identities. In particular, using the classification models described in Section 4.2, we attempt to differentiate between real and portrayed racial identities at different levels of granularity: for all examples in our dataset, for only Black identity examples, only White identity examples, and each of Black women, White women, Black men, and White men individually. The accuracy of our classifiers for each of these settings are shown in Table 3, and compared to the majority class baseline. Except for the baseline, all models shown in the table use an SVM classifier. All values are statistically significant, using the Wilcoxon signed-rank test, in comparison with the baseline ($p < 0.001$).

Nearly all the models perform better than the baseline, except the SVMs with lexical diversity and readability features, showing that there is a significant difference in individuals' writing as their own racial identity versus a different racial identity. In addition, the N-gram and TF-IDF features perform very well compared to the others; this suggests that distinguishing between real and portrayed identities relies significantly on specific keywords or phrases, which could be related to stereotypes. We look at some of these keywords in Section 7. In general, we also see higher accuracy scores for participants whose real race is White than for participants whose real race is Black; in fact, for *every* model (excluding those trained with lexical diversity and readability features, which did not perform better than the baseline), the score for White individuals – women, men, and both combined – is higher than that for the corresponding Black individuals, by an average of 10.35% and a largest difference of 20.06%. This means it is easier to determine whether a White individual is writing as themselves, or from the perspective of a

Table 1: Sample responses

	Real Identity	**Portrayed Identity**
	Black Woman	White Woman
Participant 1: (self-identifying) Black woman	I am a Black woman born and raised in Alabama. I have been married for 5 years to my husband and we have a 2 year old daughter and 1 year old son. I am a stay at home mom. I am also disabled. I have two autoimmune disorders. My dimpled smile is my favorite thing about me...	I'm a southern soccer mom who loves to live, laugh, love! I have a wonderful husband and a dear son and daughter. I'm a blond little thing with a big smile and bigger hugs. I am the life of the party after a few glasses of wine. I go to yoga every morning to keep in shape...
	White Man	Black Man
Participant 2: (self-identifying) White man	Born in Iowa, raised in las Vegas, lives in Minneapolis. Only child, not much family. 37 years of age, body feels more like 50. Married for 5 years. Home owner, no kids, steady full time job. Interests are record collecting and Tiki culture. My wife has 2 cats so u guess I have 2 cats...	Ight, check it. I finna roll up on this couch with my kicks up high. Da f*** the remote at? Who be calling my phone, dats my girl, yo. Lemme spit this out. My girl be hungry, but I'm like naw. I ate and s***. Yo check it, these cats is hungry. Then Simpsons be playing while I trip out and s***...

Table 2: Dataset statistics. For average word count, "real" indicates the author writing as themselves, and "portrayed" indicates the author writing as an alternate identity.

Real Identity	**Number Responses**	**Average Word Count (real)**	**Average Word Count (portrayed)**	**Percent with College Degree**	**Average Age**
Black Women	154	232.08 ± 76.14	224.44 ± 44.54	59.1%	34.2
White Women	152	234.05 ± 53.28	223.39 ± 47.37	59.1%	36.8
Black Men	151	224.33 ± 56.57	224.69 ± 50.22	75.4%	31.8
White Men	158	217.51 ± 34.47	218.69 ± 34.55	65.8%	38.1

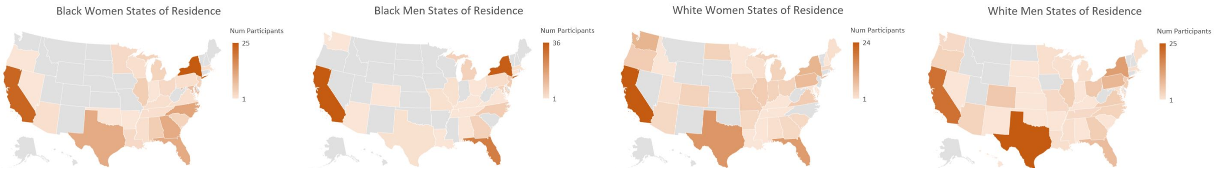
Black Women States of Residence
Num Participants
25
1
Black Men States of Residence
Num Participants
36
1
White Women States of Residence
Num Participants
24
1
White Men States of Residence
Num Participants
25
1

Figure 1: Crowd-sourced participants' locations of residence. Darker colors represent more participants, and gray indicates zero participants.

portrayed identity, which suggests White individuals write in a manner more significantly different as a Black identity than vice versa. This could imply that White participants see Black individuals as more different from themselves than Black participants see White individuals, potentially hinting at racial bias, or an inclusion of stereotypes.

6 RQ 2: STEREOTYOES *TOWARDS* RACIAL IDENTITIES

Our second research question asks whether we can identify stereotypes *about* specific groups of people by predicting an author's true racial identity. In order to explore this question, we conduct a classification experiment where we attempt to predict the true racial identity of an author given a response with a supposed racial identity; for example, given a supposed "Black" response, the classifier must predict whether the author identifies as Black, or is a White author portraying a Black identity.

As in the previous experiment, we use various granularities of data: all "Black" responses and all "White" responses (both real and portrayed), as well as split by our two gender identities. In other words, we had six different tasks; predicting the author's true racial identity knowing the text is *supposedly* written from the perspective

Table 3: Accuracy values for classification between real or portrayed identity for an individual participant response, significant in comparison with the baseline at $p < 0.001$. The best-performing SVM for each classification setting is bolded.

Model & Feature	All Identities	Real Black			Real White		
		All	Woman	Man	All	Woman	Man
Baseline	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Ngrams	75.77	73.11	72.40	**63.58**	79.52	78.62	77.85
Ngrams + POS	76.99	72.13	75.00	62.58	79.35	79.93	77.53
LIWC	61.38	60.16	62.34	52.98	68.55	65.46	69.62
Word2Vec	72.03	67.21	68.83	60.60	76.13	76.97	75.00
Lex Div	49.84	53.11	47.72	51.66	51.94	49.34	53.16
Readability	53.33	49.34	48.70	49.34	50.00	49.34	52.21
TF-IDF	**78.37**	**74.10**	**78.90**	61.92	**84.52**	**83.88**	**81.01**
All features	**78.37**	72.30	73.05	61.59	81.61	79.93	81.65

of a 1) Black individual, 2) White individual, 3) Black woman, 4) White woman, 5) Black man, and 6) White man. We conducted the classification experiments using the models described in Section 4.2. The accuracy scores we obtain for each of these experiments, as well as the corresponding majority class baseline values, are shown in Table 4. Except the baseline, all other models in the table are SVMs. All values are statistically significant in comparison with the baseline ($p < 0.001$).

Similar to the first experiment, every model, except the SVMs trained with lexical diversity and readability features, performs better than the baseline. Similarly to our previous experiment, the models that use N-gram and TF-IDF features perform very well compared to other features like Word2Vec and LIWC, which suggests that differences between real and false portrayals of a racial identity – the area where generalizations or stereotypes might be made – are based more on word occurrence and particular keywords. The models achieve higher accuracy scores when predicting portrayed Black identities – women, men, and combined – than we do when predicting portrayed White identities. As we can see in Table 4, the SVM models perform better for portrayed Black identities than portrayed White ones in 15/18 settings (not including lexical diversity or readability features) by an average of 5.65%, and a largest difference of 12.3%. This indicates that it is easier to identify a White author posing as a Black author than vice versa, which could mean our data contains more stereotypes about Black individuals.

7 RQ 3: ANALYSIS OF MOST PREDICTIVE LINGUISTIC FEATURES ASSOCIATED WITH DIFFERENT RACIAL IDENTITIES

Our final research question asks how we can learn about stereotypes from differences in language between people of different racial identities. To explore this question, we analyze differences in word usage between each racial identity, both at the word level, and word category level.

7.1 Analysis of Language at Word Level

At the word level, we obtain the most predictive TF-IDF features from our author prediction task, described in the previous section. This analysis helps explain our classification results and allows for further insights by showing which words the classifier deemed most important in predicting the racial identity of the author of a given text. We choose to use TF-IDF features here since our SVM trained on these features was generally the best-performing model.

The most predictive features, as well as their feature weights given by the SVM, are listed in Tables 5, 6, 7, which correspond to the most predictive features for all individuals of each portrayed racial identity, the most predictive features for *women* of each portrayed racial identity, and the most predictive features for *men* of each portrayed racial identity, respectively. We define *the most predictive features* as the k words (we use $k = 10$) to which the classifier assigns the highest weight, i.e., the words that are the most important (predictive) in helping the classifier predict the true racial identity of an author that is claiming to be Black or White. Given our method of classification, note that each set of classification weights (e.g., Real Black and White as Black) are comparable, but these are not comparable between sets. We group the most predictive features that we consider to fall under our definition of stereotype into the following categories: appearance, interests, residence, and identity. We provide a qualitative analysis on each below. Given the nature of this analysis, we emphasize that there are multiple ways of interpreting some of these results, though we only provide a few possible explanations here.

Appearance. A person's physical appearance has long been known to be a basis for stereotypes [33]. The term that falls predominantly in this category is "hair," which we see as predictive when White individuals write as Black individuals, and also specifically when White women write as Black women. The fact that "hair" is one of the most predictive features that an author of portrayed Black writing is actually White seems to indicate that while White people consider hair to be an important aspect in a Black person's life, Black people do not consider it significant enough to write about it. This can be shown in the frequency of "hair" in the responses; while Black participants use the term 18 (12 for women, 6 for men) times, White participants (while portraying a Black identity) use it 54 times (51 for women, 3 for men) times. This confirms well-known stereotypes about Black hair being viewed as inherently separate

Table 4: Accuracy values for classification of true racial identity given an identity response, significant in comparison with the baseline at $p < 0.001$. The best-performing SVM for each classification setting is bolded.

Model & Feature	Portrayed Black			Portrayed White		
	Total	Woman	Man	Total	Woman	Man
Baseline	50.41	50.33	51.13	50.41	50.33	51.13
Ngrams	86.18	78.76	88.03	82.28	79.74	75.73
Ngrams + POS	85.69	78.43	83.50	79.84	78.10	73.79
LIWC	67.97	68.95	69.26	65.04	66.01	61.49
Word2Vec	75.45	74.51	77.67	74.15	76.47	68.93
Lex Div	52.03	50.65	49.19	47.80	47.71	48.22
Readability	53.50	47.39	57.28	53.50	51.96	51.46
TF-IDF	**89.11**	81.70	**89.32**	**84.39**	**85.29**	**77.67**
All features	86.99	**83.33**	84.79	78.54	78.10	73.46

Table 5: Most predictive TFIDF features (with weights) in classification of true racial identity given a particular racial identity, for all responses.

Real Black		White as Black		Real White		Black as White	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
american	1.65	we	1.46	movies	1.06	the	1.21
typical	1.36	friends	1.34	year	1.05	with	1.18
african	1.14	basketball	1.28	old	1.04	work	1.17
reading	1.13	bed	1.26	small	1.01	at	1.13
also	1.11	hair	1.14	time	1.00	call	1.06
an	1.04	college	1.13	dinner	0.96	bath	1.06
new	1.01	up	1.12	male	0.95	order	0.92
evening	0.98	race	1.09	chores	0.92	wine	0.91
two	0.97	hard	1.07	many	0.89	white	0.90
shower	0.95	when	0.99	animals	0.85	race	0.90

or different from other hair types, non-conforming to traditional beauty standards [46].

Interests. Perhaps one of the most common categories of interests is sports, and we see multiple features that reflect this; "basketball" is one of the most predictive features of White individuals (and specifically White men) writing as a Black identity, highlighting a well-known stereotype [8] that has been used to appeal to Black Americans in domains such as advertising [2, 22]. Conversely, we see "golf" as a predictive feature for Black men writing as White men, emphasizing the stereotype that White men like to golf.

Focusing on leisure activities, the term "bath" is a top predictive feature of Black individuals writing as White, though Black women use the word the same number of times when writing from their own perspective and that of a White woman (28 times), and Black men use it more when writing from their own perspective (22 compared to 12). In contrast, White women use "bath" only eight times, and White men only three times. This seems to indicate that "bath" is more highly relevant in Black individuals' vocabulary. We also see "call", which appears to be an assumption that White individuals make a lot of phone calls (91% of the uses of "call" for Black women writing as White women is for making phone calls). It also appears to be a stereotype that White men play Call of Duty (one-third of the usages of "call" in Black men writing as White men are for this video game). For Black women writing as White women, we also see "wine" and "starbucks" as predictive features, surfacing common pop-culture stereotypes that White women drink wine and Starbucks beverages [19, 50]. Both White women and Black women writing as themselves consider the same sort of activities in their daily lives, such as playing "games" and "reading"; indicating that for self-descriptions among women, leisure activities appear to be similar regardless of racial identity.

In addition, "work" comes up as a top predictive feature for Black men writing as White men; while not unusual in itself (the prompt asked to describe your day after work), that it is predictive of this portrayed identity seems to indicate a stereotype that work (an occupation) is important in the lives of White men (e.g., "I work at a big firm, which is my dream job..." and "Right now i work as a manager for this really big fortune five hundred company").

Residence. The stereotypes in this category are centered on home-life. The word "local" comes up as a predictive feature for White

Table 6: Most predictive TFIDF features (with weights) in classification of true racial identity given a particular racial identity, for women.

Real Black		White as Black		Real White		Black as White	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
american	1.04	we	1.55	watch	0.84	wine	1.01
prepare	0.97	hair	1.26	or	0.79	at	0.88
reading	0.96	college	1.24	animals	0.77	call	0.79
games	0.83	friends	1.02	and	0.75	off	0.79
usually	0.78	bed	0.84	sometimes	0.73	with	0.78
online	0.77	when	0.81	playing	0.72	starbucks	0.73
also	0.74	up	0.80	games	0.72	me	0.71
african	0.73	hard	0.80	also	0.71	bath	0.71
typical	0.73	race	0.74	dinner	0.69	glass	0.70
then	0.72	they	0.73	will	0.68	city	0.68

Table 7: Most predictive TFIDF features (with weights) in classification of true racial identity given a particular racial identity, for men.

Real Black		White as Black		Real White		Black as White	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
typical	1.18	get	1.39	year	1.05	the	1.45
evening	1.04	basketball	1.01	to	1.02	with	0.97
american	1.00	we	0.89	male	0.99	work	0.87
an	0.97	race	0.85	old	0.97	white	0.86
two	0.89	skin	0.77	movies	0.89	man	0.83
also	0.81	white	0.76	like	0.81	you	0.77
take	0.78	local	0.74	many	0.81	race	0.73
new	0.78	friends	0.72	time	0.78	hanging	0.73
day	0.71	watch	0.72	small	0.77	black	0.71
computer	0.69	color	0.67	cook	0.74	golf	0.68

men writing as Black men, in contexts such as the "local pool", "local grocery store", "local gym", "local bar", etc. This could imply an implicit assumption that Black men contain themselves to their locality, or neighborhood. While this might be generally true of all individuals (e.g. most people get groceries from the store nearest their home), the word "local" as a predictive feature might suggest that White men believe Black men value a more tight-knit community, perhaps because minorities tend to group together when in predominantly white settings [26, 37].

The word "order" is a predictive feature of Black individuals portraying White individuals, and is used mostly in the context of ordering food, drinks (e.g., coffee), a car ride (i.e., ride-share services) or online products (i.e., online shopping). This could be a stereotype that White individuals do not cook much, shop online, or are wealthy enough to order takeout or coffee on a daily basis. In addition, the word "city" suggests that our Black women participants assume White women generally live in more urban areas (e.g., "I moved to the city after growing up in a small town..." and "I live in New York City"), while White women writing as themselves rarely use the word (two of the three times they used "city", it was qualified by "small").

Identity. The features in this category are stereotypes related to a person's racial identity. Some of the predictive words of White participants portraying Black participants are stereotypes that relate to what they perceive as important to a Black identity, such as "race", "hard", and "college". White people assume, correctly or not, that their race plays an important part in Black people's lives (e.g., "relations with police and people of my race are tense...", and "life is just not fair to me because of my race"), that aspects of life are hard and Black individuals must work hard (e.g., "my life is very hard work"and "I'm a hard working person who doesn't like to give up"), and that college is significant (e.g., "I am a first generation college student" and "I'm very involved on my college campus as well as in my community"). This gives us an idea of what White participants think is worth discussing in the life of a Black person, or the attributes that are highly visible to White individuals in their own lives and through media.

Another noteworthy word is "white"; White men include this while writing from the perspective of Black men, evoking a sense of "other" (e.g., "I did not get along much with the white kids"), and illuminating how White men believe Black men see people of their own racial identity. Indeed, prior work has shown that White individuals are aware that racial minorities see them as prejudiced and close-minded [5]. Similar to "white" being predictive of White men writing as Black men, "black" is predictive of Black men writing as White men, mostly used in the context of bias against themselves (e.g., "I... hate to deal with the black employees..."), which elucidates how Black men believe the stereotypical White man sees them.

In addition, "african" and "american" are top features when Black participants write as themselves, but we do not see anything corresponding (e.g., "white", "caucasian", etc.) when White participants write as themselves (though we do see "white" when Black individuals write as White individuals). This seems to agree with sociology studies that show that White participants are far less likely to think that race is very important to how they think about themselves than other racial groups, i.e., they have a lower salience of racial identity [25, 49]. The term "male" is also one of the most predictive features for White individuals writing as themselves, despite the fact that this includes both men and women. This seems to suggest that White men consider "male" to be an important aspect of their identity, while Black men (and White women) do not consider gender or sex to be as salient to their identities.

Discussion. Despite this categorization, it is clear that some features do not fall neatly into exactly one category, and that these categories do not cover all stereotypes. However, as described in Section 3, we sought to surface stereotypes with our prompts about (1) identity, personality, and/or intelligence, and (2) work, home-life, family, and/or leisure activities. As shown above, we were indeed able to identify stereotypes related to identity, home-life, and leisure activities. We have not surfaced stereotypes related to personality and intelligence, potentially because such stereotypes are more explicitly racist.

7.2 Analysis of Language at Word Category Level

Following the analysis of individual words, we look at how word *categories* occur across the study's "real" identities and "portrayed" identities. We conduct our analysis based on the semantic classes in the LIWC lexicon. Table 8 shows the top semantic classes of the responses in each identity group described above. The top semantic classes for each identity group are the most predictive TF-IDF features in the classification of racial identities given a portrayed identity (the experiment described as part of RQ 2).

Our analysis reveals consistent differences in word category use between participant groups. Individuals writing as themselves tend to focus on content that is more concrete, and when writing as an alternate identity, focus on content that is more abstract. For example, Black men writing as White men tend to focus on words that fall into more abstract categories, like "insight", and "social", while White men writing as themselves focus on more concrete things, like "leisure", "sleep", and "home". Black women writing as White women also center on words related to "social" and "insight", as well as "humans", while White women writing as themselves focus on topics like "leisure", "tv", "sleep", and "job". This seems to suggest that Black individuals might imagine White individuals to be more reflective, or abstractive, instead of focused on everyday issues. Characterizing out-group individuals more abstractly aligns with known effects of linguistic intergroup bias, as well as the fact that language abstraction promotes generalization of a single member's behavior to the whole group [3], also known as stereotyping.

For more quantitative results, we calculate the Spearman's correlation between LIWC categories for each identity group. The values are calculated using the same weights for each LIWC class per identity group as described earlier in this section. These correlations are shown in Table 9. Our highest significant correlations are between White men as themselves and White men as Black men (0.5623), Black men as White women and Black men as White men (0.4839), and White men and Black men as White men (0.4438). The first correlation merely shows that White men write similarly as both identities, and the second appears to also suggest that people of the same race use somewhat similar words when describing the other race. The third correlation here might imply that Black men participants are more accurate writing from the perspective of White men than the other way around. Note that this is true for women as well; Black women participants write more accurately as White women than vice versa (the correlation between Black women as White women and White women is 0.4246 versus a correlation of 0.1996 for the other way around). This parallels our earlier results in the prediction tasks; it is easier to differentiate real and portrayed Black writing than it is to differentiate real and portrayed White writing. This seems to imply Black participants are better able to anticipate the writing style of White participants in general. However, it is important to note that all the correlation values are relatively low; our highest correlation is just above 0.5. This suggests that in general, all identities, including portrayed identities, generally used different language.

8 LIMITATIONS

The results of this work should be considered in light of several limitations. First, it is important to note that we use a relatively small dataset, with several hundred data samples. Similarly, we only focus on two racial and two gender identities. In future work, we might increase the demographic dimensions to include multiple racial and gender identities. In addition, the participants in each identity group come from a variety of backgrounds, with different ages, education, and locations. Since we are unable to control for these, they might have affected some of our results. Another potential limitation arises from deception; although we removed highly deceptive features from our data, these features were based on models trained on a dataset primarily focused on age and gender deception. Since we asked participants to keep age and gender constant when writing as a portrayed identity, it is possible removing these features did not necessarily fully account for deception in our results.

We also emphasize that much of our analysis is generalized, and the statements we make do not pertain to all of our participants, much less all members of a particular racial identity. For example, when we say "basketball" is a predictive feature of White men portraying a Black identity and is therefore a stereotype, or that Black individuals potentially imagine White individuals to be more reflective rather than focused on day-to-day issues based on their word usage patterns, it goes without saying that this is not necessarily true of all of our participants, rather it is simply an analysis based on all the responses together. In other words, the results we

Table 8: Most predictive LIWC categories for each identity, in order of highest to lowest weight

Real Black		White as Black		Real White		Black as White	
Woman	Man	Woman	Man	Woman	Man	Woman	Man
TIME	I	SOCIAL	PAST	TENTAT	TIME	PREPS	ARTICLE
I	INCL	OTHREF	EXCL	LEISURE	LEISURE	ARTICLE	SOCIAL
PHYSCAL	SELF	PAST	PRESENT	SEE	HOME	SOCIAL	INSIGHT
AFFECT	MOTION	PRESENT	SPACE	INCL	SIMILES	INSIGHT	COGMECH
SPORTS	NUMBER	PRONOUN	OTHREF	TIME	TV	HUMANS	NEGATE
EATING	TIME	WE	LEISURE	TV	SLEEP	I	INCL
POSEMO	ARTICLE	SPACE	WE	JOB	POSFEEL	PRONOUN	BODY
ARTICLE	CAUSE	EXCL	DISCREP	SLEEP	DISCREP	MOTION	CAUSE
NUMBER	ACHIEVE	HOME	COGMECH	OCCUP	EXCL	PRESENT	JOB
ACHIEVE	CERTAIN	CAUSE	BODY	PAST	AFFECT	PHYSCAL	MOTION

Table 9: Spearman's correlation between LIWC classes among identities. * indicates significance at $p < 0.05$, ** indicates significance at $p < 0.001$. BW stands for "Black Women", "WMaBM" stands for "White Men as Black Men", etc.

	BW	WW	BM	WM	BWa WW	WWaBW	BMaWM	WMaBM
BW	1	0.2919*	0.1053	0.2546*	0.2219	0.1996	0.2417*	0.0410
WW		1	0.0392	0.2742*	0.4246**	0.1884	0.0900	0.1242
BM			1	0.2604*	0.3176**	0.1242	0.2484*	0.0714
WM				1	0.4438**	0.3727*	0.3767*	0.5623**
BWaWW					1	0.4246**	0.4839**	0.3777*
WWaBW						1	0.1278	0.2142
BMaWM							1	0.3454*
WMaBM								1

show here represent a single snapshot of stereotypes that may be different from those that emerge in everyday language. In addition, note that the collected dataset is not intended to be used to train NLP models that can generate harmful stereotypes.

9 CONCLUSION

In this paper, we explored how we can computationally identify racial stereotypes in text. We designed an experiment to elicit implicit stereotypes, creating a dataset consisting of responses from individuals of four racial and gender identities, and corresponding portrayed identities. We described a methodology for surfacing stereotypes, and an analysis of the linguistic features associated with them, through classification tasks and language usage patterns.

Using these methods, we are able to answer three research questions. Our first research question asked whether we can differentiate between authors writing from different identity perspectives. Through our first experiment, we found that we can indeed do this significantly better than a majority class baseline, and it is easier to distinguish between a White author's identity perspectives than a Black author's, showing that our White participants might view Black individuals as more significantly different from themselves. Our second question looked at whether we can predict an author's true identity given written text. We found that we can achieve nearly 90% accuracy for certain identities, and that generally it is easier to predict an author's true identity when the writing is portrayed as Black than when portrayed as White, suggesting more generalizations or implicit stereotypes for Black individuals. Our third research question focused on what we can learn about racial stereotypes from language differences between different groups of people. Through analysis of predictive words and usage patterns, we found that the features most related to an author portraying a different racial identity contain stereotypes, and they reveal how people of different identities see themselves and others, in regards to appearance, interests, home-life, and identity. In particular, we surfaced stereotypes about everyday activities like sports and leisure activities, about places of residence and work, and about the stereotypes people assume other identity groups have about their own racial identity. In addition, we find that Black participants write more accurately as a White identity than vice versa, drawing parallels with our results from the first two research questions and suggesting White participants are more likely to use inaccurate stereotypes than Black participants.

This work is merely a starting point in examining methods of surfacing racial stereotypes and biases. However, this paper shows that the methodology of identity portrayal can be used to extract meaningful stereotypes. From our findings, we also learned about the stereotypes Americans often express about each other. Although our sample may not be fully representative of the United States population, we are able to confirm known stereotypes, and more significantly, we can *detect* them by having individuals place themselves in the shoes of someone of a different race. By identifying possible linguistic stereotypes, our work can be used to identify stereotyping behavior in other online discussions, and potentially to train people in decision-making positions about the possibility of stereotypes in their writing.

The dataset used in this paper can be downloaded from http://lit.eecs.umich.edu/downloads.html

ACKNOWLEDGMENTS

We are grateful to Nick Camp from the University of Michigan for his input and pointers to relevant research work. This material is based in part upon work supported by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

REFERENCES

[1] Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. 2020. Automated Detection of Racial Microaggressions using Machine Learning. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2477–2484. https://doi.org/10.1109/SSCI47803.2020.9308569

[2] Ketra L. Armstrong. 1999. Nike's Communication with Black Audiences: A Sociological Analysis of Advertising Effectiveness via Symbolic Interactionism. *Journal of Sport and Social Issues* 23, 3 (Aug. 1999), 266–286. https://doi.org/10.1177/0193723599233003 Publisher: SAGE Publications Inc.

[3] Yvette Assilaméhou, Nadia Lepastourel, and Benoit Testé. 2013. How the Linguistic Intergroup Bias Affects Group Perception: Effects of Language Abstraction on Generalization to the Group. *The Journal of Social Psychology* 153, 1 (Jan. 2013), 98–108. https://doi.org/10.1080/00224545.2012.711380

[4] Zinzi D. Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T. Bassett. 2017. Structural racism and health inequities in the USA: evidence and interventions. *The Lancet* 389, 10077 (2017), 1453–1463. https://doi.org/10.1016/S0140-6736(17)30569-X

[5] Hilary B. Bergsieker, J. Nicole Shelton, and Jennifer A. Richeson. 2010. To be liked versus respected: Divergent goals in interracial interactions. *Journal of Personality and Social Psychology* 99, 2 (Aug. 2010), 248–264. https://doi.org/10.1037/a0018474

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

[7] James J. Bradac and Randall Wisegarver. 1984. Ascribed Status, Lexical Diversity, and Accent: Determinants of Perceived Status, Solidarity, and Control of Speech Style. *Journal of Language and Social Psychology* 3, 4 (Dec. 1984), 239–255. https://doi.org/10.1177/0261927X8400300401 Publisher: SAGE Publications Inc.

[8] Jennifer E. Bruening. 2005. Gender and Racial Analysis in Sport: Are All the Women White and All the Blacks Men? *Quest* 57, 3 (Aug. 2005), 330–349. https://doi.org/10.1080/00336297.2005.10491861

[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (14 April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[10] Miguel Ángel Cano, Seth J. Schwartz, David P. MacKinnon, Brian T. H. Keum, Guillermo Prado, Flavio F. Marsiglia, Christopher P. Salas-Wright, Cory L. Cobb, Luz M. Garcini, Mario De La Rosa, Mariana Sánchez, Abir Rahman, Laura M. Acosta, Angelica M. Roncancio, and Marcel A. de Dios. 2021. Exposure to ethnic discrimination in social media and symptoms of anxiety and depression among Hispanic emerging adults: Examining the moderating role of gender. *Journal of Clinical Psychology* 77, 3 (2021), 571–586. https://doi.org/10.1002/jclp.23050 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jclp.23050.

[11] Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L. Kern, Anneke E. K. Buffone, Lyle Ungar, and Martin E. P. Seligman. 2017. Real Men Don't Say "Cute": Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychological and Personality Science* 8, 3 (April 2017), 310–322. https://doi.org/10.1177/1948550616671998

[12] Cindy Chung and James Pennebaker. 2007. The Psychological Functions of Function Words. In *Social communication*. Psychology Press, New York, NY, US, 343–359.

[13] D Anthony Clark, Lisa Spanierman, Tamilia Reed, Jason Soble, and Sharon Cabana. 2011. Documenting Web log Expressions of Racial Microaggressions That Target American Indians. *Journal of Diversity in Higher Education* 4 (March 2011), 39–50. https://doi.org/10.1037/a0021762

[14] Kimberle Crenshaw. [n.d.]. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. ([n. d.]), 31.

[15] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–11. https://doi.org/10.1145/3313831.3376488

[16] Jessie Daniels. 2013. Race and racism in Internet Studies: A review and critique. *New Media & Society* 15, 5 (Aug. 2013), 695–719. https://doi.org/10.1177/1461444812462849

[17] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[19] Ashley Fetters. 2020. *The Many Faces of the 'Wine Mom'*. https://www.theatlantic.com/family/archive/2020/05/wine-moms-explained/612001/

[20] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. *arXiv:2106.11410 [cs]* (July 2021). http://arxiv.org/abs/2106.11410 arXiv: 2106.11410.

[21] Gilbert C. Gee and Chandra L. Ford. 2011. STRUCTURAL RACISM AND HEALTH INEQUITIES: Old Issues, New Directions. *Du Bois Review: Social Science Research on Race* 8, 1 (2011), 115–132. https://doi.org/10.1017/S1742058X11000130

[22] Joelle Sano Gilmore and Amy Jordan. 2012. Burgers and basketball: Race and stereotypes in food and beverage advertising aimed at children in the US. *Journal of Children and Media* 6, 3 (Aug. 2012), 317–332. https://doi.org/10.1080/17482798.2012.673498

[23] Samuel R Gross, Maurice Possley, and Klara Stephens. [n.d.]. Race and Wrongful Convictions in the United States. ([n. d.]).

[24] McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, Brenda Curtis, et al. 2021. Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research* 23, 5 (2021), e26933.

[25] Juliana Menasce Horowitz, Anna Brown, and Kiana Cox. 2019. The role of race and ethnicity in Americans' lives. https://www.pewresearch.org/social-trends/2019/04/09/the-role-of-race-and-ethnicity-in-americans-personal-lives/

[26] Brandon A. Jackson and Mary Margaret Hui. 2017. Looking for Brothers: Black Male Bonding at a Predominantly White Institution. *Journal of Negro Education* 86, 4 (2017), 463–478. https://muse.jhu.edu/article/802718 Publisher: Journal of Negro Education.

[27] May Jiang and Christiane Fellbaum. 2020. Interdependencies of Gender and Race in Contextualized Word Embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain (Online), 17–25. https://aclanthology.org/2020.gebnlp-1.2

[28] Kenneth Joseph and Jonathan Morgan. 2020. When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4392–4415. https://doi.org/10.18653/v1/2020.acl-main.405

[29] Anastacia Kurylo. 2012. What Are *They* Like? Non-Expert Definitions of Stereotypes and Their Implications for Stereotype Maintenance. *Qualitative Research in Psychology* 9, 4 (Oct. 2012), 337–350. https://doi.org/10.1080/14780887.2010.500517

[30] Ryan Lavalley and Khalilah Robinson Johnson. 2020. Occupation, injustice, and anti-Black racism in the United States of America. *Journal of Occupational Science* (Sept. 2020), 1–13. https://doi.org/10.1080/14427591.2020.1810111

[31] Michael Lepori. 2020. Unequal Representations: Analyzing Intersectional Biases in Word Embeddings Using Representational Similarity Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1720–1728. https://doi.org/10.18653/v1/2020.coling-main.151

[32] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, Intimacy and Self-Disclosure in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 3857–3869. https://doi.org/10.1145/2858036.2858414

[33] C. Neil Macrae, Charles Stangor, Miles Hewstone, and Leslie A Zebrowitz. 1996. *Physical appearance as a basis of stereotyping*. Guilford Press, 79–120.

[34] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias

in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 615–621. https://doi.org/10.18653/v1/N19-1062

[35] Henrika McCoy. 2020. Black Lives Matter, and Yes, You are Racist: The Parallelism of the Twentieth and Twenty-First Centuries. *Child and Adolescent Social Work Journal* 37, 5 (Oct. 2020), 463–475. https://doi.org/10.1007/s10560-020-00690-4

[36] Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6355–6361. https://doi.org/10.18653/v1/D19-1666

[37] James Moody. 2001. Race, School Integration, and Friendship Segregation in America. *Amer. J. Sociology* 107, 3 (Nov. 2001), 679–716. https://doi.org/10.1086/338954 Publisher: The University of Chicago Press.

[38] Samuel D. Museus and Kimberly A. Truong. 2013. Racism and Sexism in Cyberspace: Engaging Stereotypes of Asian American Women and Men to Facilitate Student Learning and Development. *About Campus: Enriching the Student Learning Experience* 18, 4 (Sept. 2013), 14–21. https://doi.org/10.1002/abc.21126

[39] Anna P. Goddu, Katie J. O'Conor, Sophie Lanzkron, Mustapha O. Saheed, Somnath Saha, Monica E. Peek, Carlton Haywood, and Mary Catherine Beach. 2018. Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *Journal of General Internal Medicine* 33, 5 (May 2018), 685–691. https://doi.org/10.1007/s11606-017-4289-2

[40] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 446–457. https://doi.org/10.1145/3351095.3372843

[41] J. Pennebaker, Cindy K. Chung, Molly Ireland, A. Gonzales, and R. Booth. 2011. The Development and Psychometric Properties of LIWC2007.

[42] Rosentene B. Purnell. 1982. Teaching Them to Curse: Racial Bias in Language, Pedagogy and Practices. *Phylon (1960-)* 43, 3 (1982), 231–241. https://doi.org/10.2307/274820 Publisher: Clark Atlanta University.

[43] Verónica Pérez-Rosas, Quincy Davenport, Anna Mengdan Dai, Mohamed Abouelenien, and Rada Mihalcea. 2017. Identity Deception Detection. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 885–894. https://www.aclweb.org/anthology/I17-1089

[44] Douglas Rice, Jesse H Rhodes, and Tatishe Nteta. 2019. Racial bias in legal language. *Research & Politics* 6, 2 (2019), 2053168019848930.

[45] Douglas Rice, Jesse H. Rhodes, and Tatishe Nteta. 2019. Racial bias in legal language. *Research & Politics* 6, 2 (April 2019), 2053168019848930. https://doi.org/10.1177/2053168019848930 Publisher: SAGE Publications Ltd.

[46] Cynthia L. Robinson. 2011. Hair as Race: Why "Good Hair" May Be Bad for Black Females. *Howard Journal of Communications* 22, 4 (Oct. 2011), 358–376. https://doi.org/10.1080/10646175.2011.617212

[47] Zick Rubin. 1975. Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology* 11, 3 (1975), 233–260. https://doi.org/10.1016/S0022-1031(75)80025-4

[48] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[49] Laura West Steck, Druann Maria Heckert, and D. Alex Heckert. 2003. The salience of racial identity among African-American and white students. *Race and Society* 6, 1 (Jan. 2003), 57–73. https://doi.org/10.1016/j.racsoc.2004.09.005

[50] Amelia Tait. 2018. *Karen, Sharon, Becky, and Chad: How it feels when your name becomes a meme.* https://www.newstatesman.com/science-tech/2018/01/karen-sharon-becky-and-chad-how-it-feels-when-your-name-becomes-meme

[51] Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114, 25 (June 2017), 6521–6526. https://doi.org/10.1073/pnas.1702413114

[52] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

[53] Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of Internet memes. *Computers in Human Behavior* 63 (Oct. 2016), 424–432. https://doi.org/10.1016/j.chb.2016.05.067

[54] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4134–4145. https://doi.org/10.18653/v1/2020.acl-main.380