Interdependencies of Gender and Race in Contextualized Word Embeddings

May Jiang
Computer Science
Princeton University
Princeton, NJ
mayjiang@princeton.edu

Christiane Fellbaum
Computer Science, Linguistics
Princeton University
Princeton, NJ
fellbaum@princeton.edu

Abstract

Recent years have seen a surge in research on the biases in word embeddings with respect to gender and, to a lesser extent, race. Few of these studies, however, have given attention to the critical intersection of race and gender. In this case study, we analyze the dimensions of gender and race in contextualized word embeddings of given names, taken from BERT, and investigate the nature and nuance of their interaction. We find that these demographic axes, though typically treated as physically and conceptually separate, are in fact interdependent and thus inadvisable to consider in isolation. Further, we show that demographic dimensions predicated on default settings in language, such as in pronouns, may risk rendering groups with multiple marginalized identities invisible. We conclude by discussing the importance and implications of intersectionality for future studies on bias and debiasing in NLP.

1 Introduction

In recent years, the rapid growth of natural language processing (NLP) has been accompanied by increasing attention to biases in NLP systems, with several studies investigating the demographic biases inherited by word embeddings from human text corpora. The vast majority of these studies have focused on measuring or mitigating bias with respect to gender, while fewer have concentrated on stereotypes with respect to race (Rozado, 2020). Even fewer of these studies have considered the intersection of gender and race.

Intersectionality, however, is a framework as critical as ever. Coined by black feminist scholar Kimberlé Crenshaw in 1989, intersectionality has since gained momentum as a vital framework to highlight and understand the powerful ways that different dimensions of an individual's identity combine and interact to create unique experiences of discrimination; Crenshaw elucidates the intersectional experience associated with race and sex as greater than the sum of racism and sexism, such that "any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordinated" (Crenshaw, 1989).

One consequence illuminated by this framework is that, because "people with multiple subordinate-group identities (e.g., ethnic minority women) do not fit the prototypes of their respective identity groups (e.g., ethnic minorities, women)," they and their unique experiences are often erased from the conversation – they experience 'intersectional invisibility' (Purdie-Vaughns and Eibach, 2008). In the field of psychology, due to systematic underrepresentation of certain groups in participant samples, "much of what is known about women in psychology is based on responses from women who are White and often middle class" (Cole, 2009). At an institutional level, "law enforcement, the government, and research institutions measure 'gender' as 'white women' and 'race' as 'African-American men'," and because of the prevalence of this "dualistic" approach in the social sciences, African-American women receive weaker attention in studies focusing solely on race or gender (Brown, 2010).

Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 17–25
Barcelona, Spain (Online), December 13, 2020.

We find this gap to be true of the social and linguistic properties characterizing contextual word embeddings as well. In this case study, we investigate the concepts of intersectionality inherent in the relationships and geometries of contextualized word embeddings. Specifically, this paper:

- examines the dimensions of gender and race in English-language BERT embeddings of European American and African American male and female names, finding that the concept of gender learned by the embeddings is closely dependent on race,
- explores the relationships between race and gender as captured by names and pronouns to show how default settings can contribute to the erasure of marginalized groups, and
- discusses the implications of these interdependencies and the vital importance of intersectionality for future studies on bias in NLP.

Bias Statement

In this work, we explore the intersection between race and gender, and our study of bias centers on the disparity between racial groups in the performance of the word embedding model on learning the concept of gender. If a system fails to recognize certain significant human and demographic characteristics of the non-prototypical members of a group, this results in a representational harm by which groups with multiple marginalized identities are erased. In particular, we draw on the description of 'coded exposure' given by sociologist Ruha Benjamin as the problematic phenomenon that "some technologies fail to see Blackness, while others render Black people hypervisible and expose them to systems of racial surveillance" – to be "watched, but not seen" (Benjamin, 2019). If gender and race are intertwined rather than independent, those at the intersection of marginalized racial and gender identities may be exposed to harmful biases related to the intersection and interaction of these identities, without being recognized for their identities themselves. Meanwhile, approaches that study, utilize, or debias these representations, but treat race and gender as orthogonal and isolated dimensions, would be poorly adapted for these groups.

Various studies have contended that because word embeddings are used as representations for text in a wide range of NLP tasks, any biases carried in the embeddings may be propagated to or even amplified in those downstream applications (Bolukbasi et al., 2016; Zhao et al., 2017; Bordia and Bowman, 2019). Certainly, this should be cause for concern, with the rising number of real-world applications and potential to impact lives. That said, we concur with Blodgett et al. (2020) that, independent of any allocational harms, the existence of these representational biases poses a critical problem in and of itself.

2 Related Work

The vast majority of studies of bias in word embeddings and contextual representations have focused on binary gender in isolation (Bolukbasi et al., 2016; Zhao et al., 2017; Zhao et al., 2019; Basta et al., 2019; Bordia and Bowman, 2019; Gonen and Goldberg, 2019). This is likely due to the advantage that gender pronouns are marked in language in a visible way that other demographic attributes are not. 'Bias', in these studies, is often measured as the similarity of words that are by definition gender-neutral – typically profession words – to words that are explicitly male or female gendered, such as pronouns. One popular method proposed for mitigating bias in word embeddings involves reducing the projection of gender-neutral words in the direction of a gender dimension vector defined by the difference between the vectors of 'he' and 'she' (Bolukbasi et al., 2016). Another measure of bias in word embeddings, the Word Embedding Association Test (WEAT), was introduced by Caliskan et al. (2017), adapted from the Implicit Association Test (IAT) of humans' reaction times to pairs of words (Greenwald et al., 1998).

More recently, studies have begun to factor intersectionality into their analyses of bias. Studying bias in sentence encoders, May et al. (2019) generalized the WEAT to sentence embeddings, using bleached sentence templates to test a number of hypotheses. Though several of their hypotheses are centered on gender, they also include a test of the "Angry Black Woman" stereotype and find significant evidence confirming its presence (May et al., 2019).

Tan and Celis (2019) similarly build on the WEAT, extending it to contextualized word embeddings and measuring intersectional bias as the association of European American and African American male

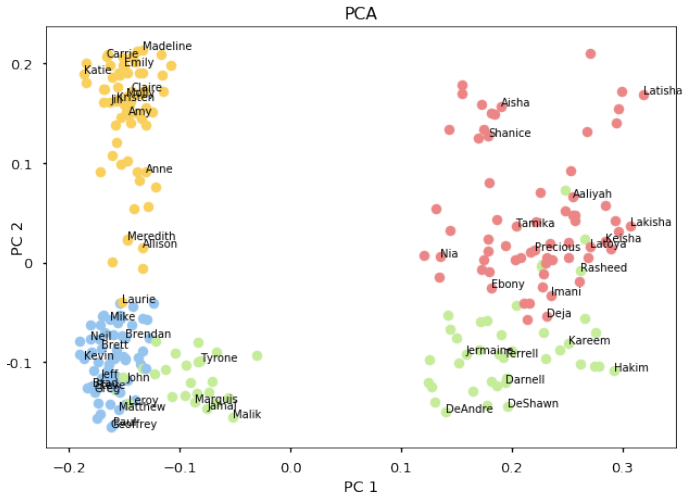


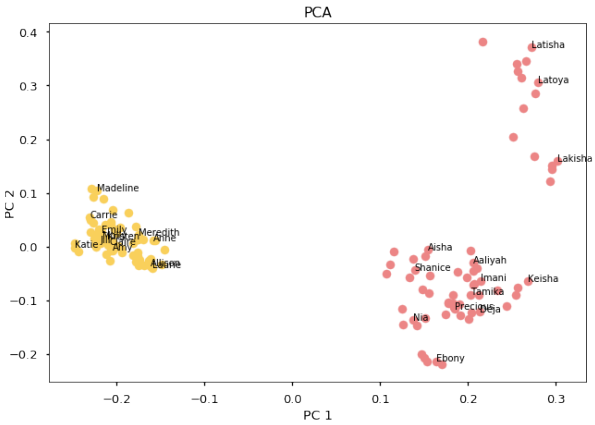
Figure 1: African American (AA) and European American (EA) male and female names projected onto their first and second principal components (AA female names in red, EA female names in yellow, AA male names in green, EA male names in blue)

and female names with pleasant and unpleasant words; they find significant evidence of bias, noting that the highest measured difference in bias is that between the names from the group with multiple privileged identities – European American men – against the names from the group with multiple marginalized identities – African American women. Further, Guo and Caliskan (2020) propose a method for identifying words that are unique to an intersectional group, and find numerous words pertaining to negative stereotypes associated significantly more strongly with an intersectional identity than with its constituent identities (Guo and Caliskan, 2020). To our knowledge, however, to date there has not been any systematic study or exploration of the conceptualization of gender and race learned by word embeddings, how these demographic dimensions relate or interact, or how bias may be embedded in gendered words such as pronouns themselves; the objective of this paper is to begin to fill in these gaps.

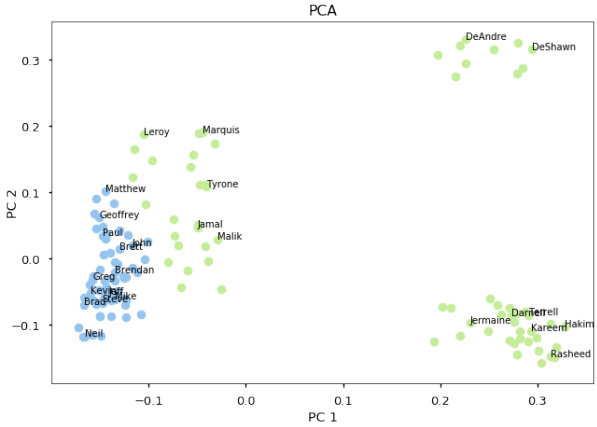
3 Learning Gender and Race in Word Embeddings from Names

To understand how and to what extent the concepts of gender and race are learned by contextualized word embeddings, in this case study we use contextualized embeddings of African American and European American male and female names from the pretrained BERT base-cased model (Devlin et al., 2019), with the sets of popular American given names borrowed from previous studies (Tan and Celis, 2019; May et al., 2019; Caliskan et al., 2017). These sets consist of 13 names for each race-gender pair, and were selected from Greenwald et al. (1998)'s original set of names used in introducing the IAT (Caliskan et al., 2017). We base our analysis on the principal component analysis (PCA) of these names, a common approach to identifying, visualizing, and understanding dimensions of gender and race in a word embedding vector space (Bolukbasi et al., 2016; Sedoc and Ungar, 2019; Manzini et al., 2019).

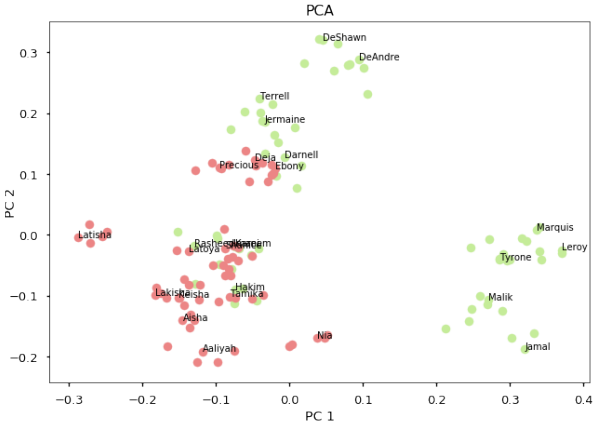
Figure 1 plots these names projected onto the first and second principal components of the PCA of all the names in the four sets. On inspection, the first principal component decidedly distinguishes the race of the names; all of the European American names have negative projections onto the component and all African American names, with the exception of a few male names for which racial association may have been more ambiguous, have positive projections. The second principal component appears to distinguish gender, cleanly separating the European American female names from the European American male names along the axis. However, though the PCA was performed on an equal number of name embeddings from each group, this gender component is significantly less successful in capturing the gender of African American names. African American female names, and male names to a lesser extent, have projections on this gender dimension close to zero, and there is substantial overlap in that region between the male and female names in a way that is in visible contrast to that of the European American names.



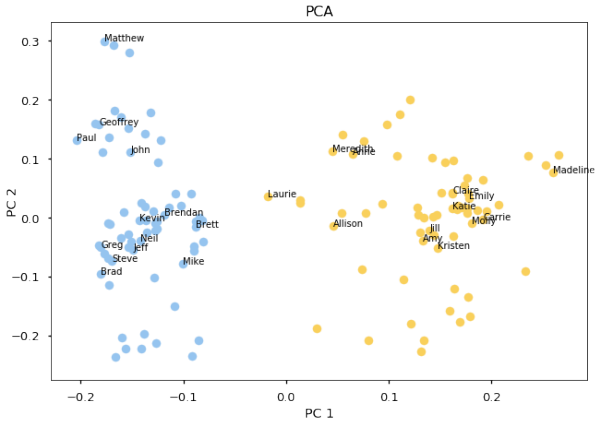
(a) EA female names (yellow), AA female names (red)



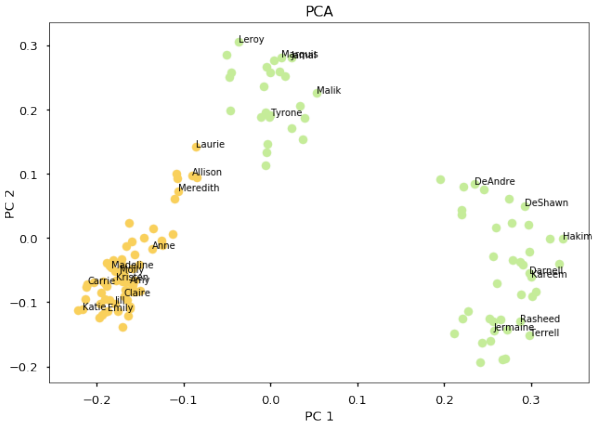
(b) EA male names (blue), AA male names (green)



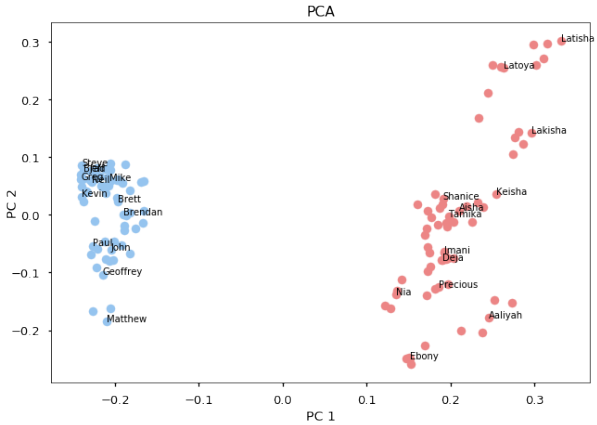
(c) AA female names (red), AA male names (green)



(d) EA male names (blue), EA female names (yellow)



(e) EA female names (yellow), AA male names (green)



(f) EA male names (blue), AA female names (red)

Figure 2: EA and AA male and female names projected onto their first and second principal components

Figure 2a plots the projections of the contextualized embeddings for the European American and African American female names on the first and second principal components of the PCA performed on names solely from those two sets, and the subsequent subfigures likewise plot the projections for European American and African American male names, African American male and female names, European American male and female names, African American male names and European American female names, and European American male names and African American female names.

Through the results of PCA on these subsets of the names, a few key themes emerge. First, it is immediately apparent that the concepts of gender and race are not learned uniformly across demographic groups; the juxtaposition of Figure 2c with Figure 2d illuminates this clearly, as the first principal component of the European American names linearly separates them by gender while neither the first nor the second principal component of the African American names results in a clean separation. To treat gender as a concept independent of race, then, risks the erasure of African Americans.

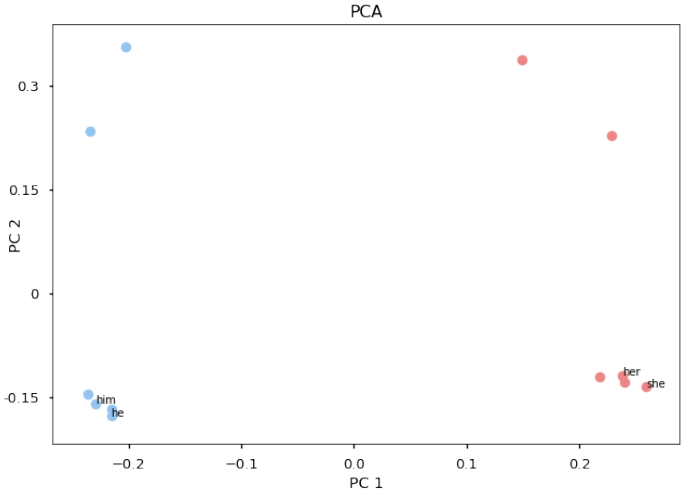
Further, we find that these concepts are not learned uniformly even within a single demographic group. That a small set of African American male names associate more closely with European American names as in Figure 2b is revealing of the heterogeneity of race, but even among the African American names that project onto the same direction of the first principal component – the race dimension – there is much more dispersion along that axis. It is worth noting that even given an equal number of names of each gender-racial group, the first principal components of the names' contextualized embeddings – the components that explain the most variance in the embeddings – are dimensions of race and gender that are better adapted to European American names.

Finally, across these experiments we find that the influence of race dominates that of gender in the embedding subspace. In Figure 1, the first principal component definitively distinguishes race, while the second principal component, more ambiguously for African American female names, distinguishes gender. Likewise, in Figure 2, the comparative ease with which the first principal components distinguish African American names from European American names is significant, as African American women's names associate much more closely with African American men's names than with European American women's names. This is striking, particularly in light of sociological studies on the centrality of race in understanding oppression, and the primacy of the role that race plays, even with the influence of factors such as race and class (Gillborn, 2015). Moreover, the weakness of the embedding model to recognize the femininity of African American women's names mirrors the biases in society that systematically marginalize African American women. Taken together, these themes underscore that the concepts and dimensions of race and gender in the BERT embedding space are not independent, and that ignoring their interaction may have harmful consequences for vulnerable populations.

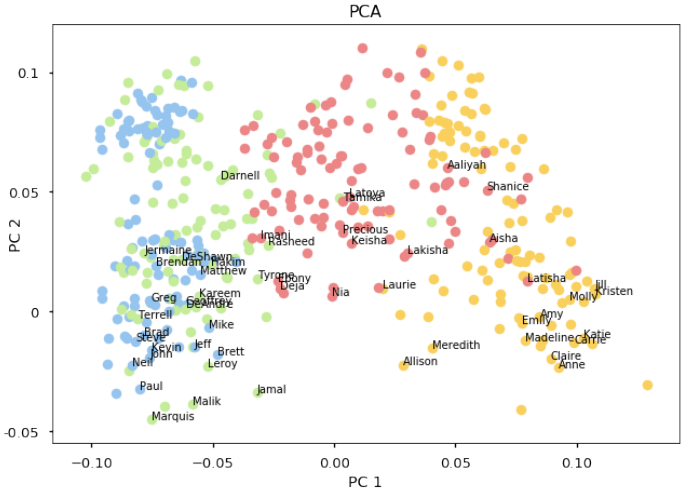
4 Default Settings and Intersectional Invisibility

Studies of gender bias in word embeddings typically use either male and female names or pronouns for measurement and debiasing. As we have found, however, the understanding of gender learned by the model and applied to names depends significantly on the race of the names. In this section, we investigate whether pronouns – by definition neutral with respect to race – are, in fact, independent of race.

Figure 3a shows the PCA of the contextualized embeddings of the gender pronouns, visibly separated by gender along the first principal component axis. More illustratively, Figure 3b shows the European American and African American male and female names projected onto the first and second principal component dimensions obtained from the PCA of the gender pronouns in Figure 3a. We find that while European American names and African American men's names align definitively with the directions of their corresponding gendered pronouns, the names of African American women do not project significantly onto that gender dimension in the direction of the female pronouns. Since the default setting of race learned by the model is European American, 'she' is taken to mean European American women, and as the default setting of gender learned by the model is male, African American men become race-prototypical. However, because the race and gender dimensions are not independent, and the race dimension dominates the gender dimension such that the African American women's names associate more closely with the African American men's names, African American women – at the intersection of



(a) Male (blue) and female (red) pronouns projected onto their first and second principal components



(b) AA and EA male and female names projected onto the first and second principal components of the gender pronouns from plot (a) above (AA female names in red, EA female names in yellow, AA male names in green, EA male names in blue)

Figure 3: Projections of pronouns and names on gender dimension from PCA of gender pronouns

marginalized gender and racial identities – experience intersectional invisibility to the language model as it is unable to associate their names with female pronouns.

Table 1 shows the cosine similarity between gender and race dimensions as identified using the first principal components of PCA on sets of names or pronouns. The gender dimension, for instance, is the first principal component of PCA on the words 'he', 'him', 'she', and 'her', with female pronouns' projections on that dimension in the positive direction. The race dimension using male names, analogously, is the first principal component of the PCA on the European American and African American male names, with African American names projecting on the component in the positive direction.

The dimensions sharing the highest similarity are the race dimensions for male and female names with a cosine similarity of 0.5536, which reinforces that race is the predominant demographic attribute captured by the word embeddings of the names. Likewise, there is a high cosine similarity of 0.5005 between the gender dimension derived from the gender pronouns and that derived from the European American male and female names. In contrast, the cosine similarity between the gender dimension derived from the pronouns and the gender dimension derived from the African American male and female names is close to zero. Further, while the gender dimension from the pronouns has a cosine similarity of only 0.036 with the race dimension from the European and African American male names, it shares a noticeable negative cosine similarity of -0.1461 with the race dimension derived from the European

Dimension	Dimension	Cosine similarity
Gender (male-female pronouns)	Gender (male-female EA names)	0.5005
Gender (male-female pronouns)	Gender (male-female AA names)	-0.0279
Gender (male-female pronouns)	Race (EA-AA male names)	0.0360
Gender (male-female pronouns)	Race (EA-AA female names)	-0.1461
Gender (male-female EA names)	Gender (male-female AA names)	-0.0177
Race (EA-AA female names)	Race (EA-AA male names)	0.5536

Table 1: Cosine similarity of gender and race dimensions identified by PCA using pronouns and names

and African American female names. That is, the gender dimension vector in the direction male-female has a negative relationship with the race dimension vector in the direction European American-African American, with the ultimate result that the names of African American women are marginalized from both the perspective of invisibility to the concept of gender as determined by pronouns, and conversely from the erroneous and unwanted exposure to the male direction of the gender dimension.

In sum, we find strong evidence that the concept and biases of race not only are not independent of the language model's understanding of gender as captured by pronouns, but are in fact embedded in that gender dimension itself.

5 Discussion and Conclusion

Above all, our findings reveal that language models cannot consider race and gender completely in isolation; since the word embedding space mirrors the reality of a society where racial and gender dimensions are intertwined, it is critically important to consider the significant interactive influences of gender and race. To fail to do so may be to fail to fully recognize both the unique experiences of those with intersecting identities and their positions as members of their constituent groups.

Further, the existing definitions, measures, and methods handling bias may be too narrow to encapsulate the nuances of diverse and multifaceted identities. We find that the human biases inherited by word embeddings are not merely manifested in the association of gendered words with gender-neutral terms relating to people, professions, or pleasantness and unpleasantness – as typically used in the measurement and analysis of bias – but are also able to fundamentally shape the relationships between the language, the labels, and the dimensions of identity themselves, as well as what these dimensions learn or fail to learn about groups of people.

This has profound implications for the way that bias is studied and measured, as well as mitigated. While previous studies have shown that the word embeddings for a set of names representing multiple marginalized identities are indeed shaped by biases in a unique and potentially more potent way compared to the embeddings for sets of names pertaining to each of their constituent identities, our finding that the racial and gender dimensions within word embedding subspaces are not independent implies that those most marginalized and vulnerable may be least seen and protected by methods that 'debias' by operating on isolated dimensions of race or gender built on default settings. Future work may extend our case study and analysis to specific demographic biases rooted in human studies, to other intersectional identities, considering potential biases with respect to demographic attributes such as religion, nationality, or age, to other words beyond proper names that may also carry biases, such as professions and adjectives, as well as to other embeddings and language models across different languages and cultures.

Ultimately, the weakness of word embeddings or language models to adequately, fairly, or uniformly learn the concept of gender for different racial groups and the dearth of research on intersecting identities and resulting biases concerning African American women and other multiply marginalized groups reveal an important blind spot in NLP. These manifestations of bias not only highlight the underlying complexities of inequality inherent in these models but also shed light on the parallel nuances of inequality in society and in the lived experiences of those individuals, further validating the intersectional framework as a crucial piece to understanding and unraveling the biases embedded in technology and society.

References

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Ruha Benjamin. 2019. *Race after technology: abolitionist tools for the new Jim code*. Polity, Medford, MA.

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 ACL*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Geneva Brown. 2010. The intersectionality of race, gender, and reentry: Challenges for african-american women. *Issue Brief.. Washington, DC: American Constitution Society*.

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

Elizabeth R. Cole. 2009. Intersectionality and research in psychology. *American Psychologist*, 64(3):170–180.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

David Gillborn. 2015. Intersectionality, critical race theory, and the primacy of racism: Race, class, gender, and disability in education. *Qualitative Inquiry*, 21(3):277–287.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614, Minneapolis, Minnesota, June. 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.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *arXiv preprint arXiv:2006.03955*.

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*, 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*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Valerie Purdie-Vaughns and Richard P. Eibach. 2008. Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles*, 59:377–391.

David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLOS ONE*, 15(4):1–26, 04.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy, August. Association for Computational Linguistics.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13230–13241.

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 EMNLP*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.