A Survey of Race, Racism, and Anti-Racism in NLP

Anjalie Field
Carnegie Mellon University
anjalief@cs.cmu.edu

Su Lin Blodgett
Microsoft Research
sulin.blodgett@microsoft.com

Zeerak Waseem
University of Sheffield
z.w.butt@sheffield.ac.uk

Yulia Tsvetkov
University of Washington
yuliats@cs.washington.edu

Abstract

Despite inextricable ties between race and language, little work has considered race in NLP research and development. In this work, we survey 79 papers from the ACL anthology that mention race. These papers reveal various types of race-related bias in all stages of NLP model development, highlighting the need for proactive consideration of how NLP systems can uphold racial hierarchies. However, persistent gaps in research on race and NLP remain: race has been siloed as a niche topic and remains ignored in many NLP tasks; most work operationalizes race as a fixed single-dimensional variable with a ground-truth label, which risks reinforcing differences produced by historical racism; and the voices of historically marginalized people are nearly absent in NLP literature. By identifying where and how NLP literature has and has not considered race, especially in comparison to related fields, our work calls for inclusion and racial justice in NLP research practices.

1 Introduction

Race and language are tied in complicated ways. Raciolinguistics scholars have studied how they are mutually constructed: historically, colonial powers construct linguistic and racial hierarchies to justify violence, and currently, beliefs about the inferiority of racialized people's language practices continue to justify social and economic exclusion (Rosa and Flores, 2017).[1] Furthermore, language is the primary means through which stereotypes and prejudices are communicated and perpetuated (Hamilton and Trolier, 1986; Bar-Tal et al., 2013).

However, questions of race and racial bias have been minimally explored in NLP literature. While researchers and activists have increasingly drawn attention to racism in computer science and academia, frequently-cited examples of racial bias in AI are often drawn from disciplines other than NLP, such as computer vision (facial recognition) (Buolamwini and Gebru, 2018) or machine learning (recidivism risk prediction) (Angwin et al., 2016). Even the presence of racial biases in search engines like Google (Sweeney, 2013; Noble, 2018) has prompted little investigation in the ACL community. Work on NLP and race remains sparse, particularly in contrast to concerns about gender bias, which have led to surveys, workshops, and shared tasks (Sun et al., 2019; Webster et al., 2019).

In this work, we conduct a comprehensive survey of how NLP literature and research practices engage with race. We first examine 79 papers from the ACL Anthology that mention the words 'race', 'racial', or 'racism' and highlight examples of how racial biases manifest at all stages of NLP model pipelines (§3). We then describe some of the limitations of current work (§4), specifically showing that NLP research has only examined race in a narrow range of tasks with limited or no social context. Finally, in §5, we revisit the NLP pipeline with a focus on how *people* generate data, build models, and are affected by deployed systems, and we highlight current failures to engage with people traditionally underrepresented in STEM and academia.

While little work has examined the role of race in NLP specifically, prior work has discussed race in related fields, including human-computer interaction (HCI) (Ogbonnaya-Ogburu et al., 2020; Rankin and Thomas, 2019; Schlesinger et al., 2017), fairness in machine learning (Hanna et al., 2020), and linguistics (Hudley et al., 2020; Motha, 2020). We draw comparisons and guidance from this work and show its relevance to NLP research. Our work differs from NLP-focused related work on gender bias (Sun et al., 2019), 'bias' generally

[1] We use *racialization* to refer the process of "ascribing and prescribing a racial category or classification to an individual or group of people . . . based on racial attributes including but not limited to cultural and social history, physical features, and skin color" (Hudley, 2017).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1905–1925
August 1–6, 2021.

(Blodgett et al., 2020), and the adverse impacts of language models (Bender et al., 2021) in its explicit focus on race and racism.

In surveying research in NLP and related fields, we ultimately find that *NLP systems and research practices produce differences along racialized lines*. Our work calls for NLP researchers to consider the social hierarchies upheld and exacerbated by NLP research and to shift the field toward "greater inclusion and racial justice" (Hudley et al., 2020).

2 What is race?

It has been widely accepted by social scientists that race is a social construct, meaning it "was brought into existence or shaped by historical events, social forces, political power, and/or colonial conquest" rather than reflecting biological or 'natural' differences (Hanna et al., 2020). More recent work has criticized the "social construction" theory as circular and rooted in academic discourse, and instead referred to race as "colonial constituted practices", including "an inherited western, modern-colonial practice of violence, assemblage, superordination, exploitation and segregation" (Saucier et al., 2016).

The term race is also multi-dimensional and can refer to a variety of different perspectives, including *racial identity* (how you self-identify), *observed race* (the race others perceive you to be), and *reflected race* (the race you believe others perceive you to be) (Roth, 2016; Hanna et al., 2020; Ogbonnaya-Ogburu et al., 2020). Racial categorizations often differ across dimensions and depend on the defined categorization schema. For example, the United States census considers Hispanic an ethnicity, not a race, but surveys suggest that 2/3 of people who identify as Hispanic consider it a part of their racial background.[2] Similarly, the census does not consider 'Jewish' a race, but some NLP work considers anti-Semitism a form of racism (Hasanuzzaman et al., 2017). Race depends on historical and social context—there are no 'ground truth' labels or categories (Roth, 2016).

As the work we survey primarily focuses on the United States, our analysis similarly focuses on the U.S. However, as race and racism are global constructs, some aspects of our analysis are applicable to other contexts. We suggest that future studies on racialization in NLP ground their analysis in the appropriate geo-cultural context, which may result in findings or analyses that differ from our work.

3 Survey of NLP literature on race

3.1 ACL Anthology papers about race

In this section, we introduce our primary survey data—papers from the ACL Anthology[3]—and we describe some of their major findings to emphasize that *NLP systems encode racial biases*. We searched the anthology for papers containing the terms 'racial', 'racism', or 'race', discarding ones that only mentioned race in the references section or in data examples and adding related papers cited by the initial set if they were also in the ACL Anthology. In using keyword searches, we focus on papers that explicitly mention race and consider papers that use euphemistic terms to not have substantial engagement on this topic. As our focus is on NLP and the ACL community, we do not include NLP-related papers published in other venues in the reported metrics (e.g. Table 1), but we do draw from them throughout our analysis.

Our initial search identified 165 papers. However, reviewing all of them revealed that many do not deeply engage on the topic. For example, 37 papers mention 'racism' as a form of abusive language or use 'racist' as an offensive/hate speech label without further engagement. 30 papers only mention race as future work, related work, or motivation, e.g. in a survey about gender bias, "Non-binary genders as well as racial biases have largely been ignored in NLP" (Sun et al., 2019). After discarding these types of papers, our final analysis set consists of 79 papers.[4]

Table 1 provides an overview of the 79 papers, manually coded for each paper's primary NLP task and its focal goal or contribution. We determined task/application labels through an iterative process: listing the main focus of each paper and then collapsing similar categories. In cases where papers could rightfully be included in multiple categories, we assign them to the best-matching one based on stated contributions and the percentage of the paper devoted to each possible category. In the Appendix we provide additional categorizations of the papers

[2] https://bit.ly/3r9J1fO, https://pewrsr.ch/36vlUEl

[3] The ACL Anthology includes papers from all official ACL venues and some non-ACL events listed in Appendix A, as of December 2020 it included 6,200 papers

[4] We do not discard all papers about abusive language, only ones that exclusively use racism/racist as a classification label. We retain papers with further engagement, e.g. discussions of how to define racism or identification of racial bias in hate speech classifiers.

	Collect Corpus	Analyze Corpus	Develop Model	Detect Bias	Debias	Survey/Position	Total
Abusive Language	6	4	2	5	2	2	21
Social Science/Social Media	2	10	6	1	-	1	20
Text Representations (LMs, embeddings)	-	2	-	9	2	-	13
Text Generation (dialogue, image captions, story gen.)	-	-	1	5	1	1	8
Sector-specific NLP applications (edu., law, health)	1	2	-	-	1	3	7
Ethics/Task-independent Bias	1	-	1	1	1	2	6
Core NLP Applications (parsing, NLI, IE)	1	-	1	1	1	-	4
Total	11	18	11	22	8	9	79

Table 1: 79 papers on race or racism from the ACL anthology, categorized by NLP application and focal task.

according to publication year, venue, and racial categories used, as well as the full list of 79 papers.

3.2 NLP systems encode racial bias

Next, we present examples that identify racial bias in NLP models, focusing on 5 parts of a standard NLP pipeline: data, data labels, models, model outputs, and social analyses of outputs. We include papers described in Table 1 and also relevant literature beyond the ACL Anthology (e.g. NeurIPS, PNAS, Science). These examples are not intended to be exhaustive, and in §4 we describe some of the ways that NLP literature has failed to engage with race, but nevertheless, we present them as evidence that *NLP systems perpetuate harmful biases along racialized lines*.

Data A substantial amount of prior work has already shown how NLP systems, especially word embeddings and language models, can absorb and amplify social biases in data sets (Bolukbasi et al., 2016; Zhao et al., 2017). While most work focuses on gender bias, some work has made similar observations about racial bias (Rudinger et al., 2017; Garg et al., 2018; Kurita et al., 2019). These studies focus on *how* training data might describe racial minorities in biased ways, for example, by examining words associated with terms like 'black' or traditionally European/African American names (Caliskan et al., 2017; Manzini et al., 2019). Some studies additionally capture *who* is described, revealing under-representation in training data, sometimes tangentially to primary research questions: Rudinger et al. (2017) suggest that gender bias may be easier to identify than racial or ethnic bias in Natural Language Inference data sets because of data sparsity, and Caliskan et al. (2017) alter the Implicit Association Test stimuli that they use to measure biases in word embeddings because some African American names were not frequent enough in their corpora.

An equally important consideration, in addition to whom the data describes is *who authored the data*. For example, Blodgett et al. (2018) show that parsing systems trained on White Mainstream American English perform poorly on African American English (AAE).[5] In a more general example, Wikipedia has become a popular data source for many NLP tasks. However, surveys suggest that Wikipedia editors are primarily from white-majority countries,[6] and several initiatives have pointed out systemic racial biases in Wikipedia coverage (Adams et al., 2019; Field et al., 2021).[7] Models trained on these data only learn to process the type of text generated by these users, and further, only learn information about the topics these users are interested in. The *representativeness* of data sets is a well-discussed issue in social-oriented tasks, like inferring public opinion (Olteanu et al., 2019), but this issue is also an important consideration in 'neutral' tasks like parsing (Waseem et al., 2021). The type of data that researchers choose to train their models on does not just affect what *data* the models perform well for, it affects what *people* the models work for. NLP researchers cannot assume models will be useful or function for marginalized people unless they are trained on data

[5] We note that conceptualizations of AAE and the accompanying terminology for the variety have shifted considerably in the last half century; see King (2020) for an overview.

[6] https://bit.ly/2Yv07IL

[7] https://bit.ly/3j2weZA

generated by them.

Data Labels Although model biases are often blamed on raw data, several of the papers we survey identify biases in the way researchers categorize or obtain data annotations. For example:

- **Annotation schema** Returning to Blodgett et al. (2018), this work defines new parsing standards for formalisms common in AAE, demonstrating how parsing labels themselves were not designed for racialized language varieties.
- **Annotation instructions** Sap et al. (2019) show that annotators are less likely to label tweets using AAE as offensive if they are told the likely language varieties of the tweets. Thus, how annotation schemes are designed (e.g. what contextual information is provided) can impact annotators' decisions, and failing to provide sufficient context can result in racial biases.
- **Annotator selection** Waseem (2016) show that feminist/anti-racist activists assign different offensive language labels to tweets than figure-eight workers, demonstrating that annotators' lived experiences affect data annotations.

Models Some papers have found evidence that model instances or architectures can change the racial biases of outputs produced by the model. Sommerauer and Fokkens (2019) find that the word embedding associations around words like 'race' and 'racial' change not only depending on the model architecture used to train embeddings, but also on the specific model *instance* used to extract them, perhaps because of differing random seeds. Kiritchenko and Mohammad (2018) examine gender and race biases in 200 sentiment analysis systems submitted to a shared task and find different levels of bias in different systems. As the training data for the shared task was standardized, all models were trained on the same data. However, participants could have used external training data or pre-trained embeddings, so a more detailed investigation of results is needed to ascertain which factors most contribute to disparate performance.

Model Outputs Several papers focus on model outcomes, and how NLP systems could perpetuate and amplify bias if they are deployed:

- Classifiers trained on common abusive language data sets are more likely to label tweets containing characteristics of AAE as offensive (Davidson et al., 2019; Sap et al., 2019).
- Classifiers for abusive language are more likely to label text containing identity terms like 'black' as offensive (Dixon et al., 2018).
- GPT outputs text with more negative sentiment when prompted with AAE -like inputs (Groenwold et al., 2020).

Social Analyses of Outputs While the examples in this section primarily focus on racial biases in trained NLP systems, other work (e.g. included in 'Social Science/Social Media' in Table 1) uses NLP tools to analyze race in society. Examples include examining how commentators describe football players of different races (Merullo et al., 2019) or how words like 'prejudice' have changed meaning over time (Vylomova et al., 2019).

While differing in goals, this work is often susceptible to the same pitfalls as other NLP tasks. One area requiring particular caution is in the interpretation of results produced by analysis models. For example, while word embeddings have become a common way to measure semantic change or estimate word meanings (Garg et al., 2018), Joseph and Morgan (2020) show that embedding associations do not always correlate with human opinions; in particular, correlations are stronger for beliefs about gender than race. Relatedly, in HCI, the recognition that authors' own biases can affect their interpretations of results has caused some authors to provide self-disclosures (Schlesinger et al., 2017), but this practice is uncommon in NLP.

We conclude this section by observing that when researchers have looked for racial biases in NLP systems, they have usually found them. This literature calls for proactive approaches in considering how data is collected, annotated, used, and interpreted to prevent NLP systems from exacerbating historical racial hierarchies.

4 Limitations in where and how NLP operationalizes race

While §3 demonstrates ways that NLP systems encode racial biases, we next identify gaps and limitations in how these works have examined racism, focusing on *how* and *in what tasks* researchers have considered race. We ultimately conclude that prior NLP literature has marginalized research on race and encourage deeper engagement with other fields, critical views of simplified classification schema,

and broader application scope in future work (Blodgett et al., 2020; Hanna et al., 2020).

4.1 Common data sets are narrow in scope

The papers we surveyed suggest that research on race in NLP has used a very limited range of data sets, which fails to account for the multi-dimensionality of race and simplifications inherent in classification. We identified 3 common data sources:[8]

- 9 papers use a set of tweets with inferred probabilistic topic labels based on alignment with U.S. census race/ethnicity groups (or the provided inference model) (Blodgett et al., 2016).
- 11 papers use lists of names drawn from Sweeney (2013), Caliskan et al. (2017), or Garg et al. (2018). Most commonly, 6 papers use African/European American names from the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), which in turn draws data from Greenwald et al. (1998) and Bertrand and Mullainathan (2004).
- 10 papers use explicit keywords like 'Black woman', often placed in templates like "I am a _____" to test if model performance remains the same for different identity terms.

While these commonly-used data sets can identify performance disparities, they only capture a narrow subset of the multiple dimensions of race (§2). For example, none of them capture self-identified race. While observed race is often appropriate for examining discrimination and some types of disparities, it is impossible to assess potential harms and benefits of NLP systems without assessing their performance over text generated by and directed to people of different races. The corpus from Blodgett et al. (2016) does serve as a starting point and forms the basis of most current work assessing performance gaps in NLP models (Sap et al., 2019; Blodgett et al., 2018; Xia et al., 2020; Xu et al., 2019; Groenwold et al., 2020), but even this corpus is explicitly not intended to infer race.

Furthermore, names and hand-selected identity terms are not sufficient for uncovering model bias. De-Arteaga et al. (2019) show this in examining gender bias in occupation classification: when overt indicators like names and pronouns are scrubbed from the data, performance gaps and potential allocational harms still remain. Names also generalize poorly. While identity terms can be examined across languages (van Miltenburg et al., 2017), differences in naming conventions often do not translate, leading some studies to omit examining racial bias in non-English languages (Lauscher and Glavaš, 2019). Even within English, names often fail to generalize across domains, geographies, and time. For example, names drawn from the U.S. census generalize poorly to Twitter (Wood-Doughty et al., 2018), and names common among Black and white children were not distinctly different prior to the 1970s (Fryer Jr and Levitt, 2004; Sweeney, 2013).

We focus on these 3 data sets as they were most common in the papers we surveyed, but we note that others exist. Preoţiuc-Pietro and Ungar (2018) provide a data set of tweets with self-identified race of their authors, though it is little used in subsequent work and focused on demographic prediction, rather than evaluating model performance gaps. Two recently-released data sets (Nadeem et al., 2020; Nangia et al., 2020) provide crowd-sourced pairs of more- and less-stereotypical text. More work is needed to understand any privacy concerns and the strengths and limitations of these data (Blodgett et al., 2021). Additionally, some papers collect domain-specific data, such as self-reported race in an online community (Loveys et al., 2018), or crowd-sourced annotations of perceived race of football players (Merullo et al., 2019). While these works offer clear contextualization, it is difficult to use these data sets to address other research questions.

4.2 Classification schemes operationalize race as a fixed, single-dimensional U.S.-census label

Work that uses the same few data sets inevitably also uses the same few classification schemes, often without justification. The most common explicitly stated source of racial categories is the U.S. census, which reflects the general trend of U.S.-centrism in NLP research (the vast majority of work we surveyed also focused on English). While census categories are sometimes appropriate, repeated use of classification schemes and accompanying data sets without considering who defined these schemes and whether or not they are appropriate for the current context risks perpetuating the misconception that race is 'natural' across geo-cultural contexts. We refer to Hanna et al. (2020) for a more thorough

[8] We provide further counts of what racial categories papers use and how they operationalize them in Appendix B.

overview of the harms of "widespread uncritical adoption of racial categories," which "can in turn re-entrench systems of racial stratification which give rise to real health and social inequalities." At best, the way race has been operationalized in NLP research is only capable of examining a narrow subset of potential harms. At worst, it risks reinforcing racism by presenting racial divisions as natural, rather than the product of social and historical context (Bowker and Star, 2000).

As an example of questioning who devised racial categories and for what purpose, we consider the pattern of re-using names from Greenwald et al. (1998), who describe their data as sets of names "judged by introductory psychology students to be more likely to belong to White Americans than to Black Americans" or vice versa. When incorporating this data into WEAT, Caliskan et al. (2017) discard some judged African American names as too infrequent in their embedding data. Work subsequently drawing from WEAT makes no mention of the discarded names nor contains much discussion of how the data was generated and whether or not names judged to be white or Black by introductory psychology students in 1998 are an appropriate benchmark for the studied task. While gathering data to examine race in NLP is challenging, and in this work we ourselves draw from examples that use Greenwald et al. (1998), it is difficult to interpret what implications arise when models exhibit disparities over this data and to what extent models without disparities can be considered 'debiased'.

Finally, almost all of the work we examined conducts single-dimensional analyses, e.g. focus on race or gender but not both simultaneously. This focus contrasts with the concept of *intersectionality*, which has shown that examining discrimination along a single axis fails to capture the experiences of people who face marginalization along multiple axes. For example, consideration of race often emphasizes the experience of gender-privileged people (e.g. Black men), while consideration of gender emphasizes the experience of race-privileged people (e.g. white women). Neither reflect the experience of people who face discrimination along both axes (e.g. Black women) (Crenshaw, 1989). A small selection of papers have examined intersectional biases in embeddings or word co-occurrences (Herbelot et al., 2012; May et al., 2019; Tan and Celis, 2019; Lepori, 2020), but we did not identify mentions of intersectionality in any other NLP research areas. Further, several of these papers use NLP technology to examine or validate theories on intersectionality; they do not draw from theory on intersectionality to critically examine NLP models. These omissions can mask harms: Jiang and Fellbaum (2020) provide an example using word embeddings of how failing to consider intersectionality can render invisible people marginalized in multiple ways. Numerous directions remain for exploration, such as how 'debiasing' models along one social dimension affects other dimensions. Surveys in HCI offer further frameworks on how to incorporate identity and intersectionality into computational research (Schlesinger et al., 2017; Rankin and Thomas, 2019).

4.3 NLP research on race is restricted to specific tasks and applications

Finally, Table 1 reveals many common NLP applications where race has not been examined, such as machine translation, summarization, or question answering.[9] While some tasks seem inherently more relevant to social context than others (a claim we dispute in this work, particularly in §5), *research on race is compartmentalized to limited areas of NLP even in comparison with work on 'bias'*. For example, Blodgett et al. (2020) identify 20 papers that examine bias in co-reference resolution systems and 8 in machine translation, whereas we identify 0 papers in either that consider race. Instead, race is most often mentioned in NLP papers in the context of abusive language, and work on detecting or removing bias in NLP models has focused on word embeddings.

Overall, our survey identifies a need for the examination of race in a broader range of NLP tasks, the development of multi-dimensional data sets, and careful consideration of context and appropriateness of racial categories. In general, race is difficult to operationalize, but NLP researchers do not need to start from scratch, and can instead draw from relevant work in other fields.

5 NLP propagates marginalization of racialized people

While in §4 we primarily discuss race as a topic or a construct, in this section, we consider the role, or more pointedly, the absence, of traditionally underrepresented people in NLP research.

[9] We identified only 8 relevant papers on Text Generation, which focus on other areas including chat bots, GPT-2/3, humor generation, and story generation.

5.1 People create data

As discussed in §3.2, data and annotations are generated by people, and failure to consider who created data can lead to harms. In §3.2 we identify a need for diverse training data in order to ensure models work for a diverse set of people, and in §4 we describe a similar need for diversity in data that is used to assess algorithmic fairness. However, gathering this type of data without consideration of the people who generated it can introduce privacy violations and risks of demographic profiling.

As an example, in 2019, partially in response to research showing that facial recognition algorithms perform worse on darker-skinned than lighter-skinned people (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019), researchers at IBM created the "Diversity in Faces" data set, which consists of 1 million photos sampled from the the publicly available YFCC-100M data set and annotated with "craniofacial distances, areas and ratios, facial symmetry and contrast, skin color, age and gender predictions" (Merler et al., 2019). While this data set aimed to improve the fairness of facial recognition technology, it included photos collected from a Flickr, a photo-sharing website whose users did not explicitly consent for this use of their photos. Some of these users filed a lawsuit against IBM, in part for "subjecting them to increased surveillance, stalking, identity theft, and other invasions of privacy and fraud."[10] NLP researchers could easily repeat this incident, for example, by using demographic profiling of social media users to create more diverse data sets. While obtaining diverse, representative, real-world data sets is important for building models, data must be collected with consideration for the people who generated it, such as obtaining informed consent, setting limits of uses, and preserving privacy, as well as recognizing that some communities may not want their data used for NLP at all (Paullada, 2020).

5.2 People build models

Research is additionally carried out by people who determine what projects to pursue and how to approach them. While statistics on ACL conferences and publications have focused on geographic representation rather than race, they do highlight under-representation. Out of $2,695$ author affiliations associated with papers in the ACL Anthology for 5 major conferences held in 2018, only 5 (0.2%) were from Africa, compared with $1,114$ from North America (41.3%).[11] Statistics published for 2017 conference attendees and ACL fellows similarly reveal a much higher percentage of people from "North, Central and South America" (55% attendees / 74% fellows) than from "Europe, Middle East and Africa" (19%/13%) or "Asia-Pacific" (23%/13%).[12] These broad regional categories likely mask further under-representation, e.g. percentage of attendees and fellows from Africa as compared to Europe. According to an NSF report that includes racial statistics rather than nationality, 14% of doctorate degrees in Computer Science awarded by U.S. institutions to U.S. citizens and permanent residents were awarded to Asian students, $< 4\%$ to Black or African American students, and 0% to American Indian or Alaska Native students (National Center for Science and Engineering Statistics, 2019).[13]

It is difficult to envision reducing or eliminating racial differences in NLP systems without changes in the researchers building these systems. One theory that exemplifies this challenge is *interest convergence*, which suggests that people in positions of power only take action against systematic problems like racism when it also advances their own interests (Bell Jr, 1980). Ogbonnaya-Ogburu et al. (2020) identify instances of interest convergence in the HCI community, primarily in diversity initiatives that benefit institutions' images rather than underrepresented people. In a research setting, interest convergence can encourage studies of incremental and surface-level biases while discouraging research that might be perceived as controversial and force fundamental changes in the field.

Demographic statistics are not sufficient for avoiding pitfalls like interest convergence, as they fail to capture the lived experiences of researchers. Ogbonnaya-Ogburu et al. (2020) provide several examples of challenges that non-white HCI researchers have faced, including the invisible labor of representing 'diversity', everyday microaggres-

[10] https://bit.ly/3r3LuIk https://nbcnews.to/3j5hI39 IBM has since removed the "Diversity in Faces" data set as well as their "Detect Faces" public API and stopped their use of and research on facial recognition. https://bit.ly/3j2Jv4i

[11] http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/

[12] https://www.aclweb.org/portal/content/acl-diversity-statistics

[13] Results exclude respondents who did not report race or ethnicity or were Native Hawaiian or Other Pacific Islander.

sions, and altering their research directions in accordance with their advisors' interests. Rankin and Thomas (2019) further discuss how research conducted by people of different races is perceived differently: "Black women in academia who conduct research about the intersections of race, gender, class, and so on are perceived as 'doing service,' whereas white colleagues who conduct the same research are perceived as doing cutting-edge research that demands attention and recognition." While we draw examples about race from HCI in the absence of published work on these topics in NLP, the lack of linguistic diversity in NLP research similarly demonstrates how representation does not necessarily imply inclusion. Although researchers from various parts of the world (Asia, in particular) do have some numerical representation among ACL authors, attendees, and fellows, NLP research overwhelmingly favors a small set of languages, with a heavy skew towards European languages (Joshi et al., 2020) and 'standard' language varieties (Kumar et al., 2021).

5.3 People use models

Finally, NLP research produces technology that is used by people, and even work without direct applications is typically intended for incorporation into application-based systems. With the recognition that technology ultimately affects people, researchers on ethics in NLP have increasingly called for considerations of whom technology might harm and suggested that there are some NLP technologies that should not be built at all. In the context of perpetuating racism, examples include criticism of tools for predicting demographic information (Tatman, 2020) and automatic prison term prediction (Leins et al., 2020), motivated by the history of using technology to police racial minorities and related criticism in other fields (Browne, 2015; Buolamwini and Gebru, 2018; McIlwain, 2019). In cases where potential harms are less direct, they are often unaddressed entirely. For example, while low-resource NLP is a large area of research, a paper on machine translation of white American and European languages is unlikely to discuss how continual model improvements in these settings increase technological inequality. Little work on low-resource NLP has focused on the realities of structural racism or differences in lived experience and how they might affect the way technology should be designed.

Detection of abusive language offers an informative case study on the danger of failing to consider people affected by technology. Work on abusive language often aims to detect racism for content moderation (Waseem and Hovy, 2016). However, more recent work has show that existing hate speech classifiers are likely to falsely label text containing identity terms like 'black' or text containing linguistic markers of AAE as toxic (Dixon et al., 2018; Sap et al., 2019; Davidson et al., 2019; Xia et al., 2020). Deploying these models could censor the posts of the very people they purport to help.

In other areas of statistics and machine learning, focus on *participatory design* has sought to amplify the voices of people affected by technology and its development. An ICML 2020 workshop titled "Participatory Approaches to Machine Learning" highlights a number of papers in this area (Kulynych et al., 2020; Brown et al., 2019). A few related examples exist in NLP, e.g. Gupta et al. (2020) gather data for an interactive dialogue agent intended to provide more accessible information about heart failure to Hispanic/Latinx and African American patients. The authors engage with healthcare providers and doctors, though they leave focal groups with patients for future work. While NLP researchers may not be best situated to examine how people interact with deployed technology, they could instead draw motivation from fields that have stronger histories of participatory design, such as HCI. However, we did not identify citing participatory design studies conducted by others as common practice in the work we surveyed. As in the case of researcher demographics, participatory design is not an end-all solution. Sloane et al. (2020) provide a discussion of how participatory design can collapse to 'participation-washing' and how such work must be context-specific, long-term, and genuine.

6 Discussion

We conclude by synthesizing some of the observations made in the preceding sections into more actionable items. First, NLP research needs to explicitly incorporate race. We quote Benjamin (2019): *"[technical systems and social codes] operate within powerful systems of meaning that render some things visible, others invisible, and create a vast array of distortions and dangers."*

In the context of NLP research, this philosophy implies that all technology we build works in service of some ideas or relations, either by upholding

them or dismantling them. Any research that is not actively combating prevalent social systems like racism risks perpetuating or exacerbating them. Our work identifies several ways in which NLP research upholds racism:

- Systems contain representational harms and performance gaps throughout NLP pipelines
- Research on race is restricted to a narrow subset of tasks and definitions of race, which can mask harms and falsely reify race as 'natural'
- Traditionally underrepresented people are excluded from the research process, both as consumers and producers of technology

Furthermore, while we focus on race, which we note has received substantially less attention than gender, many of the observations in this work hold for social characteristics that have received even less attention in NLP research, such as socioeconomic class, disability, or sexual orientation (Mendelsohn et al., 2020; Hutchinson et al., 2020).

Nevertheless, none of these challenges can be addressed without direct engagement with marginalized communities of color. NLP researchers can draw on precedents for this type of engagement from other fields, such as participatory design and value sensitive design models (Friedman et al., 2013). Additionally, numerous organizations already exist that serve as starting points for partnerships, such as Black in AI, Masakhane, Data for Black Lives, and the Algorithmic Justice League.

Finally, race and language are complicated, and while readers may look for clearer recommendations, no one data set, model, or set of guidelines can 'solve' racism in NLP. For instance, while we draw from linguistics, Hudley et al. (2020) in turn call on linguists to draw models of racial justice from anthropology, sociology, and psychology. Relatedly, there are numerous racialized effects that NLP research can have that we do not address in this work; for example, Bender et al. (2021) and Strubell et al. (2019) discuss the environmental costs of training large language models, and how global warming disproportionately affects marginalized communities. We suggest that readers use our work as one starting point for bringing inclusion and racial justice into NLP.

Acknowledgements

We gratefully thank Hanna Kim, Kartik Goyal, Artidoro Pagnoni, Qinlan Shen, and Michael Miller Yoder for their feedback on this work. Z.W. has been supported in part by the Canada 150 Research Chair program and the UK-Canada Artificial Intelligence Initiative. A.F. has been supported in part by a Google PhD Fellowship and a GRFP under Grant No. DGE1745016. This material is based upon work supported in part by the National Science Foundation under Grants No. IIS2040926 and IIS2007960. 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 NSF.

7 Ethical Considerations

We, the authors of this work, are situated in the cultural contexts of the United States of America and the United Kingdom/Europe, and some of us identify as people of color. We all identify as NLP researchers, and we acknowledge that we are situated within the traditionally exclusionary practices of academic research. These perspectives have impacted our work, and there are viewpoints outside of our institutions and experiences that our work may not fully represent.

References

Julia Adams, Hannah Brückner, and Cambria Naslund. 2019. Who counts as a notable sociologist on Wikipedia? gender, race, and the "professor test". *Socius*, 5.

Silvio Amir, Mark Dredze, and John W. Ayers. 2019. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*.

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France. European Language Resources Association.

Daniel Bar-Tal, Carl F Graumann, Arie W Kruglanski, and Wolfgang Stroebe. 2013. *Stereotyping and prejudice: Changing conceptions*. Springer Science & Business Media.

Francesco Barbieri and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in Twitter. In *Proceedings of the Seventh*

Joint Conference on Lexical and Computational Semantics, pages 101–106, New Orleans, Louisiana. Association for Computational Linguistics.

Derrick A Bell Jr. 1980. Brown v. board of education and the interest-convergence dilemma. *Harvard law review*, pages 518–533.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, page 610–623, New York, NY, USA. Association for Computing Machinery.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1019, Atlanta, Georgia. Association for Computational Linguistics.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

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*, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, New York, NY, USA. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Michael Castelle. 2018. The linguistic ideologies of deep abusive language classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada. Association for Computational Linguistics.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(8).

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*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 120–128, New York, NY, USA. Association for Computing Machinery.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 67–73, New York, NY, USA. Association for Computing Machinery.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, Oregon, USA. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field, Chan Young Park, and Yulia Tsvetkov. 2021. Controlled analyses of social biases in Wikipedia bios. *Computing Research Repository*, arXiv:2101.00078. Version 1.

Batya Friedman, Peter Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael Gorman, editors, *Early engagement and new technologies: Opening up the laboratory*, volume 16. Springer, Dordrecht.

Roland G Fryer Jr and Steven D Levitt. 2004. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, 119(3):767–805.

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Nabeel Gillani and Roger Levy. 2019. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 94–99, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia. 2020. Heart failure education of African American and Hispanic/Latino patients: Data collection and analysis. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 41–46, Online. Association for Computational Linguistics.

David L Hamilton and Tina K Trolier. 1986. Stereotypes and stereotyping: An overview of the cogni-

tive approach. In J. F. Dovidiom and S. L. Gaertner, editors, *Prejudice, discrimination, and racism*, pages 127—163. Academic Press.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 501–512, New York, NY, USA. Association for Computing Machinery.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54, Avignon, France. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Anne H. Charity Hudley. 2017. Language and Racialization. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*, pages 381–402. Oxford University Press.

Anne H Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, 96(4):e200–e235.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

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*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

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*, pages 4392–4415, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Anna Kasunic and Geoff Kaufman. 2018. Learning to listen: Critically considering the role of AI in human storytelling and character creation. In *Proceedings of the First Workshop on Storytelling*, pages 1–13, New Orleans, Louisiana. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Sharese King. 2020. From African American Vernacular English to African American Language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics*, 6(1):285–300.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning. International Conference on Machine Learning Workshop.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

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*, pages 1720–1728, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.

Charlton D. McIlwain. 2019. *Black Software: The Internet and Racial Justice, from the AfroNet to Black Lives Matter*. Oxford University Press, Incorporated.

J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55.

Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *Computing Research Repository*, arXiv:1901.10436. Version 6.

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), pages 6355–6361, Hong Kong, China. Association for Computational Linguistics.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ehsan Mohammady and Aron Culotta. 2014. Using county demographics to infer attributes of Twitter users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 7–16, Baltimore, Maryland. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Leigh Yeh, Mohammad Atari, Brendan Kennedy, Gwenyth Portillo Wightman, Elaine Gonzalez, Natalie Delong, Rhea Bhatia, Arineh Mirinjian, Xiang Ren, and Morteza Dehghani. 2019. Reporting the unreported: Event extraction for analyzing the local representation of hate crimes. 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)*, pages 5753–5757, Hong Kong, China. Association for Computational Linguistics.

Suhanthie Motha. 2020. Is an antiracist and decolonizing applied linguistics possible? *Annual Review of Applied Linguistics*, 40:128–133.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *Computing Research Repository*, arXiv:2004.09456. Version 1.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

National Center for Science and Engineering Statistics. 2019. Doctorate recipients from U.S. universities. National Science Foundation.

Safiya U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Julia Parish-Morris. 2019. Computational linguistics for enhancing scientific reproducibility and reducing healthcare inequities. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 94–102, Minneapolis, Minnesota. Association for Computational Linguistics.

Amandalynne Paullada. 2020. How Does Machine Translation Shift Power? In *Proceedings of the First Workshop on Resistance AI*.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for NLP. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024, Austin, Texas. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435.

Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678, Vancouver, Canada. Association for Computational Linguistics.

Yolanda A. Rankin and Jakita O. Thomas. 2019. Straighten up and fly right: Rethinking intersectionality in HCI research. *Interactions*, 26(6):64–68.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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)*, pages 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647.

Wendy D Roth. 2016. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Wesley Santos and Ivandré Paraboni. 2019. Moral stance recognition and polarity classification from Twitter and elicited text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.

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*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

P.K. Saucier, T.P. Woods, P. Douglass, B. Hesse, T.K. Nopper, G. Thomas, and C. Wun. 2016. *Conceptual Aphasia in Black: Displacing Racial Formation*. Critical Africana Studies. Lexington Books.

Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional hci: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5412–5427, New York, NY, USA. Association for Computing Machinery.

Tyler Schnoebelen. 2017. Goal-oriented design for ethical machine learning and NLP. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 88–93, Valencia, Spain. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125, Online. Association for Computational Linguistics.

Sima Sharifirad and Stan Matwin. 2019. Using attention-based bidirectional LSTM to identify different categories of offensive language directed toward female celebrities. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 46–48, Florence, Italy. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

M Sloane, E Moss, O Awomolo, and L Forlano. 2020. Participation is not a design fix for machine learning. *Computing Research Repository*, arXiv:2007.02423. Version 3.

Harold Somers. 2006. Language engineering and the pathway to healthcare: A user-oriented view. In *Proceedings of the First International Workshop on Medical Speech Translation*, pages 28–35, New York, New York. Association for Computational Linguistics.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems*, volume 32, pages 13230–13241. Curran Associates, Inc.

Rachael Tatman. 2020. What I Won't Build. Workshop on Widening NLP.

Rocco Tripodi, Massimo Warglien, Simon Levis Sullam, and Deborah Paci. 2019. Tracing antisemitic language through diachronic embedding projections: France 1789-1914. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 115–125, Florence, Italy. Association for Computational Linguistics.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, and Isabelle Bingel, Joachim Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *Computing Research Repository*, arXiv:2101.11974. Version 1.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy. Association for Computational Linguistics.

Michael Wojatzki, Saif Mohammad, Torsten Zesch, and Svetlana Kiritchenko. 2018. Quantifying qualitative data for understanding controversial issues. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, and Mark Dredze. 2018. Predicting Twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 105–111, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Zach Wood-Doughty, Michael Smith, David Broniatowski, and Mark Dredze. 2017. How does Twitter user behavior vary across demographic groups? In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 83–89, Vancouver, Canada. Association for Computational Linguistics.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

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*, pages 4134–4145, Online. Association for Computational Linguistics.

Jieyu Zhao and Kai-Wei Chang. 2020. LOGAN: Local group bias detection by clustering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1968–1977, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

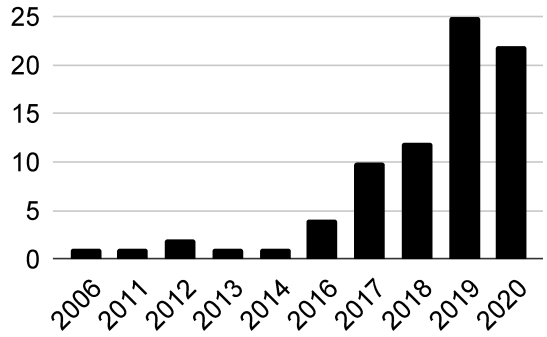

Figure 1: Year of publication of 79 papers that mention "racial" or "racism". More papers have been published in recent years (2019-2020).

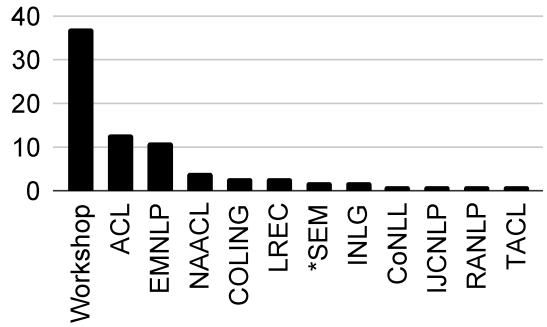

Figure 2: Venue of publication of 79 papers that mention "racial" or "racism". About half (46.8%) were published in workshops.

A ACL Anthology Venues

ACL events: AACL, ACL, ANLP, CL, CoNLL, EACL, EMNLP, Findings, NAACL, SemEval, *SEM, TACL, WMT, Workshops, Special Interest Groups

Non-ACL events: ALTA, AMTA, CCL, COLING, EAMT, HLT, IJCNLP, JEP/TALN/RECITAL, LILT, LREC, MUC, PACLIC, RANLP, ROCLING/IJCLCLP, TINLAP, TIPSTER

B Additional Survey Metrics

We show three additional breakdowns of the data set: Figure 1 shows the number of papers published each year, Figure 2 shows the number of papers published in each venue, and Table 2 shows how papers have operationalized race. As expected, given the growth of NLP research in general and the increasing focus on social issues (e.g. "Ethics and NLP" track was added to ACL in 2020) more work has been published on race in more recent years (2019, 2020). In Figure 2, we consider if work on race has been siloed into or out of specific

	Census-aligned	Crowd-sourced	Explicit keywords	External/Public	Names	Predicted	Self-reported	Total
4+			5	2		1	5	13
BW	7		2	1	8	1	1	20
BWAH	1					3		4
{BWAH}	1	1	3	1	2			8
W/non-W		1			1			2
Total	9	2	10	4	11	5	6	47

Table 2: Racial categories used by ACL Anthology papers. BWAH stand for Black, White, Asian, and Hispanic. {BWAH} denotes any incomplete subset of BWAH other than BW (e.g. Black and Hispanic). 4+ denotes that the paper used ≥ 4 racial categories, often including "other", "mixed", or an open-ended text box. Papers with multiple schema are counted as separate data points.

venues. The majority of papers were published in workshops, which is consist with the large number of workshop papers. In 2019, approximately 2,038 papers were published in workshops[14] and 1,680 papers were published in conferences (ACL, EMNLP, NAACL, CONLL, CICLing), meaning 54.8% were published in workshops. In our data set, 46.8% of papers surveyed were published in workshops. The most number of papers were published in the largest conferences: ACL and EMNLP. Thus, while Table 1 suggests that discussions of race have been siloed to particular NLP applications, Figure 2 does not show evidence that they have been siloed to particular venues.

In Table 2, for all papers that use categorization schema to classify race, we show what racial categories they use. If a paper uses multiple schemes (e.g. collects crowd-sourced annotations of stereotypes associated with different races and also asks annotators to self-report their race), we report each scheme as a separate data point. This table does not include papers that do not specify racial categories (e.g. examine "racist language" without specifying targeted people or analyze semantic change of topics like "racism" and "prejudice"). Finally, we map terms used by papers to the ones in Table 2, e.g. papers examining African American vs. European American names are included in BW.

The majority of papers focus on binary

[14] https://www.aclweb.org/anthology/venues/ws/

Black/white racial categories. While many papers draw definitions from the U.S. census, very few papers consider less-commonly-selected census categories like Native American or Pacific Islander. The most common method for identifying people's race uses first or last names (10 papers) or explicit keywords like "black" and "white" (10 papers).

C Full List of Surveyed Papers

	Year	Venue	NLP Task	Task Type
Assimakopoulos et al. (2020)	2020	LREC	Abusive Language	Collect Corpus
Bommasani et al. (2020)	2020	ACL	Text Representations	Detect Bias
Chakravarthi (2020)	2020	Workshop	Abusive Language	Collect Corpus
Groenwold et al. (2020)	2020	EMNLP	Text Generation	Detect Bias
Gupta et al. (2020)	2020	Workshop	Sector-spec. NLP apps.	Collect Corpus
Huang et al. (2020)	2020	LREC	Abusive Language	Detect Bias
Jiang and Fellbaum (2020)	2020	Workshop	Text Representations	Detect Bias
Joseph and Morgan (2020)	2020	ACL	Text Representations	Detect Bias
Kennedy et al. (2020)	2020	ACL	Abusive Language	Debias
Kurrek et al. (2020)	2020	Workshop	Abusive Language	Collect Corpus
Lepori (2020)	2020	COLING	Text Representations	Detect Bias
Liu et al. (2020)	2020	COLING	Text Generation	Debias
Meaney (2020)	2020	Workshop	Social Science/Media	Survey/Position
Nangia et al. (2020)	2020	EMNLP	Text Representations	Detect Bias
Roy and Goldwasser (2020)	2020	EMNLP	Social Science/Media	Analyze Corpus
Sap et al. (2020)	2020	ACL	Abusive Language	Collect Corpus
Shah et al. (2020)	2020	ACL	Ethics/Task-indep. Bias	Survey/Position
Shahid et al. (2020)	2020	Workshop	Social Science/Media	Analyze Corpus
Tan et al. (2020)	2020	ACL	Ethics/Task-indep. Bias	Develop Model
Xia et al. (2020)	2020	Workshop	Abusive Language	Debias
Zhang et al. (2020)	2020	ACL	Abusive Language	Detect Bias
Zhao and Chang (2020)	2020	EMNLP	Ethics/Task-indep. Bias	Detect Bias
Amir et al. (2019)	2019	Workshop	Sector-spec. NLP apps.	Analyze Corpus
Davidson et al. (2019)	2019	Workshop	Abusive Language	Detect Bias
Demszky et al. (2019)	2019	NAACL	Social Science/Media	Analyze Corpus
Gillani and Levy (2019)	2019	Workshop	Text Representations	Analyze Corpus
Jurgens et al. (2019)	2019	ACL	Abusive Language	Survey/Position
Karve et al. (2019)	2019	Workshop	Text Representations	Debias
Kurita et al. (2019)	2019	Workshop	Text Representations	Detect Bias
Lauscher and Glavaš (2019)	2019	Workshop	Text Representations	Detect Bias
Lee et al. (2019)	2019	Workshop	Text Generation	Detect Bias
Liu et al. (2019)	2019	CoNLL	Social Science/Media	Develop Model
Manzini et al. (2019)	2019	NAACL	Text Representations	Debias
May et al. (2019)	2019	ACL	Text Representations	Detect Bias
Mayfield et al. (2019)	2019	Workshop	Sector-spec. NLP apps.	Survey/Position
Merullo et al. (2019)	2019	EMNLP	Social Science/Media	Analyze Corpus
Mostafazadeh Davani et al. (2019)	2019	EMNLP	Core NLP Applications	Develop Model
Parish-Morris (2019)	2019	Workshop	Sector-spec. NLP apps.	Survey/Position
Romanov et al. (2019)	2019	NAACL	Sector-spec. NLP apps.	Debias
Santos and Paraboni (2019)	2019	RANLP	Social Science/Media	Collect Corpus
Sap et al. (2019)	2019	ACL	Abusive Language	Detect Bias
Sharifirad and Matwin (2019)	2019	Workshop	Abusive Language	Analyze Corpus
Sommerauer and Fokkens (2019)	2019	Workshop	Text Representations	Detect Bias
Tripodi et al. (2019)	2019	Workshop	Text Representations	Analyze Corpus
Vylomova et al. (2019)	2019	Workshop	Social Science/Media	Analyze Corpus
Wallace et al. (2019)	2019	EMNLP	Text Generation	Detect Bias
Xu et al. (2019)	2019	INLG	Text Generation	Develop Model
Barbieri and Camacho-Collados (2018)	2018	*SEM	Social Science/Media	Analyze Corpus

Blodgett et al. (2018)	2018	ACL	Core NLP Applications	Debias
Castelle (2018)	2018	Workshop	Abusive Language	Analyze Corpus
de Gibert et al. (2018)	2018	Workshop	Abusive Language	Collect Corpus
Elazar and Goldberg (2018)	2018	EMNLP	Ethics/Task-indep. Bias	Debias
Kasunic and Kaufman (2018)	2018	Workshop	Text Generation	Survey/Position
Kiritchenko and Mohammad (2018)	2018	*SEM	Social Science/Media	Detect Bias
Loveys et al. (2018)	2018	Workshop	Sector-spec. NLP apps.	Analyze Corpus
Preoţiuc-Pietro and Ungar (2018)	2018	COLING	Social Science/Media	Develop Model
Sheng et al. (2019)	2018	EMNLP	Text Generation	Detect Bias
Wojatzki et al. (2018)	2018	LREC	Social Science/Media	Collect Corpus
Wood-Doughty et al. (2018)	2018	Workshop	Social Science/Media	Develop Model
Clarke and Grieve (2017)	2017	Workshop	Abusive Language	Analyze Corpus
Gallagher et al. (2017)	2017	TACL	Social Science/Media	Develop Model
Hasanuzzaman et al. (2017)	2017	IJCNLP	Abusive Language	Develop Model
Ramakrishna et al. (2017)	2017	ACL	Social Science/Media	Analyze Corpus
Rudinger et al. (2017)	2017	Workshop	Core NLP Applications	Detect Bias
Schnoebelen (2017)	2017	Workshop	Ethics/Task-indep. Bias	Survey/Position
van Miltenburg et al. (2017)	2017	INLG	Image Processing	Detect Bias
Waseem et al. (2017)	2017	Workshop	Abusive Language	Survey/Position
Wood-Doughty et al. (2017)	2017	Workshop	Social Science/Media	Analyze Corpus
Wright et al. (2017)	2017	Workshop	Abusive Language	Analyze Corpus
Blodgett et al. (2016)	2016	EMNLP	Ethics/Task-indep. Bias	Collect Corpus
Pavlick et al. (2016)	2016	EMNLP	Core NLP Applications	Collect Corpus
Waseem (2016)	2016	Workshop	Abusive Language	Detect Bias
Waseem and Hovy (2016)	2016	Workshop	Abusive Language	Collect Corpus
Mohammady and Culotta (2014)	2014	Workshop	Social Science/Media	Develop Model
Bergsma et al. (2013)	2013	NAACL	Social Science/Media	Develop Model
Herbelot et al. (2012)	2012	Workshop	Social Science/Media	Analyze Corpus
Warner and Hirschberg (2012)	2012	Workshop	Abusive Language	Develop Model
Eisenstein et al. (2011)	2011	ACL	Social Science/Media	Analyze Corpus
Somers (2006)	2006	Workshop	Sector-spec. NLP apps.	Survey/Position